Heuristic Molecular Lipophilicity Potential for Computer-Aided Rational Drug Design

A Thesis

Submitted to the College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

in the Department of Chemistry

University of Saskatchewan

Saskatoon

By

Qishi Du

Spring 1998

Acknowledgments

I express sincere thanks to Dr. P.G. Mezey, for his excellent supervision, academic guidance, and emotional encouragement throughout my graduate study in this department. In particular, his financial support has been the necessary condition for me to finish the two-year study at the University of Saskatchewan. His motivation, initiative, industrious and rigorous style penetrates all parts of my thesis. He has been not only a trusted supervisor, but also a very respected friend during my stay in Saskatoon.

I also want to thank Mrs. Heather Acton for carefully checking the English of my thesis. This is very beneficial in improving my English and is important for conveying my ideas correctly.

Special thanks are sent to Dr. Ron Verrall for carefully checking and reading my thesis. His thoughtful and knowledgeable classes of solution theory are greatly helpful for my model HMLP. The long discussions with Dr. Verrall solve many questions in the model HMLP and give me the right direction in my thesis research. Many new ideas come from his classes and discussions.

Thanks are also sent to Dr. Gustavo A. Arteca (Laurentian University) for his enlightening discussions and inspiring encouragement.

I thank external examiner Dr. Mariusz Klobukowski, University of Alberta, and other members of my Thesis Committee, Dr. S. Pedras, Dr. S. Reid, and Dr. M. Khoshkam, for their time and energies spent on listening to my reports and examining my Thesis. I am very grateful to both the Dean's Office and the School of Graduate Studies and Research for their valuable Graduate Fellowship awarded to me. I extend thanks to all staff and students of the Chemistry Department, for the academic atmosphere they provide.

Publications in Thesis Research

1. **Qishi Du**, Paul G. Mezey, and Gustavo A. Arteca, "Heuristic molecular lipophilicity potential for computer-aided rational drug design", *Journal of Computer-Aided Molecular Design*, **11**(1997), 503-515.

2. **Qishi Du**, Paul G. Mezey, "Heuristic molecular lipophilicity potential for computer-aided rational drug design: Optimizations of screening functions", Accepted by *Journal of Computer-Aided Molecular Design*, (1998).

3. **Qishi Du**, Paul G. Mezey, "Heuristic molecular lipophilicity potential for computer-aided rational drug design: An Application to the Small Molecular System of Pyrazole Derivatives", *Journal of Computer-Aided Molecular Design*, sent for review, (1998).

4. **Qishi Du** and Gustavo A. Arteca, "Modeling lipophilicity from the distribution of electrostatic potential on a molecular surface", *Journal of Computer-Aided Molecular Design*, **10**(1996), 133-144.

5. **Qishi Du** and Gustavo A. Arteca, "Derivation of Fused-Sphere Molecular Surfaces from the Electrostatic Potential Distribution", *Journal of Computational Chemistry*, **17**(1996), 1258-1268.

6. Isabel Rozas, **Qishi Du**, Gustavo A. Arteca, "Interrelation Between Electrostatic and Lipophilicity Potentials on Molecular Surfaces", *Journal of Molecular Graphics*, **13**(1995), 98-108.

ABSTRACT

In my thesis research, I suggest a heuristic molecular lipophilicity potential (HMLP), a structure-based technique requiring no empirical indices of atomic lipophilicity. The input data used in this approach are molecular geometries and molecular surfaces. The HMLP is a modified electrostatic potential, combined with the averaged influences from the molecular environment. Quantum mechanics is used in calculating the electron density function $\rho(\mathbf{r})$ and the electrostatic potential $V(\mathbf{r})$, and from this information a lipophilicity potential $L(\mathbf{r})$ is generated. The HMLP is a unified lipophilicity and hydrophilicity potential. The interactions of dipole and multipole moments, hydrogen bonds, and charged atoms in molecules are included in the hydrophilic interactions in this model. The HMLP is used to study hydrogen bonds and water-octanol partition coefficients in several examples. The calculated results show that HMLP gives qualitatively and quantitatively correct, as well as chemically reasonable results in cases where comparisons are available. These comparisons indicate that the HMLP has advantages over the empirical lipophilicity potential in many aspects. Three possible screening functions and parameters used in them are tested and optimized in this research. Power screening function, $b_i/||\mathbf{R}_i-\mathbf{r}||^r$, and exponential screening function, $b_i exp(-||\mathbf{R}_i-\mathbf{r}||/d_0)$, give satisfactory results. A new strategy for drug design and combinatory chemistry is presented based on HMLP, and is used in the study of a small molecular system, pyrazole and its derivatives. The mechanism of inhibition of LADH caused by pyrazole and its derivatives is explained based on the calculation results of HMLP indices. Good results are achieved in this example. Further improvements of screening function and visualization of HMLP by computer graphics are discussed. I suggest two possible visualization approaches of HMLP: a two-color system and a three-color system. Their possible applications are discussed. HMLP is suggested as a potential tool in computer-aided three-dimensional drug design, studies of 3D-QSAR, active structure of proteins, combinatory chemistry, and other types of molecular interactions.

Table of Contents

Chapter 1: Theoretical and Experimental Background of Molecular Lipophilicity

1.1 Hydration and Molecular Lipophilicity

Solvation, hydration, molecular lipophilicity and hydrophilicity are the main subjects of my thesis. They are related but different concepts. This section is an introduction to the above concepts. Section 1.1 also introduces the structure and properties of liquid water molecules, the most important solvent in the natural world, and for which a good understanding of molecular lipophilicity and hydrophilicity is necessary.

1.1.1 Solvation and Hydration

Solvation is a topic as old as physical chemistry. In an aqueous solution, solvation is called hydration. Traditionally, the term *solvation* means a solute is being solvated by solvents. Solvation has been studied from two different aspects: its macroscopic and microscopic features; and by two different approaches: thermodynamic properties and interaction mechanics.

In the thermodynamic approach certain thermodynamic quantities, such as standard free energy (or enthalpy or entropy) of solution, are used as measures of the corresponding functions of the solvation of a given solute in a given solvent. For many years solvation thermodynamics have traditionally been treated in the context of classical thermodynamics alone. However, solvation is a process at the molecular level based on local rather than macroscopic properties of the system. Therefore the statistical mechanical approach has to be combined with thermodynamics in the study of solvation. Ben-Naim suggests a precise definition for the solvation process of a molecule s in a fluid l as the process of transferring the molecule s from a fixed position in an ideal gas phase g into a fixed position in the fluid or liquid phase l. The process is carried out at constant temperature T and pressure P. Also, the composition of the system is unchanged [Ben-

Naim 1987]. When such a process is carried out, the molecule *s* is being solvated by the liquid phase *l*. Sometimes the solute molecule *s*, the solvation of which is being studied, is called the *solvaton*.

In the above definition, the solvaton is a particular molecule *s* which has been chosen to be placed at a fixed position and to study its solvation properties. This new concept was introduced merely to distinguish between the particular molecule *s* being studied and all other molecules in the system. Of course, such a distinction cannot be made in practice—one cannot tag a specific molecule. However, theoretically, one can do it. One can always write the partition function of a system having one solvaton at some fixed position. From the point of view of this system (excluding the solvaton), this partition function is equivalent to a partition function of a system subjected to an "external" field of force produced by the solvaton at a fixed position [Ben-Naim 1987 p.190].

After the definition of the process of solvation is introduced, the corresponding thermodynamic quantities can be introduced: solvation entropy, solvation energy, solvation volume, and so on, which refer to the changes in the corresponding thermodynamic quantities associated with the solvation process as defined above. The Gibbs energy of solvation of *s* in *l* is defined as

$$\Delta G^*_s = \mu_s^{*l} - \mu_s^{*g}, \tag{1-1}$$

where μ_s^{*l} and μ_s^{*g} are the pseudo chemical potential of *s* in the liquid and in an ideal-gas phase, respectively. In eq. (1-1) ΔG^*_s is the Gibbs energy change for transferring *s* from a fixed position in an ideal-gas phase into a fixed position in the liquid phase *l*. The pseudo chemical potential of *s* in the ideal gas phase μ_s^{*l} is expressed in statistics,

$$\mu_s^{*l} = -kT \ln q_s, \tag{1-2}$$

where q_s is the internal partition function (including rotational, vibrational, electronic, and nuclear contributions). It means that ΔG^*_s in eq. (1-1) includes all the effects due to the interaction between s and its entire environment. Theoretically, the Gibbs energy ΔG^*_s can be divided into two parts: the interaction Gibbs energy between l and s and the effect of l on the internal degrees of freedom of s. However, experimentally, it is not easy to distinguish the two parts in the experimental data.

A simple situation arises when the solvaton has no internal degrees of freedom or when these are effectively unaffected by the surroundings. In these cases the pseudo chemical potential of s in the liquid phase μ_s^{*l} is simply written

$$\mu_s^{*l} = W(s|l) - kT \ln q_s, \tag{1-3}$$

where q_s is exactly the same as in eq. (1-2). Hence the solvation Gibbs energy is reduced to

$$\Delta G_s^* = W(s|l), \tag{1-4}$$

which is simply the work of transferring the solvaton from s to l.

1.1.2 Molecular Interactions in Solvation

On the other hand, solvation is inherently a molecular process. Solvation is a phenomenon of the molecular interactions between a solute molecule and all solvent molecules. Solvation is the result of all types of molecular interactions: DLVO forces and non-DLVO forces [Israelachvili 1992]. DLVO forces include van der Waals interactions and double-layer interactions [Derjaguin and Landau 1941, Verwey and Overbeek 1948]. Non-DLVO forces are also called solvation forces and, in aqueous solution, hydration forces. Non-DLVO forces have a structural origin.

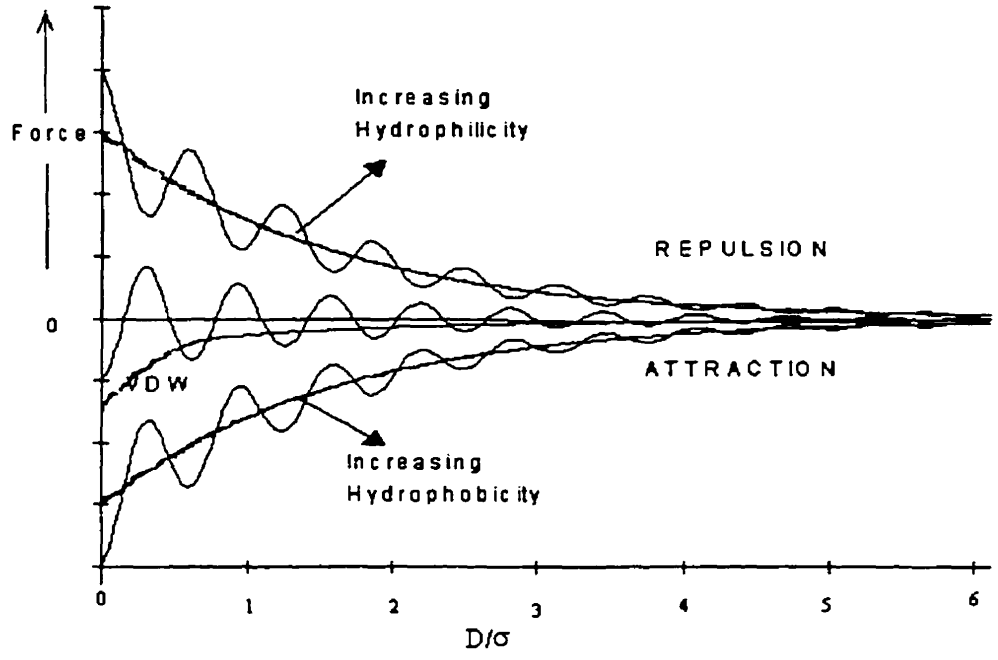


Figure 1-1. Oscillatory solvation force superimposed on a monotonic solvation force. D is the distance and σ is the dimension of a solvent molecule.

In a continuous medium model there is no solvation force. However, at distances near the dimension, σ, of solvent molecules, a continuous model no longer holds, and a discrete model must be used. Solvation forces arise at a short distance ($D<3\sigma$) whenever liquid molecules are induced to order into quasi-discrete layers. Surface-solvent interactions can induce a positional or orientational order in the adjacent liquid, therefore, solvation forces have a mainly geometric origin. Additional non-DLVO forces may also arise from the disruption of the liquid hydrogen-bonding network in the solution system, from electrostatic ion-binding and ion-correlation effects, and from molecular 'bridging' effects [Israelachvili 1992]. A solvation force is oscillatory as shown in Fig. 1-1. An oscillatory solvation force arises once there is an oscillatory change in the liquid density and orientation between the smooth and infinite hard surfaces as they approach each other. As the distance increases, $\rho_s(D)$ approaches the value for isolated surfaces $\rho_s(\infty)$ and oscillation approaches zero. Considering the properties of surfaces of macromolecules, the oscillatory solvation force is superimposed on the monotonic

solvation force. This type of interaction often arises in aqueous solutions where hydrogen-bond correlation effects can give rise to an additional monotonically decaying 'hydration' force (in addition to any oscillatory and DLVO force). For hydrophilic surfaces the monotonic component is repulsive, whereas for hydrophobic surfaces it is attractive [Israelachvili 1992].

The conventional explanation of why hydrophilic surfaces and macromolecules remain well separated in water is that they experience a monotonically repulsive hydration force owing to the structuring of water molecules at the surface. Based on recent experiments and theoretical results, Israelachvili and Wennerstrom [1996] suggest an alternative interpretation in which hydration forces are either attractive or oscillatory, and where repulsive forces originate from the properties of surfaces: the roughness and the flexibility of surface [Suresh and Walz 1996].

1.1.3 Structure of Water Molecules in Liquid

Molecular lipophilicity and hydrophilicity are the properties of solute molecules in aqueous solutions [Lemieux 1996]. Water is the most important and common solvent in the natural world and possesses many unique properties such as small size, high density, very high boiling and freezing points, tetrahedral charge distributions, and two hydrogen-bond donors and two acceptors. Figure 1-2 shows the widely used water model ST2 [Stillinger and Rahman 1974] in molecular modeling.

In the ST2 model, the water molecule is modeled with charges of +0.24e centered on each hydrogen atom and two compensating charges of -0.24e on the opposite side of the oxygen atom, representing the two unshared electron pairs. The four charges are located along four tetrahedral arms radiating from the center of the O atom. The interaction between two water molecules is assumed to involve an isotropic Lennard-Jones potential and 16 Coulombic terms representing the interactions between four point charges on one molecule with four on the other.

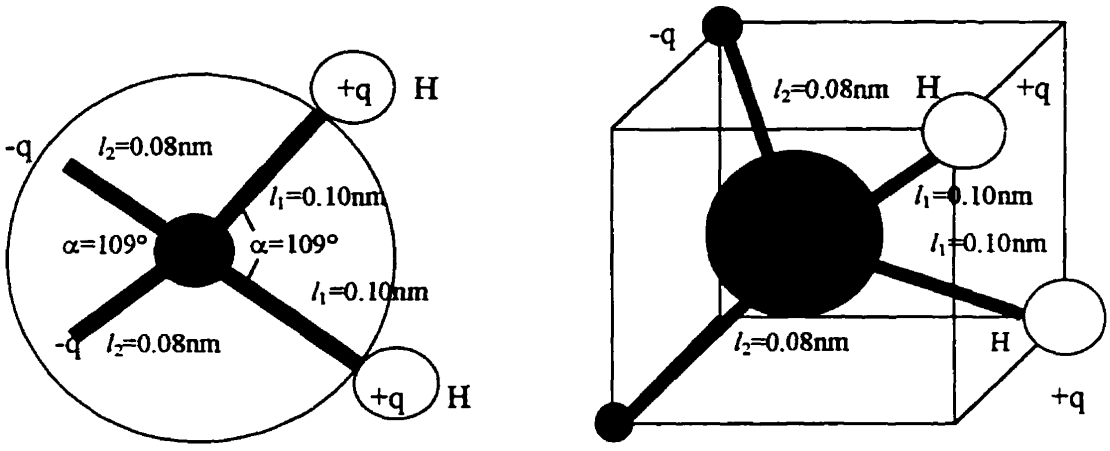


Figure 1-2. ST2 model of a water molecule; q=0.24e, l_1=0.1 nm, l_2=0.08nm, and α=109°.

Water is a highly associated liquid because of hydrogen bonds. In liquid water the tendency to remain in the ice-like tetrahedral network remains, but the ice structure in liquid water is distorted and labile. The average number of nearest neighbors per molecule rises to about five (hence the higher density of water upon melting), but the mean number of hydrogen bonds per molecule falls to about 3.5 with lifetimes of about 10^{-11} s. Around an inert solute molecule the water molecules actually have a higher coordination (of 4 hydrogen-bonds) and, thus, are even more ordered than in the bulk liquid. Both theoretical and experimental studies indicate that the reorientation, or restructuring, of water around non-polar solutes or surfaces is entropically very unfavorable, since it disrupts the existing water structure and imposes a new and more ordered structure on the surrounding water molecules [Israelachvili 1992, p. 129]. This is the origin of lipophilic effects.

The tetrahedral coordination of a water molecule, much more than the hydrogen bonds themselves, is at the heart of the unusual properties of water. Molecules that can participate in only two H bonds can link into a one-dimensional chain or ring (*e.g.*, HF and alcohols). Likewise, atoms that can participate in three bonds (*e.g.*, arsenic, antimony and carbon in graphite), can form two-dimensional sheets or layered structures held

together by weaker van der Waals forces. Only the tetrahedral, or higher coordination allows for a three-dimensional network to form.

1.1.4 Molecular Lipophilicity and Hydrophilicity

Molecular lipophilicity (or hydrophobicity) and hydrophilicity (or lipophobicity) are two opposite properties of solute molecules or particles in aqueous solutions. One means the tendency to attracting oil, the other means the tendency to attract water. Lipophilicity and hydrophilicity are used widely and frequently, and it is well known which molecules, even atoms, are lipophilic or hydrophilic. It seems that lipophilicity and hydrophilicity are *priori* properties. However, the concepts and natures of lipophilicity and hydrophilicity are not so clear [Israelachvili, 1992]. There are no precise definitions for these two concepts. Israelachvili [1992] gives an explanation for these two terms. He said "The immiscibility of inert substances with water, and the mainly entropic nature of this incompatibility is known as the *hydrophobic effects*" [Israelachvili 1992]. On the other hand, he said "While there is no phenomenon actually known as the hydrophilic effect or the hydrophilic interaction, such effects can be recognized in the propensity of certain molecules and groups to be soluble and to repel each other strongly in water, in contrast to the strong attraction exhibited by hydrophobic groups" [Israelachvili 1992].

It is clear that the above explanations are not precise definitions for molecular hydrophilicity and hydrophobicity, and seem to be recycled explanations. However, maybe these are the best explanations so far for these two concepts. The difficulty of the definitions and explanations of lipophilicity and hydrophilicity arises from the complexity of these two phenomena. Hydrophobic effects have a complex nature involving all types of interactions between solute molecules and a huge number of water molecules. Molecular lipophilicity and hydrophilicity are determined by both the properties of water molecules and the nature of solute molecules.

A basic principle of molecular modeling is that all types of molecular properties are decided by the molecular structure, including both electronic structure and geometric

structure. Suppose that the properties of water are the same for all types of solute molecules, then molecular lipophilicity and hydrophilicity are the properties of solute molecules. In my thesis research, I focus on the molecular structure of solutes, and try to find a unified measuring system and explanation for molecular lipophilicity and hydrophilicity. A model of molecular lipophilicity potential will be established based on molecular structure using molecular theoretical properties from quantum *ab initio* calculations.

1.2 Molecular Electrostatic Potential and Force Field

Molecular electrostatic potential (MEP) is the best physical quantity used in the study of molecular interactions [Tomasi 1981, Tasi and Pálinkó 1995]. Another relative and widely used theoretical property in the study of molecular interaction is molecular electrostatic force field (MEF) [Mishra and Kumar 1995].

1.2.1 Molecular Electrostatic Potential

A well-established and important approach to the study of molecular interactions is the molecular electrostatic potential (MEP). This 3D function, MEP, has a precise quantum mechanical definition. The electrostatic potential $V(\mathbf{r})$ is created in the space around a molecule by its nuclei and electrons. According to quantum mechanics, MEP is defined by the following equation,

$$V(\mathbf{r}) = \sum_{\alpha} \frac{Z_\alpha}{\|\mathbf{R}_\alpha - \mathbf{r}\|} - \int_{\infty} \frac{\rho(\mathbf{r}')}{\|\mathbf{r}' - \mathbf{r}\|} d\mathbf{r}' \,, \qquad (1\text{-}5)$$

where Z_α is the charge on nucleus α located at $\mathbf{R}_\alpha$, and $\rho(\mathbf{r})$ is the electron density function [Politzer 1981, Politzer and Murray 1991]. In eq. (1-5), the first term is the nuclear potential, and the second term is the contribution of electrons. The electrostatic potential is directly and rigorously related to the electronic density, both by eq. (1-5) and also by Poisson's equation,

$$\nabla^2 V(\mathbf{r}) = 4\pi\rho(\mathbf{r}). \qquad (1\text{-}6)$$

Eqs. (1-5) and (1-6) show a close relationship between electron density $\rho(\mathbf{r})$ and electrostatic potential $V(\mathbf{r})$. From the density functional concept that the energy of a system can be expressed as a function of its charge density, eq. (1-5) and (1-6) suggest the possible existence of a relationship between the energy of a system and its electrostatic potential [Politzer 1981].

Besides a rigorous theoretical foundation, MEP also has a solid experimental background. Both electron density and electrostatic potential are real physical properties and can be determined experimentally with scattering techniques. X-ray diffraction is used to determine electron density and electron diffraction is used to determine electrostatic potential [Fink and Bonham 1981]. Actually, in electron diffraction, the electrical potential of the target material scatters the incident electrons. However, when the energies of an electron beam are high enough, in the 20-50 kV range, the electrical potential is well approximated by electrostatic potential defined by eq. (1-5). The availability of reliable relationships between electrostatic potential and total energies or interaction energies would therefore make it possible to go directly from the quantities obtained with scattering experiments.

MEP is exactly equal in magnitude to the electrostatic interaction energy between the static (*i.e.*, unperturbed) charge distribution of the system and a positive unit point charge located at $\mathbf{r}$ as defined by eq. (1-5). MEP has a simple physical meaning: $V(\mathbf{r})$ is the interaction energy between the molecule and the unit probe point charge at position $\mathbf{r}$. This physical feature of MEP makes it a powerful tool in the study of molecular interactions on both fundamental and applied levels.

1.2.2 The Role of MEP in Molecular Interactions

In principle, most molecular interactions are initiated by electrostatic interaction between molecules. For chemists, there are various models to choose: from the very

simple point-multipole classical models to sophisticated quantum mechanical calculations. Among these models, electrostatic potential is the most useful "simple model" in understanding non-covalent intermolecular interactions.

As mentioned in the earlier section, the electrostatic interaction between a molecule and a unit point charge placed at **r** is simply given by $V(\mathbf{r})$. However, in the more general case of an interaction between two complex molecules, A and B, the electrostatic potential energy is not a simple function of the electrostatic potential. The energy components for their interaction are [Morokuma and Kitaura 1981]

(a) electrostatic ΔE_{es}: the interaction between the unperturbed charge distributions of the molecules,

(b) polarization ΔE_{pol}: the energy associated with the polarization of the charge in B by the electric field of A and vice versa,

(c) exchange repulsion ΔE_{ex}: Pauli principle repulsion between the electrons of monomer A with those of B,

(d) charge transfer ΔE_{ct}: the transfer of electrons from one monomer to the other, and

(e) dispersion ΔE_{disp}: the instantaneous dipole-dipole attraction observed even in rare gas atom interactions.

Only (a) - (d) can be calculated at the single-configuration SCF level [Kollman 1981] and the electrostatic term ΔE_{es} is the major contributor to the total interaction energy in most cases. Kollman [1977] has carried out an analysis for a series of hydrogen-bonding and other Lewis acid-Lewis base interactions using quantum ab initio calculations. He found that the electrostatic interaction energy is a function of the product of the corresponding electrostatic potentials at the reference positions,

$$\Delta E_{es}=kV_{es}(\mathbf{R}_A)V_{es}(\mathbf{R}_B), \qquad (1\text{-}7)$$

where $\mathbf{R}_A$ and $\mathbf{R}_B$ are reference positions for Lewis acid and base, respectively, and k is a constant with dimension of length. Based on the ab initio calculations at RHF/4-31G level, in atomic units, they found k=10.4 a_o (1 a_o=0.5292 Å).

Kollman [1981] gives a simple explanation for eq. (1-7) based on classical physics. If molecule A and B are treated as simple ions with charges q_A and q_B, each located at the reference distance from the other, then

$$V_{es}(\mathbf{R}_A)V_{es}(\mathbf{R}_B) = \frac{q_A q_B}{\mathbf{R}_A \mathbf{R}_B}. \tag{1-8}$$

If the distance to the reference positions are both the same and are approximately the A-B distance $\mathbf{R}_{AB}$ in the complex, then the product of the potentials is $q_A q_B / \mathbf{R}^2_{AB}$. Under the same assumptions, the electrostatic interaction energy is

$$\Delta E_{es} = \frac{q_A q_B}{\mathbf{R}_{AB}}. \tag{1-9}$$

Thus, the constant k in eq. (1-7) is merely $\mathbf{R}_{AB}$.

In the same way, for the interaction of point dipoles, in classical physics, electrostatic interaction energy is

$$\Delta E_{es} = \frac{2\mu_A \mu_B}{\mathbf{R}^3_{AB}}. \tag{1-10}$$

The product of the electrostatic potential of two point dipoles at reference position $\mathbf{R}_A$ and $\mathbf{R}_B$ is

$$V_{es}(\mathbf{R}_A)V_{es}(\mathbf{R}_B) = \frac{\mu_A \mu_B}{\mathbf{R}_{AB}^4}. \qquad (1\text{-}11)$$

Comparing with eq. (1-7), for point dipoles, the constant k is $2\mathbf{R}_{AB}$.

In a solution system, suppose A is a solute molecule and B is a solvent molecule. According to eq. (1-7), one might think that electrostatic potential $V(\mathbf{R}_A)$ is the measure of the interaction ability of molecule A with solvent molecule B at reference position $\mathbf{R}_A$. However, there are a huge number of solvent molecules. It is not a simple pair interaction of A and B according to eq. (1-11). In my thesis research, I will discuss this question.

1.2.3 Molecular Electric Force Field

Molecular electric force field (MEF) is a quantity related to the electrostatic potential as shown in the following equation,

$$E(\mathbf{r}) = -\nabla V(\mathbf{r}). \qquad (1\text{-}12)$$

Despite the fact that MEP and MEF are related by eq. (1-12), their spatial distributions may be quite different. Due to the vectorial character of MEF, its magnitude and direction can both be employed to illustrate molecular interactions and to evaluate molecular similarity. This property sometimes makes MEF more rigorous and more useful than MEP [Mishra and Kumar 1995].

Suppose the magnitude of the electric field, due to the charge distribution of a molecule at a point in its vicinity is E. If a point dipole having moment p is placed at that point, its potential energy of interaction with the electric field would be given by

$$W = -p\,E\cos\theta, \qquad (1\text{-}13)$$

where θ is the angle between the direction of the point dipole and that of the electric field. If the dipole is allowed to rotate freely, it would orient itself along the minimum energy direction (θ = 0) which would also be the direction of the field. Then

$$E = -\frac{W}{p}. \tag{1-14}$$

If E_A and E_B are the MEF of two molecules A and B at a corresponding point, one may use their scalar or vector product to evaluate similarity between two molecules [Dughan *et al.* 1991]. The vectorial aspect of MEF makes it a much better tool to describe the direction feature of hydrogen bonds than MEP [Mishra and Kumar 1995].

1.3 Ab Initio Calculation of MEP

In quantum mechanics, electrostatic potential is the expectation value of operator 1/**r**,

$$V(r) = \left\langle \Psi \left| \frac{1}{r} \right| \Psi \right\rangle, \tag{1-15}$$

where Ψ is the normalized wave function of a molecule. A serious question for the application of MEP in the study of molecular interactions is: how to find the wave functions of large biomolecules by ab initio calculations? Mezey [1995 a, b] has made great contributions to this topic.

1.3.1 Density Matrix and MEP

As shown in the definition eq. (1-5) of MEP, the key issue in the calculation of MEP is the electron density ρ(**r**). The second term in eq. (1-5) is the contribution of electrons to the MEP. After selecting a suitable basis set $\{\phi_\mu\}$, the contribution of electrons can be expressed by the following equation,

$$\int \frac{\rho(r')}{|r'-r|} dr' = \sum_{\mu,\nu} p_{\mu\nu} \int_{\infty} \frac{\phi_\mu^*(r')\phi_\nu(r')}{|r'-r|} dr', \tag{1-16}$$

where $p_{\mu\nu}$'s are the elements of density matrix **P**. Molecular orbitals are represented as the linear combination of basis functions,

$$\psi_\mu = \sum_i a_{\mu i} \phi_i . \tag{1-17}$$

For a closed shell molecule, the elements of density matrix are obtained by the following equation,

$$p_{kl} = 2\sum_{\nu}^{occ} a_{\nu k} a_{\nu l}, \tag{1-18}$$

where the summation is over all occupied molecular orbitals, and *occ* means the highest occupied molecular orbital.

Usually, ab initio quantum chemistry can only solve small and middle size molecules. Therefore, the limitation in the calculations of MEP by ab initio quantum chemistry is that of how to get the density matrix **P** for large macromolecules. This question was solved by Mezey [1995 a, b] based on the fuzzy additive electron density fragmentation principle.

1.3.2 Fuzzy Electron Density Fragmentation Principle

According to the Hartree-Fock-Roothaan-Hall SCF LCAO ab initio representation of a molecular wave function with respect to a fixed nuclear arrangement *K*, the electron density $\rho(\mathbf{r})$ of the molecule is defined in terms of a set of *n* atomic orbitals $\varphi_i(\mathbf{r})$,

$$\rho(r) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}\varphi_i(r)\varphi_j(r),\ \text{i=1, 2, ···, } n. \tag{1-19}$$

According to the fragmentation scheme, the set of nuclei of a molecule M is divided into *m* mutually exclusive groups,

$$f_1, f_2, \cdots, f_k, \cdots f_m \cdot \tag{1-20}$$

These nuclear families serve as AO reference locations when generating the corresponding *density fragments*,

$$F_1, F_2, \cdots, F_k, \cdots F_m\,, \tag{1-21}$$

or fragment density functions

$$\rho^1(\mathbf{r}), \rho^2(\mathbf{r}), \cdots, \rho^k(\mathbf{r}), \cdots \rho^m(\mathbf{r}), \tag{1-22}$$

defined in terms of the AO set of the molecule M and the family of fragment density matrices

$$\mathbf{P}^1, \mathbf{P}^2, \cdots, \mathbf{P}^k, \cdots \mathbf{P}^m, \tag{1-23}$$

respectively.

For an additive fuzzy density fragment, $\rho^k(\mathbf{r})$, the electron density $\rho(\mathbf{r})$ of molecule M is specified by an arbitrary subset *k* of nuclei and their "share" $\mathbf{P}^k$ of the density matrix **P** of the molecule. In practice it is advantageous to select nuclear families where the nuclei within a family are near one another. Using the simplest version of the additive fuzzy fragmentation method, the *k*th fuzzy electron density fragment $\rho^k(\mathbf{r})$ is calculated in terms of Mezey's additive fragment density matrix $\mathbf{P}^k$, defined as follows,

$$\begin{aligned} \mathbf{P}_{ij}^{k} &= \mathbf{P}_{ij} && \text{if both } \varphi_i(\mathbf{r}) \text{ and } \varphi_j(\mathbf{r}) \text{ are AO's centered} \\ & && \text{on nuclei of the } k\text{th fragment,} \\ &= 0.5\mathbf{P}_{ij} && \text{if only one of } \varphi_i(\mathbf{r}) \text{ and } \varphi_j(\mathbf{r}) \text{ is centered} \\ & && \text{on nuclei of the } k\text{th fragment,} \\ &= 0 && \text{otherwise.} \end{aligned} \tag{1-24}$$

Both the density matrix **P** of a complete molecule, and the additive fragment density matrix $\mathbf{P}^k$ of kth fragment have the same $n \times n$ dimensions. In terms of the full AO set of the molecule and the fragment density matrix $\mathbf{P}^k$, the electron density of Mezey's kth additive fuzzy density fragment, $\rho^k(\mathbf{r})$ is defined as

$$\rho^k(r) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}{}^k \varphi_i(r)\varphi_j(r), \ \text{i=1, 2, } \cdots, n. \tag{1-25}$$

If the nuclear families $f_1, f_2, \cdots, f_k, \cdots f_m$ are mutually exclusive, and if they collectively contain all the nuclei of molecule M, then eq. (1-24) defining the matrix elements $\mathbf{P}_{ij}^k$ implies that the sum of the fragment density matrices $\mathbf{P}^k$ is equal to the density matrix **P** of molecule M:

$$p_{ij} = \sum_{k=1}^{m} p_{ij}^{k}, \tag{1-26}$$

and

$$P = \sum_{k=1}^{m} P^k. \tag{1-27}$$

That is, the total molecular density matrix is the sum of all fragment density matrices. Furthermore, the linearity of the electron density expressions (1-19) to (1-25) in the matrix elements $\mathbf{P}_{ij}$ and $\mathbf{P}_{ij}^k$ of the molecular density matrix **P** and fragment density

matrices $\mathbf{P}^k$ implies that the sum of the fragment densities $\rho^k(\mathbf{r})$ is equal to the density $\rho(\mathbf{r})$ of molecule M:

$$\rho(r) = \sum_{k=1}^{m} \rho^k(r). \tag{1-28}$$

Consequently, at any given ab initio HF-LCAO level, the Mulliken-Mezey electron density decomposition scheme is an exactly additive, fuzzy electron density fragmentation scheme.

1.3.3 Additive Fuzzy Density Fragmentation Scheme

Three practical schemes for the construction of large biomolecules are suggested by Mezey [1995 a, b] and Walker [1993, 1994] based on the Mulliken-Mezey additivity of fuzzy electron density fragmentation, eq. (1-19)-(1-25). The three schemes are for different purposes and work on different levels.

1) Molecular Electron Density Lego Assembler (MEDLA)

The MEDLA technique [Walker and Mezey 1993, 1994] uses a numerical electron density MEDLA databank, containing pre-calculated electron density fragments obtained from calculations of smaller "parent" molecules containing "custom-made" nuclear geometry. In other words, MEDLA first calculates the smaller "parent" molecules, and saves the values of electron density at the cubic grid of molecular space in the databank. Then MEDLA uses the values of fragments saved in the databank to build the numerical electron density distribution of large molecules in the molecular space.

MEDLA saves a lot of computer CPU time and can build the electron density of huge macromolecules. However, the electron density of MEDLA is a set of discrete values on a cubic grid. MEDLA provides a good visual representation of the electron density distribution of macromolecules, but cannot provide an analytical electron density function $\rho(\mathbf{r})$. It is difficult to calculate MEP and other molecular properties based on a

set of discrete values of electron density. Another shortcoming is that a MEDLA databank occupies too much volume of the hard disk.

2) Adjustable Local Density Assembler (ALDA)

The additive fuzzy electron density fragmentation method can be used for the computation of macromolecular electron densities without relying on a numerical electron density fragment database. In other words, it can provide analytical density function $\rho(\mathbf{r})$. ALDA [Mezey 1995 b] only generates the fragment density matrices, and the actual density fragment density contributions are computed when they are needed. No numerical electron density database is generated, hence, there is no need for the storage of electron density values at several million grid points for each fragment. Instead, the ALDA method uses a much smaller ALDA database that stores the actual fragment density matrix elements for each $\mathbf{P}^k$, as well as associated nuclear geometry and basis set information. Evidently, this requires much less memory than a MEDLA database generating comparable electron densities.

The ALDA method is slower than the MEDLA method. The computer time requirement is determined by the number of fragments, and also increases linearly with the number of fragments, consequently, the overall computer time required for the ALDA method grows linearly with the size of the molecule. The disadvantage of the slower, but still linear, performance of the ALDA method is compensated by several advantages as follows.

(a) The ALDA database is much smaller than the MEDLA database, since the ALDA database contains only fragment density matrices, nuclear coordinates of parent molecules, and basis set information.

(b) ALDA can produce the analytical density function $\rho(\mathbf{r})$ for large macromolecules. Therefore it can provide enhanced resolution at some interesting location of the target molecule. The analytical density function provided by ALDA makes the calculations of macromolecular properties, such as MEP and MEF, much easier and more accurate than by the MEDLA method.

(c) Another important advantage is the versatility in the rapid, approximate computation of macromolecular electron densities for nuclear arrangements slightly distorted with respect to the arrangements found in the ALDA database.

3) Adjustable Density Matrix Assembler (ADMA)

The ADMA [Mezey 1995 b] macromolecular density matrix **P** is obtained by combining appropriately defined, mutually compatible, additive fragment density matrices $\mathbf{P}^k$. Mutual compatibility involves two conditions:

a) AO basis set orientation constraints, and
b) fragment choices fulfilling a compatible target-parent fragmentation condition.

The ADMA method uses a fragment density matrix database, similar to that of the ALDA method, however, these fragment density matrices fulfill the second of the above two compatibility conditions. By a suitable transformation, the fragment density matrices can be converted to physically equivalent fragment density matrices defined with respect to AO basis sets fulfilling condition (a).

The actual ADMA macromolecular density matrix constructed from the fragment density matrices represents the same level of accuracy as the MEDLA and ALDA methods. In particular, ADMA reproduces the effects of interactions between local fragment representations to the same level of accuracy as the MEDLA and ALDA methods. The ADMA density matrix technique also has provisions for the adjustability of the calculated electron density with respect to small nuclear geometry changes of a macromolecule, a feature similar to that of the ALDA method.

1.4 Partition Coefficients as a Measure of Molecular Lipophilicity

For a long time, partition coefficients of a solute between organic and water phases have been used as the measure of molecular lipophilicity [Leo *et al.* 1971]. A large number of partition coefficients have been determined using various experimental

methods over the past hundred years. The most common choice of solvent pairs is *n*-octanol (1-octanol) and water. The corresponding partition coefficient (P_{ow}) has been widely used in drug design.

1.4.1 Partition Coefficients

A partition coefficient is the ratio of concentrations of a compound distributed in the water phase and the organic phase. However, in most cases "partition coefficient" actually means the logarithm of a partition coefficient. Usually, the *n*-octanol and water solvent pair is used in the experiments of partition coefficients. Its simplest definition involves a ratio of molar concentrations C (see Fig. 1-3):

$$\log P_{OW} = \log \frac{C \text{ (in octanol)}}{C \text{ (in water)}} \quad . \tag{1-29}$$

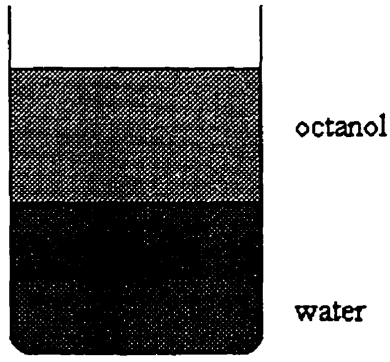


Figure 1-3. The partition coefficient of a solute distributed in the *n*-octanol phase and the water phase.

As a equilibrium constant for a two-phase system, log P_{ow} is determined by the difference between the solvation free energies of the solute in each phase. This difference is represented by the *partial molar standard free energy of transfer* $\Delta \overline{G}_{tr}^{\circ}$, *from the aqueous phase to the organic phase*,

$$\log P_{ow} = -\frac{\Delta \overline{G}^{\circ}_{tr}}{2.303RT} \quad . \tag{1-30}$$

Hydrophobicity is then represented in terms of the partition coefficient log P_{ow}. If a compound is strongly hydrophilic ("water-lover"), its concentration in the water phase is higher than in the organic phase, therefore the partition coefficient of this compound has a negative value, otherwise, a strong hydrophobic compound has a positive partition coefficient. However, log P_{ow} is an overall measure of molecular lipophilicity. It cannot tell the detailed local distributions of molecular lipophilicity in the molecular surface or space. Sometimes it is not very sensitive to the local lipophilicity change in a molecule. I will discuss this problem in Chapter 4 in more detail.

1.4.2 Partition Coefficients in ρ–σ–π Analysis

Drug design strategy, ρ–σ–π analysis, is an old technique. However, it provides a very good understanding of the role and importance of partition coefficients and molecular lipophilicity in rational drug design. The ρ–σ–π analysis was developed by Hansch and Fujita [1964] in the 1960s, and is still the basis of many new approaches to drug design. It serves as a good example for illustrating the use of lipophilicity in molecular modeling, and the relationship between molecular lipophilicity and biochemical activity.

The discovery and design of biologically active compounds can be classed into two different strategies: (1) the attempt to find new "lead" compounds, and (2) the attempt to fully exploit existing lead compounds. A "lead" compound is a molecule that has a biological activity of interest, although its activity may be weak or it may have some undesired side effects. The procedures for exploiting a lead compound are much more fully developed than those for discovering new lead compounds. In this latter aspect, there is still a great deal of work to do, to which theoretical chemists can make an

important contribution. Below, I discuss a standard technique for optimizing the activity of a lead compound taking into account its hydrophobicity.

A. Mechanism of Biological Responses

The mechanism of many biological responses caused by chemicals can be represented in simple terms as shown in Figure 1-4 below [Leo *et al.* 1971]:

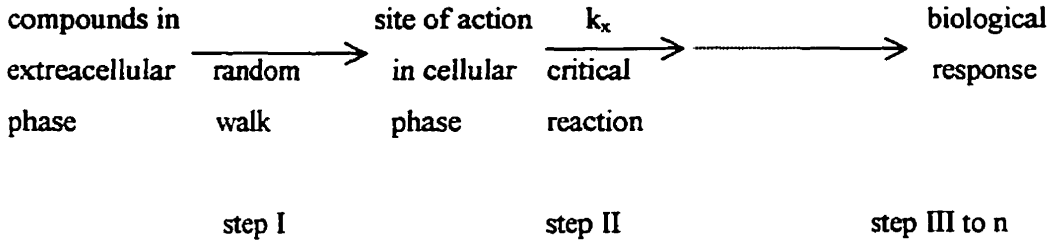


Figure 1-4. Mechanism of biological responses caused by chemicals

The first step in the above reaction scheme can be regarded as a random walk (or diffusion) process in which the molecule makes its way from a dilute solution surrounding the cell to a particular site on the cell (*e.g.*, the cell membrane or an organelle). It is well known that the internal cellular structure is very complex. A molecule would have to be partitioned between an "aqueous" phase and many different more or less "organic" phases when passing the wall membrane, then the endoplasm, and finally, the membrane of a particular organelle. It is clear that the partition coefficient between the aqueous and organic phases is a key point along these processes [Hansch and Fujita 1964].

There are several reasons for choosing the *n*-octanol-water system to model the behavior of drugs at biological interphases [Leo *et al.* 1971]. From a pragmatic viewpoint, the most compelling reason to use this system is the bulk of data available: thousands of compounds have been measured with this solvent pair. [For a brief table, see Leo *et al.* 1971.]

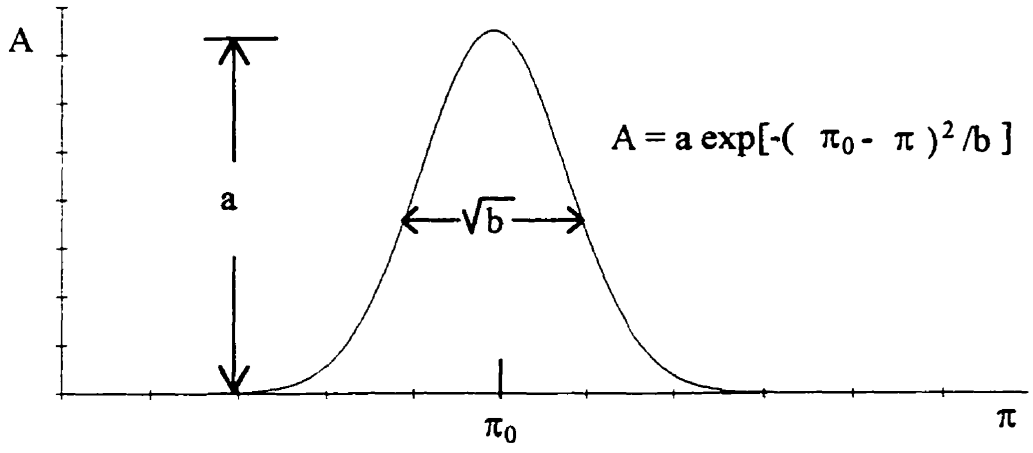


Figure 1-5. Probability factor *A* is a Gaussian function of the variable π

B. ρ–σ–π Analysis

In the research of quantitative structure-activity relationships (QSARs), linear free energy relations are a powerful tool. This technique is based on the formation of an empirical model of drug action that uses free energy-related parameters as the linear independent variables. The basic assumption of a linear free-energy relation is that all the factors involved in the biological activity triggered by a series of related compounds can be correlated with corresponding physicochemical (or structural) parameters. All physicochemical factors related to transport and receptor interactions can be broken down into hydrophobic, electrostatic, and steric components. The contribution of each of these components is expressed in substituent-dependent constants that represent the difference in properties between the lead compound and the derivative being studied.

The ρ–σ–π analysis, suggested by Hansch and Fujita [1964], involves a linear free-energy relation with three linear parameters: the Hammett parameters σ and ρ [Jaffé 1953], and the parameter π, introduced by Hansch and Fujita [1964]. It is based on two basic hypotheses. First, it is assumed that there will be one key rate-controlling reaction at the active sites of many biologically active molecules. This can be formulated as follows:

$$\text{rate of biological reaction} = d(\textit{response})/dt = AC\,k_X\,, \tag{1-31}$$

where A is the probability of the drug molecule reaching the site of action in a given time interval; C is the molar concentration of the drug outside the cell, and k_X is the reaction rate constant for the receptor-drug binding. The product of A and C is called the *effective concentration of the drug* accumulating on the active site [Leo *et al.* 1971]. The second hypothesis is that the factor A is a Gaussian function of a variable π:

$$A = f(\pi) = a \exp[-(\pi - \pi_o)^2/b]\ , \tag{1-32}$$

where a, b and π_0 are constants (see Fig. 1-5). The quantity π expresses the difference in the logarithms of the *n*-octanol–water partition coefficients, log P_S, the compound to be studied, and the log P_0 of the parent compound of the series of pharmaceutical analogues:

$$\pi = \log (P_S/P_0) = \log P_S - \log P_0\ . \tag{1-33}$$

From eqs. (1-31)–(1-33), it is clear that the transport of a drug molecule into a cell (and the biochemical action it triggers) is related to its partition coefficient log P_{OW}. Experiments indicate that there often exists an optimal partition coefficient for a biologically active series [Hansch and Fujita 1964]. The symbol π_0 is used for the value of π associated with the optimum P_{ow}. Any deviation from π_0 (*i.e.*, increasing or decreasing the partition coefficient), decreases the coefficient A. Hansch and Fujita [1964] suggested that the distribution of A would be normal, as shown in Eq. (1-32). Therefore:

$$d(\textit{response})/dt = C\,k_X\,a \exp[-(\pi - \pi_o)^2/b\,]\ . \tag{1-34}$$

Usually the drug concentration is adjusted until a particular rate of biological response is reached, *i.e.*, d(*response*)/d*t* is a constant. In this case, eq. (1-34) can be simplified by taking the logarithm and collecting constants:

$$\log(1/C) = -k\,\pi^2 + k'\,\pi\,\pi_o - k''\,\pi_o^2 + \log k_X + k''' \quad . \tag{1-35}$$

The Hammett equation [Jaffé 1953], which applies to either equilibrium or rate constants, establishes that:

$$\log k_X = \rho\sigma \quad , \tag{1-36}$$

where σ is an electronic structure parameter relating the molecule under study to the parent molecule, and ρ is a constant related to the type of reaction considered [Jaffé 1953]. Substitution of eq. (1-36) into eq. (1-35) yields

$$\log(1/C) = -k\,\pi^2 + k'\,\pi\,\pi_o - k''\,\pi_o^2 + \rho\sigma + k''' \quad . \tag{1-37}$$

Eq. (1-37) relates the hydrophobicity of a pharmaceutical compound, as expressed by the partition coefficient, with its biochemical action.

In summary, the importance of hydrophobicity as a parameter for drug design is due to its relation to steps in the pathway between the administration of a drug and its biological end point. First, hydrophobicity must be taken into account during the drug transport process. Second, hydrophobicity is related to the entropic change that accompanies the interaction between a drug and its receptor, which in most cases is a dehydration process (desolvation, in general). Finally, hydrophobicity plays a role in the quantitative estimation of the interaction energy between a drug and its receptor.

1.4.3 Molecular Lipophilicity in Advanced Drug Design Approaches

Recent strategies of drug design can be divided into two types: direct and indirect. A "direct" strategy can be used if the three-dimensional (3D) structure of the binding sites is known, allowing explicit characterization of ligand-receptor interactions. This is the case when designing many enzyme inhibitors from X-ray data of enzyme-ligand complexes. However, the 3D structure of many receptor sites is still unknown. Here, the clues for the design of new ligands are more "indirect." The strategy is then based on the analysis of the molecular properties of compounds known to have some interaction with the receptor, resulting in diverse pharmacological activities.

Loew *et al.* [1993] has reviewed some new developments in drug design. Presently, researchers recognize two qualitatively different "indirect" approaches. One is the so-called two-dimensional quantitative structure-activity relationship (2D-QSAR). The 2D-QSAR approach is represented by a relation similar to eq. (1-34):

$$\log A_b = c_h f_{hydr}(X_h) + c_e f_{elec}(X_e) + c_{st} f_{st}(X_s) + \text{constant}, \qquad (1\text{-}38)$$

where A_b is proportional to either the receptor binding affinity or a specific biological activity. Each of the terms in eq. (1-38) refers to properties that can affect either receptor recognition or activation. Typically, hydrophobic (f_{hydr}), electronic (f_{elec}), and steric (f_{st}) properties of the ligands are used. Each term is a linear or quadratic function of a corresponding physicochemical parameter X. Common X parameters are: (i) the *n*-octanol-water partition coefficient for a hydrophobic term, (ii) Taft E_S parameter for steric effect [Loew *et al.* 1993, Hansch and Fujita 1964], (iii) Hammett constants to describe electronic effects, and (iv) the molar refractivity to account for dispersion forces. The $\rho-\sigma-\pi$ analysis can be seen as an example of the 2D-QSAR approach.

The second approach is the so-called three-dimensional QSAR (3D-QSAR). This approach builds a 3D model of the receptor cavity. It should be noted that biochemical data are hard to rationalize even after the structure of the complex is known. The main

source of uncertainty is the effect of the solvent and the various entropic contributions. In order to overcome this difficulty, indirect 3D-QSAR techniques are useful. The *distance-geometry-based* 3D-QSAR method of Ghose and Crippen [1985 a,b] is an example of such an approach. Here, the receptor cavity is divided into smaller regions or pockets in order to study their interactions. The interaction energy may be modeled as a function of one or more physicochemical atomic properties. Hydrophobicity is one of these parameters. I will discuss these properties in more detail in Chapter 2.

Kellogg and Abraham [1992] have introduced a related drug design strategy, implemented in the program HINT (Hydrophobic INTeractions). The program uses a complementary hydrophobicity (or "hydropathicity") map between receptor and ligand and is based on three main subroutines. The KEY routine can use the receptor structure to model the "hydropathic" profile of an *ideal* substrate for the receptor. The LOCK routine can perform the complementary analysis, *i.e.*, use the substrate structure to model the hydrophobic character of the receptor. Finally, the LOCKSMITH subroutine highlights the common hydropathic features from a set of tentative ligands, and compares them to both the ideal substrate and the modelled receptor. Again, a quantitative measure of hydrophobicity is the key factor throughout this drug design strategy.

Another 3D-QSAR technique is the Comparative Molecular Field Analysis (CoMFA). This approach assumes that steric and electrostatic forces determine the nature of the ligand-receptor interactions [Cramer III *et al.* 1988, Klebe *et al.* 1994]. In this case, the molecular electrostatic potential is the key property in the design of the pharmacophore. This method requires a set of analogues with optimized overlap, in order to estimate the "shape" of the receptor cavity. Around the set of overlapped molecules, a cubic grid is constructed. The electric field that each molecule would exert upon a probe charge placed at a lattice point is calculated. The value of the electrostatic potential at each of those grid points is then used in a partial least-squares (PLS) [Geladi and Kowalski 1986a,b] regression analysis. In this fashion, one can extract stable QSARs from a severely overdetermined system:

$$R_{ij} = \Sigma_k C_{ik} V_{kj} \ , \qquad (1\text{-}39)$$

where R_{ij} is the bioactivity of type "i" for the j-th compound, C_{ik} is a "sensitivity" coefficient of bioactivity i at grid point k, and V_{kj} is the electrostatic potential at grid point k produced by compound j. [That is, *i* is the index for bioactivity measures, *j* is the index of compounds, and *k* is the index of grid points.] Since there are many more grid points than number of compounds, the system is overdetermined. The PLS statistics permits the determination of a linear expression (3D-QSAR) which has the minimal set of lattice points that reproduces the measured activity of the set of compounds.

Actually, in this last 3D-QSAR drug design technique there is no factor accounting for molecular lipophilicity. I believe this is an important drawback. In Chapter 6, I outline briefly how this approach can be generalized to include hydrophobic, electronic, and steric properties of the ligands.

1.5 Experimental Methods for Determining log *P*

Flask-shaking (FS) is the oldest method for determining partition coefficients. An extensive database of partition coefficients has been created by simply shaking a solute with two immiscible solvents and then analyzing the solute concentration in one or both phases [Cohn and Edsal 1943]. The method is still much in use, although it has several disadvantages, such as being tedious, time-consuming, and requiring very pure compounds and large sample sizes. Other fundamentally different techniques are also used, *e.g.*, high-performance liquid chromatography (HPLC) and micellar reversed-phase liquid chromatography (RPLC).

Since later in the Thesis I will present theoretical correlations with experimental log*P* data, I think it appropriate to review briefly here the techniques for the determination of partition coefficients. More details can be found in the literature, *e.g.*,

Leo *et al.* [1971], Tomlinson [1975], Kallszan [1984], Braumann [1986], Khaledi *et al.* [1989] and Henczi *et al.* [1994].

1.5.1 Flask-Shaking Method and Determination of Phase Concentrations

Occasionally, the ratio of *solubilities* in two separate solvents is reported as a partition coefficient [*e.g.*, Worth and Reid 1916]. This ratio is a value of P in the particular case of saturation. Under the conditions of low solute concentration and with the two solvent phases mutually saturated, the value obtained will be different. The amount of water dissolving in many solvents can be quite high and this modifies their character considerably. Rather high concentrations of organic solutes are necessary to saturate many solvents. Not only does this make for greater solute-solute interactions, but such high concentrations actually change the character of the organic phase so that one is no longer dealing with, say, octanol as the organic phase but rather with some mixed solvent.

There has been some discussion in the literature regarding how to establish the equilibrium between phases. In general, it is estimated that 5 minutes shaking at 20 shakes/min should be sufficient for most substances [Hansch 1969, Leo *et al.* 1971]. Very vigorous shaking should be avoided since this tends to produce emulsions.

In measuring about 800 partition coefficients between water and octanol, Leo *et al.* [1971] usually analyzed the solute in only one phase and obtained the concentration in the other by difference. However, this is not always true. If there is a possibility that absorption to glass may occur, both phases should be analyzed. Such absorption has been found to occur with ionic solutes [Fogh *et al.* 1954]. Absorption may also be a serious problem when using low concentrations of isotopically-labeled compounds ($<10^{-6}$ M).

The volume of solvent used plays a role in the accuracy of the log P determination. For example, if a solute has a P value of 200 (a very lipophilic substance), and 20 mg of it were partitioned between equal 100-ml volumes, the aqueous phase

would end up with only 0.1 mg. If the analytical procedure has an inherent error of 0.05 mg/100 ml, the *P* value could vary between 133 and 400. If, however, 200 ml of water and 5 ml of nonpolar solvent were used, the water layer would contain 3.5 mg and the estimation of *P* would improve to 200 ± 6. With good analytical procedures and proper choices of solvent volumes, log *P* values in the range of -5 to +5 can be measured with accuracy.

Many partitioning systems show a temperature dependence of about 0.01 log unit/deg in the room temperature range. Temperature control is essential for systems with large immiscibility. For most applications, especially biological structure-activity relationships, variations due to temperature are smaller than those caused by other factors. For this reason, most partition coefficient tables are simply characterized as "at room temperature," without any precise statement of what that might be.

Other methodologies to determine log *P* from basically the same experimental set up can also be included here:

(i) An efficient method which employs automatic titration for the determination of partition coefficients of organic bases between immiscible solvents has been described [Brandstrom 1963].

(ii) It has also been shown how a partition coefficient can be calculated from the difference between surface and interfacial tensions, but the accuracy is probably no better than an order of magnitude [Crook *et al.* 1965].

1.5.2 Micellar Reversed-Phase Liquid Chromatography

Chromatographic techniques offer several advantages in measuring physicochemical properties of solutes since they provide good accuracy, require small samples, and can easily deal with low purity compounds [Tomlinson 1975, Kallszan 1984, Braumann 1986, Khaledi *et al.* 1989, Henczi *et al.* 1994].

Weaver and co-workers have discussed a technique that combines micro shake flask with high-performance liquid chromatography [Henczi *et al.* 1994]. The method has been used to determine P_{ow} partition coefficients for a series of anticonvulsants.

Many attempts have been made to establish a correlation between the partition coefficient in octanol-water (log P_{ow}) and the retention factor in reversed-phase liquid chromatography (RPLC) with hydro-organic mobile phases. The assumption is that the extent of retention reflects the hydrophobicity of a solute. However, the correlations between log P_{ow} and RPLC retention factors are limited to very similar compounds. The addition of an organic solvent to the aqueous mobile phase in RPLC is often necessary for very hydrophobic compounds. However, high concentrations of organic solvents lead to inaccurate estimations of hydrophobicity.

The use of surfactant solutions above the critical micelle concentrations as mobile phases in liquid chromatography [LC] has attracted much attention in the past few years [Khaledi 1987, 1988; Dorsey 1987]. Micellar reversed-phase liquid chromatography has unique characteristics that can be advantageous in quantifying hydrophobicity of bioactive molecules in QSAR studies.

Micelles have long been known as simple chemical models for biomembranes [Fendler 1984]. The use of pure "bulk" solvents for modeling complex systems such as biomembranes has been criticized by many authors. It has been demonstrated that the partitioning of solutes in micelles closely resembles that of lipid bilayers and that both of these are different from the two-phase octanol-water system [Miller *et al.* 1977, Treiner 1986a,b, Diamond and Katz 1974]. Both micelles and biomembranes are amphiphilic and anisotropic. Molecular size and shape are significant factors in the partitioning of solutes in anisotropic environments, whereas they play a minor role in isotropic media (*e.g.*, *n*-octanol). Khaledi and Breyer [1989] have given interesting examples confirming the suitability of micelles for representing biomembranes as far as hydrophobic interactions are concerned. The use of micelles as biomembranes in QSAR studies, however, has received

much less attention. Perhaps the difficulties associated with measuring micelle-water partition coefficients by conventional methods have been the major obstacle in conducting QSAR research using micellar systems.

Another important aspect of micellar RPLC is the possibility of calculating micelle-water partition coefficients, P_{mw}, through an equation such as [Armstrong 1985]:

$$1/k' = C_m(P_{mw} - 1)V/P_{sw}\phi + 1/P_{sw}\phi \ , \qquad (1\text{-}40)$$

where k' is the retention factor, V is the partial molar volume of water, C_m is the micelle concentration (*i.e.*, total surfactant concentration minus critical micelle concentration then divided by mean aggregation number), P_{sw} is the solute partition coefficient between water and the stationary phase, and ϕ is the chromatographic phase ratio. A plot of $1/k'$ vs C_m is linear. The P_{mw} value can be calculated from the ratio of slope/intercept.

When compared to conventional chromatographic techniques, two-phase solvent systems like octanol-water still have some advantages. For example, the latter provides a continuous scale for measuring hydrophobicity, while retention data are unique to a given stationary phase/eluent system. In contrast, partition coefficients measured in water/micelle systems also provide a single and continuous scale.

Another advantage of a two-phase solvent system is the additive nature of its partition coefficients. On the basis of the additivity rules, Hansch and Leo [1979] have derived substituent constants for different functional groups, which allows one to estimate the log P values for new compounds. The additivity rules might also be valid for micelle-water partition coefficients. The results reported by several authors are in favor of such a viewpoint [Khaledi 1988]. However, the additivity properties for log P_{mw} should be further verified experimentally.

The reciprocal of the intercept in eq. (1-40) is the retention factor at zero micelle concentration, $k'_0 = P_{sw}\phi$. This parameter may also be useful in hydrophobicity measurements. That is, in addition to the retention factor k', two other parameters of solutes (P_{mw} and k'_0) can be obtained chromatographically.

In a micellar RPLC system, the stationary phase is modified with a constant amount of ionic surfactant and as a result, the stationary phase in micellar RPLC has both amphiphilic and anisotropic properties. It is important to note that the composition (and perhaps the conformation) of the stationary phase in an ionic micellar RPLC system is independent of the micelle concentration in the mobile phase. In other words, solutes would experience the same stationary phase environments at all micellar mobile phase compositions. Finally, the ability of micellar RPLC to deal with both ionic and neutral compounds may also be advantageous in measuring the hydrophobicity of electrolytes.

Despite the existence of certain differences in partitioning behavior in micelles as compared to that in octanol, there is a correlation between micelle-water and octanol-water partition coefficients (log P_{mw} *vs* log P_{ow}) within a group of compounds with a similar partitioning behaviour. Satisfactory correlation between the RPLC retention factor and log P_{ow} can be achieved by adjusting the lipophilic-hydrophilic balance of the chromatographic system to mimic octanol-water environments. Micelles allow some degree of control over specific interactions and hydrophobic "force" by a selection of surfactant and solvent additives. Several authors have reported P_{ow}–P_{ow} correlations. Treiner and *et al.* [1986] reported correlations for 20 polar aliphatic compounds and Gago *et al.* [1987] reported correlations between log k' in different micellar mobile phases with log P_{ow} for 11 monosubstituted benzenes.

In RPLC with hydro-organic mobile phases, the relationship between the retention factor and log P_{ow} is often expressed in the logarithmic form as

$$\log k' = a \log P_{ow} + b \quad . \tag{1-41}$$

This is a special case of the Collander equation [Hansch 1971]. This equation predicts linear relationships between the logarithm of the partition coefficients measured in two different partitioning systems (P_1 and P_2), provided that solute-solvent interactions are similar in the two systems (*i.e.*, $\log P_1 = p \log P_2 + q$).

The correlation between retention in micellar RPLC and $\log P_{ow}$ depends upon the type of solute, mobile, and stationary phases. For the micellar eluents, Khaledi *et al.* [1989] observed a better linear relationship between k', not $\log k'$, and $\log P_{ow}$:

$$k' = a' \log P_{ow} + b' \quad . \tag{1-42}$$

1.6 NMR in the Study of Lipophilicity

As mentioned in section 1.1, molecular lipophilicity is a phenomenon on the molecular level, and depends on the local condition of a molecular surface. Elucidation of the behavior requires answers to the following questions: How much water is significantly perturbed from pure water behavior and how does this amount depend on the molecular structure in terms of the charged, polar, and hydrophobic surface character, as well as the presence of associated counterions? What is the shape of the orientational probability distribution of the perturbed water with respect to the surface? What is the rate of water reorientation? How long does an average water molecule spend in the perturbed region before it diffuses away into the bulk? The experimental technique that is best suited to provide the answers to these questions is nuclear magnetic resonance (NMR) [Lee 1994].

1.6.1 Dynamic NMR

The relaxation of a collection of nuclear (or electron) spins processed in an external magnetic field will occur because of fluctuations in the local field at the species observed. For some important cases of interest, fluctuating fields are due to atomic and molecular motions which modulate the magnetic interactions operating on the spin:

anisotropic interaction with the external magnetic field, hyperfine or dipolar interaction with the external magnetic field, hyperfine or dipolar interactions with other magnetic species, and possibly other effects (zero-field splitting, spin-rotation interactions, quadrupolar, *etc.*). The relaxation time of the various field-modulating mechanisms for inducing NMR relaxation is 10^6-10^8 sec^{-1} and for ESR relaxation is 10^9-10^{11} sec^{-1}.

Structural information can be provided by NMR in a variety of ways. The parameters of NMR measured for this purpose are classified: frequency shifts (chemical shifts), changes of line intensities, and coupling constants. In a further development, time-dependence of NMR spectra through line-shape analysis of the frequency-domain signal or measurement of decay of the time-domain signal was used to obtain dynamic information at the molecular level. A static NMR description of a flexible biological molecule, which means no time parameter is enclosed, is insufficient to fully account for their physical or chemical properties due to the presence of a variety of dynamic processes. In this respect, time is often considered as the fourth dimension in NMR structure determination. The term 'stereodynamics' has been used to emphasize this necessary overlapping of structural and dynamical information to describe mobile systems.

Progress with improved resolution and sensitivity now allows NMR to study systems of higher complexity, such as synthetic polymers or biological molecules. Large molecules are intrinsically mobile and the knowledge of internal motions is necessary for an accurate description of three dimensional structures [Williams 1989]. Physico-chemical properties of their solutions are often monitored by motion at the molecular level. This is particularly true for biological molecules in which internal motions often control biological function, such as a channel opening for ion transport, fiber contraction, and the surface activity of proteins.

NMR has proved to be an invaluable tool for learning about the dynamical information of microheterogeneous systems [Lindman *et al.* 1994]. Typical examples of

such investigations are systems of higher complexity, such as synthetic or natural polymers, where the major contribution to dynamics arises from internal motions. Other systems of high complexity are organized assemblies of amphiphilic molecules found in natural surfactant phases with respect to disordered dispersions [Helfrich 1978, De Gennes and Taupin 1982].

1.6.2 Water Oxygen-17 Magnetic Relaxation

Water oxygen-17 magnetic relaxation has been used by several authors [Halle *et al.* 1981] to study protein hydration. Compared to proton and deuteron magnetic relaxation, which have been used extensively in protein hydration studies [Koenig *et al.* 1975, Hallenga and Koenig 1976], ^{17}O relaxation has at least four important advantages:

(i) The strong quadrupolar (J=5/2) interaction leads to large relaxation effects, thus permitting studies at reasonably low protein concentrations;

(ii) the intramolecular origin of the electric field gradient at the water oxygen nucleus makes the quadrupolar interaction virtually independent of the molecular environment, greatly facilitating the interpretation of relaxation data;

(iii) except for a narrow pH range around neutral, the ^{17}O relaxation is not influenced by proton (deuteron) exchange with prototropic residues on the protein, which is a serious problem in ^{1}H and ^{2}D relaxation, but can only be affected by the exchange of entire water molecules;

(iv) Cross-relaxation, which contributes significantly to ^{1}H relaxation, is unimportant for ^{17}O.

The relaxation theory for the oxygen-17 nucleus is complicated by the large spin quantum number 5/2. If the molecular motion causing quadrupolar relaxation has components with correlation times of the order of the inverse resonance frequency $1/\omega_0$ or longer, i.e., if in its spectral density $J(0)\neq J(\omega_0)$ (so called "nonextreme narrowing" conditions), then the relaxation must be described as a sum of three decaying exponentials [Hubbard 1970]. If the ^{17}O nucleus exchanges between environments with different intrinsic relaxation rates, even more exponentials are needed to describe the

decaying nuclear magnetization. For the important case of fast exchange, i.e., when the exchange rates exceed the intrinsic relaxation rates, the relaxation matrix **R** may be decomposed according to eq. (1-43), where the sum runs over all environments ('states') S, and P_S is the fraction of nuclei in state S,

$$\mathbf{R} = \sum_S P_S \mathbf{R}_S \,. \tag{1-43}$$

For spin 5/2, it is impossible to obtain general analytical expressions for the decay of the longitudinal and transverse magnetization, but numerical computations for a two-state model with one state (bulk water in this case) under "extreme narrowing" conditions show that the longitudinal magnetization decays as a single exponential in all cases of practical interest, while the transverse magnetization, under similar conditions, decays as a sum of three exponentials. However, since the preexponential factors depend on the distribution of nuclei over different sites as well as on the corresponding correlation times, the transverse magnetization will also decay exponentially for $P_S \leq 0.1$ and $\tau_c \leq 50$ ns. In fact, in common experimental conditions, the dominating exponential always exceeds 0.99 relative amplitude [Halle *et al.* 1981].

For a fast exchange two-state model the excess relaxation rate can be written as

$$R_{i,ex} = R_{i,obsd} - R_{ref} = P_{PR}(R_{i,PR} - R_{ref}), \quad (i=1, 2), \tag{1-44}$$

where $R_{i,ex}$ is the observed relaxation rate in a protein solution, R_{ref} is the relaxation rate in pure water of the same temperature, and the mole fraction P_{PR} and relaxation rate refer to those water molecules that interact detectably with the protein, *i.e.*, the "hydration water" in the ^{17}O relaxation sense.

Several NMR investigations on aqueous solutions of organic compounds have indicated that the local solvent viscosity in the neighborhood of alkyl groups is

significantly higher than the bulk viscosity of water [Herz 1973, Chan *et al.* 1979 and Howarth 1975].

1.6.3 Techniques other than NMR

Besides the NMR relaxation time measurements, there are several other experimental methods that may be used to get information on the difficult problem of the description of molecular motions in a solution [Canet and Robert 1994]. The methods most often used for the analysis of molecular motions are Infrared Absorption, Raman and Rayleigh Scattering, Coherent and Incoherent Neutron Scattering, Dielectric and Kerr Relaxation, and Fluorescence Depolarization. All these methods are related to a particular correlation function and spectral density, which characterizes the time evolution of a give parameter of the molecule. For example, in the infrared absorption method, an algebraic expression is established which exists between the line intensity shape, expressed as a function of the frequency, namely $I(\omega)$, and the molecular electric dipole moment correlation function [Gordon 1968].

Conventional NMR requires moderately high concentrations ($\geq 10^{-2}$ *M*), so only solutes containing polar functional groups can be studied. ESR methods, although workable at very low concentrantions ($\geq 10^{-5}$ *M*), require unpaired electron, which can be found only in polar molecules or ions. An accurate or realistic description of the motion of a molecule in a solution can only be reached by considering more than one experimental approach. Each of these methods has its advantages and disadvantages, and a detailed description can be found in references [Steele 1976, Williams 1978, Rotschild 1984, Madden and Kivelson 1984, Wei and Patey 1989]. However, it has been pointed out by many authors that NMR is the best experimental tool in the study of liquid structure in a solution system [Lee 1994, Canet and Robert 1994].

1.7 Measurements of Interaction Forces between Surfaces

A related and more theoretical concept to the solvation free energy is hydrophobic interaction (HI) [Ben-Naim 1980]. The definition of pairwise HI can be given by the follow equation,

$$\Delta G(\infty \rightarrow \sigma) = G(T, P, \text{solvent}, R_{12}=\sigma) - G(T, P, \text{solvent}, R_{12}=\infty). \quad (1\text{-}45)$$

Eq. (1-45) is the difference of free energy of two solute molecules at a close distance σ and at infinite separation. In the study of HI, people want to know the behavior of $\Delta G(R)$ as a function of distance R,

$$\Delta G(R) = U_{SS}(R) + \delta G^{HI}(R), \quad (1\text{-}46)$$

where $U_{SS}(R)$ is the direct solute—solute interaction, and $\delta G^{HI}(R)$ is the contribution from solvent.

1.7.1. Direct Measurements of Intermolecular and Surface Forces

The study of force laws is needed to measure the forces between molecules or particles as a function of distance. Once the force F as a function of distance D is known for the two surfaces (of radius R), the force between any other curved surfaces simply scales by R. Furthermore, the adhesion or interfacial energy E per unit area between two flat surfaces is simply related to F by the Derjaguin approximation [Derjaguin 1934, Israelachvili 1992]:

$$E = F/2\pi R \quad (2\text{-}47)$$

Three techniques that can directly measure the force laws between two bodies of macroscopic, colloidal and atomic dimesions, respectively, are Surfaces Forces Apparatus (SFA), Total Internal Reflection Microscopy (TIRM), and Atomic Force

Microscope (AFM) [Israelachvili 1992]. Table 1-1 makes a comparison of these three techniques.

Table 1-1 Comparison of Three Techniques of Force-Measurements

Technique	Application	Distance resolution	Sensitivity
SFA	Two surfaces	0.1 nm	10^{-8} N
TIRM	Colloidal Particles	~ 10 μm (diameter)	10^{-15} N
AFM	tip to surface	> 0.1 nm	10^{-9}-10^{-10} N

1.7.2. Applications of Direct Measurements of Forces

The scope of phenomena that can be studied using the SFA technique includes measurements of dynamic interactions and time-dependent effects, for example, the viscosity of liquids in very thin films [Chan and Horn 1985, Israelachvili 1986, 1989], shear and frictional forces [Israelachvili *et al.* 1985], and the fusion of lipid bilayers [Helm *et al.* 1989]. In the TIRM method, by analyzing how the reflected intensity of the light varies with time, one can thus determine the distances sampled. From this the force law can be obtained over a reasonably large range of distances on either side of the equilibrium. The TIRM technique promises to provide reliable data on a variety of interparticle interactions under conditions that closely parallel those occurring in colloidal systems. AFM is very similar to SFA. Interpreting the results of an AFM experiment is not always straightforward, because the absolute distance between the surfaces is not usually known exactly, and neither is the tip geometry. In addition, the fine tips and the surfaces are often elastic or plastic during a measurement, further complicating the interpretation of the results. However, the technology is developing rapidly so that very soon we may expect to see reliable intermolecular force laws emerging from AFM measurements [Isaerlachvili 1992].

Chapter 2: Review of Research of Molecular Lipophilicity

As discussed in Chapter 1, lipophilicity (or hydrophobicity) is a molecular property related to entropic effects caused by changes in the organization of water molecules around the solute molecule. Experimentally, the *n*-octanol/water partition coefficient is used as an overall measure of molecular lipophilicity. This phase equilibrium constant is related to the free energy $\Delta\overline{G}^{\circ}_{tr}$ of transfer of the solute from the water to the organic phase, $\log P_{ow}=-\Delta\overline{G}^{\circ}_{tr}/2.3026RT$ (cf. eqs. (1-29) and (1-30), the latter written for the $\Delta\overline{G}^{\circ}_{tr}$ of transfer in the opposite direction). In turn, $\Delta\overline{G}^{\circ}_{tr}$ is the difference of standard solvation free energies for the solute in each phase (eq. (1-30)). Therefore, the difference between solvation free energies is a natural thermodynamical measure of molecular lipophilicity.

Recent developments in computational methods [Carrupt *et al.* 1997] provide various strategies for evaluating $\Delta\overline{G}^{\circ}_{tr}$ values. Computer simulation has become an important tool in studying the behavior of complex biological systems, such as solvated proteins, nucleic acids, and protein folds. Simulation models, properly calibrated with available experimental information, can provide insights on structure and dynamics that may not be directly measurable. Many biochemically interesting systems can now be studied in more detail using these techniques. In this chapter, I will review some of the latest developments of computational methods for evaluating $\log P_{ow}$ and solvation free energy, including structure-based empirical correlations, quantum mechanical calculations, molecular dynamics simulations, Monte Carlo simulations, and two combined approaches.

2.1 Empirical Estimations of Partition Coefficients Derived from Molecular Structure

There are several promising theoretical methods for computing free energy differences (and, thus, related quantities such as partition coefficients and equilibrium constants). Among these techniques, I can include quantum-mechanical SCRF (self-

consistent reaction field) approaches [Miertuš *et al.* 1981, Miertuš and Tomasi 1982, Bonaccorsi *et al.* 1984], the thermodynamic perturbation method, molecular dynamics simulations [McCammon and Harvey 1987, van Gunsteren and Berendsen 1990], and Monte Carlo simulations [Northrup and McCammon 1980, Friesner and Levy 1984, Heermann 1990]. Nevertheless, so far no theoretical technique, based on the first principles, is advanced enough to characterize molecular hydrophobicity for realistic systems with good accuracy.

Since the rigorously theoretical calculation of hydrophobicity is difficult, many empirical and semi-empirical methods for estimating partition coefficients have been proposed. There are two different approaches in this category: empirical correlations and fragment-addition formulas. The first approach correlates partition coefficients with various experimental or theoretical parameters, such as molecular surface, volume, mass, atomic charges and electrostatic potential. The second approach considers the partition coefficient as the sum of contributions of molecular fragments or "atoms." Empirical methods can often produce very good results for homologous series of compounds.

2.1.1 Estimations of log*P* Based on Molecular Surface Information

Long ago, Némethy and Scheraga [1962] pointed out that the dominant energy source for hydrophobic behavior is the regularity of the cluster of water molecules in contact with the hydrophobic surface of the solute molecule. Later, Watanabe and Mitsui [1981] suggested that molecular hydrophobicity might be estimated from the solvent-accessible surface area (*s*).

The solvent-accessible surface (*s*) was originally defined by Lee and Richards [1971] as the area traced out by the center of a solvent molecule (assumed to be a sphere) as it is rolled over the molecular surface of the solute. Molecular van der Waals and solvent-accessible surfaces are determined from a set of atomic radii and a solvent radius [see Fig. 2-1]. The solvent-accessible surface defined with a solvent sphere of zero radius is, of course, the original van der Waals surface.

Iwase *et al.* [1985] proposed a method for estimating log P_{ow} based on the total area s of a molecule. The chosen solvent radius of water was 1.4 Å, after considering the water molecule effectively as a sphere. Two correlations were proposed by these authors. One includes the surface area of exposed hydrogen atoms, whereas another excludes them. Standard quantum chemical programs were used to optimize the molecular geometries and calculate the molecular surface area. In order to compare later with the work in this thesis (Chapter 3), it is useful to briefly discuss their results here.

For aliphatic hydrocarbons, Iwase *et al.* [1985] obtained:

$$\log P_{ow} = (2.05 \pm 0.18)\, s^* - (0.45 \pm 0.15)\, m - (1.29 \pm 0.35), \qquad (2\text{-}1)$$
$$n = 9,\ r = 0.995,\ \mathrm{s} = 0.09,$$

where s^* is the surface area excluding hydrogen atoms, m is the dipole moment, n is the number of hydrocarbons used in the correlation, r is the correlation coefficient, and s is the standard (statistical) error. A correlation of similar quality exists for aromatic hydrocarbons:

$$\log P_{ow} = (1.94 \pm 0.14)\, s^* - (1.92 \pm 0.41), \qquad (2\text{-}2)$$
$$n = 12,\ r = 0.995,\ \mathrm{s} = 0.09.$$

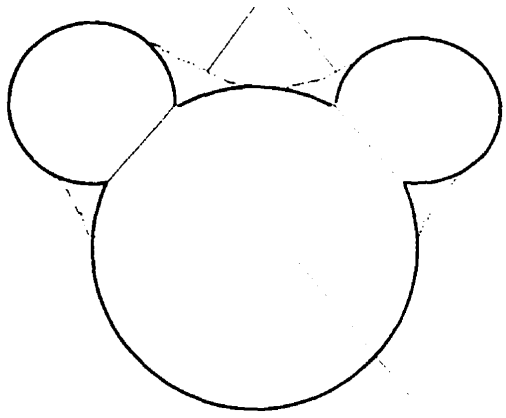


Figure 2-1: Solvent accessible surface and van der Waals surface.

For solubility in water, C (in molar concentration), for a series of 156 miscellaneous organic liquids, Iwase *et al.* [1985] have obtained a correlation equation:

$$\log(1/C) = (2.23 \pm 0.11)\, s^* - (1.31 \pm 0.05)\, s_H - (1.80 \pm 0.27), \qquad (2\text{-}3)$$
$$n = 156,\ r = 0.981,\ s = 0.26,$$

where s^* is the molecular surface without the contribution of hydrogen and s_H is the exposed surface of the hydrogen atoms in each molecule. In the correlation eqs. (2-1), (2-2) and (2-3), the areas are expressed in $100\times\text{Å}^2(=\text{nm}^2)$.

For a homologous series of aliphatic hydrocarbons and aromatic hydrocarbons, the above formulas indicate a good correlation with the surface area. However, it is clear that log$P$ of compounds with other functional groups cannot be determined based only on the surface areas. Additional structural properties should be included in the empirical formulas, as shown below.

2.1.2 Empirical Formulas Based on Atomic Charge, Surface Area, and Dipole Moment

Kantola *et al.* [1991] have proposed an atom-based structural method for the calculations of the hydrophobicity index H (proportional to log P) of molecules. The parameters used in this method are molecular surfaces and atomic charges, both of which have some dependency on the conformation adopted by the system, as well as a set of adjustable parameters that depend only on the atomic number. The four formulas they presented are as follows,

$$H = \Sigma_i [\alpha_i(N)\, s_i + \beta_i(N)\, s_i\, (\Delta q_i)^2] , \tag{2-4}$$

$$H = m + \Sigma_i [\alpha_i(N)\, s_i + \beta_i(N)\, s_i\, (\Delta q_i)^2] , \tag{2-5}$$

$$H = \Sigma_i [\alpha_i(N)\, s_i + \beta_i(N)\, s_i\, (\Delta q_i)^2 + \gamma_i(N)\, q_i] , \tag{2-6}$$

$$H = m + \Sigma_i [\alpha_i(N)\, s_i + \beta_i(N)\, s_i\, (\Delta q_i)^2 + \gamma_i(N)\, q_i] , \tag{2-7}$$

where α_i, β_i, and γ_i are atomic parameters determined by a linear regression with experimental partition coefficients. The structural atomic information used in the above formulas is: q_i, the atomic total charge on atom "i" (measured in electron units), Δq_i, the atomic net charge (*i.e.*, the difference between the number of electrons in the neutral atom and q_i) and, s_i, the contribution of atom "i" to the molecular surface area (in $Å^2$). The three terms in eq. (2-4) -(2-7) are interpreted as different contributions to hydrophobicity:

1. The term $\alpha_i(N)\, s_i$ should describe the energy required to form a cavity in the solvent. Since water has a larger internal pressure compared to any other solvent, an increase in solute size favors a larger solubility in the organic phase.
2. The term $\beta_i(N)\, s_i\, (\Delta q_i)^2$ describes the contribution to hydrophobicity that arises from the presence of a polar group.
3. The $\gamma_i(N)\, q_i$ relates to the effect of molecular polarizability.

It should be pointed out that all atomic quantities used in this approach, q_i, Δq_i, and s_i, are obtained from quantum chemical calculations based on molecular geometries.

No experimental parameters are used. There is merit to this approach, because experimental data are often unavailable for imagined molecules in drug design.

Kamlet and co-workers [Kamlet *et al.* 1981, 1988, 1986; Kamlet and Taft 1985, Taft *et al.* 1985, Abraham *et al.* 1986] have used a large number of parameters to derive correlations with octanol/water partition coefficients, P_{ow}:

$$\log P_{ow} = mV_I + s(p^* + dd) + bb_m + aa_m + k \ , \qquad (2\text{-}8)$$

where V_I is the van der Waals volume of the solute molecule, p^* is the polarizability, d is a "polarizability correction", a_m and b_m represent hydrogen-bond donating and accepting tendencies, respectively, and k is a constant. For 245 organic molecules of different types, eq. (2-8) gave a correlation coefficient r=0.996 and a standard error s=0.131. It should be noted that certain compounds led to deviations and the authors excluded them from the correlations. Troublesome compounds were pyridine and its derivatives, primary and secondary amines, and nitroalkanes. Later, Abraham [1993] has shown that these deficiencies can apparently be corrected if an alternative set of structural parameters is used. Eq. (2-8) is the best equation for the calculation of partition coefficients so far. However, it is of little use for drug design, because in drug design, people deal with designed molecules, or imaginary molecules. All experimental parameters are not available. People want to know the properties of imaginary compounds based only on their molecular structures, and this is the job of molecular modeling.

In summary, the approaches in this section are an improvement over the methods discussed in §2.1.1 based on the entire molecular surface. Besides atomic surface area s_i, more structural parameters can be included, such as the atomic total charge q_i, atomic net charge Δq_i, and the molecular dipole moment m. Other molecular properties have been used in linear and nonlinear correlations for log$P$. The literature on these types of approaches is vast and cannot be covered here. Some significant contributions are reported by Klopman and Iroff [1981], Klopman *et al.* [1985], Pearlman [1986], Bodor *et*

al. [1989], Bodor and Huang [1992], Viswanadhan *et al.* [1993]. In this thesis, I will focus mostly on correlations with structural parameters *derived from molecular electrostatic properties*. These parameters have a very clear physical meaning in terms of reactivity. I deal with these properties in §2.1.4.

2.1.3 Empirical Formulas for Hydration Free Energy

There are other molecular surface-based methods for estimating hydrophobicity. Some of them calculate the free energy of hydration directly. Ooi *et al.* [1987] described a method for estimating the effects of hydration on conformational energies of polypeptides. The free energy of hydration is composed of additive contributions from various functional groups. Ooi *et al.* [1987] assume that the extent of the interaction of any functional group "i" of a solute with the solvent is proportional to the solvent-accessible surface area s_i of group i. The reason is simple: the group can interact directly only with the water molecules that are in contact with its exposed surface. Thus, the total free energy of hydration of a solute molecule is given by

$$\Delta G^\circ_H = \sum_i g_i s_i \,, \tag{2-9}$$

where the summation extends over all atomic fragments of the solute, and s_i is the conformation-dependent accessible surface area of group i. The proportionality constant g_i represents the contribution to the free energy of the hydration of group i per unit accessible area. These constants have been evaluated for seven functional groups occurring in peptides, by least-squares regression with experimental free energies of a solution of small aliphatic and aromatic molecules with various functional groups. The calculation involves an important approximation: the surfaces $\{s_i\}$ are assumed constant over small conformational changes in the solution.

Ooi *et al.* [1987] have applied the same approach to modeling the enthalpy ΔH°_H and heat capacity of hydration ΔC_p°, assuming also that these properties can be expressed as fragment contributions proportional to the accessible surface area:

$$\Delta H^{\circ}_{H} = \sum_i h_i s_i , \qquad (2\text{-}10)$$

$$\Delta C_p^{\circ} = \sum_i c_i s_i . \qquad (2\text{-}11)$$

The free energy and enthalpy of hydration for the *N*-acetyl-*N'*-methylamides of all 20 naturally occurring amino acids have been computed with this method.

Finally, I should mention that the contributions of the various accessible minimum energy conformations to the hydration free energy have been discussed by Vásquez *et al.* [1983]. A similar approach was developed by Eisenberg and McLachlan [1986] for the estimation of solvation energy in protein folding and binding.

2.1.4 Fragmental Contribution to log *P*

The methods discussed in §2.1.2 and §2.1.3 represent hydrophobicity by a decomposition into fragmental contributions. The additivity of fragmental contributions to log *P* is an old idea. As seen before, it assumes that the total transfer free energy of a molecule is the sum of the contributions from all constitutive fragments. The partition coefficients can then be expressed as:

$$\log P_{ow} = \sum_i n_i a_i , \qquad (2\text{-}12)$$

where n_i is the number of atoms of type *i*, and a_i is the contribution of atomic type *i*. Rekker [1977] and Hansch and Leo [1979] gave the values for some standard chemical groups.

The most general approach to this task is to construct a complete set of chemically significant transferable fragments. Ghose and Crippen [1986] give a very detailed classification of hydrophobic contributions of "atom types," to be used in conjunction with their 3D-QSAR model. The classification is designed to take into account: (i) the electronic density distribution around an atom; (ii) the approachability of the solvent

molecules to the atom; and (iii) the influence of the nearest neighbors bonded to the atom. These factors are thought to account for the enthalpic and entropic contributions to the free energy of solvation.

A total of 494 octanol-water partition coefficients for compounds containing carbon, hydrogen, oxygen, nitrogen, halogens, and sulfur are considered in the work of Ghose and Crippen [1986]. These elements are classified into *90 atomic types*. For carbon alone, there are as many as 51 atom types depending on valency and neighboring atoms. A second set given by Viswanadhan *et al.* [1989] improves on the latter by using the experimental data on 893 compounds to derive the *120 atom type contributions*. (An intermediate set of parameters is discussed by Ghose *et al.* [1988].) These values can be used to test various empirical formulas that use structural information. Ideally, the correlations discussed in §§2.1.2 and 2.1.3 should provide an interpretation to the atomic fragments of Ghose and Crippen [1986].

2.1.5 Estimation of Partition Coefficients Based on Molecular Electrostatic Potential

Over the past 20 years, the molecular electrostatic potential (MEP) has been used extensively as a reliable and quantitative tool for the identification of molecular regions most susceptible to electrophilic and nucleophilic attack [Politzer and Truhlar 1981]. The electrostatic potential provides insight into the general patterns of positive and negative regions that promote or inhibit molecular interactions between drugs and receptors. The MEP is defined as the electrostatic interaction energy between the unperturbed charge distribution of the molecule and a unit positive charge placed at the point **r** in 3D space. Its quantum-mechanical expression in atomic units is [Scrocco and Tomasi 1973]:

$$V(\mathbf{r}) = \sum_{A=1}^{n} \frac{Z_A}{\|\mathbf{R}_A - \mathbf{r}\|} - \sum_{\mu,\nu} p_{\mu\nu} \int \frac{\phi_\mu^*(\mathbf{r}')\phi_\nu(\mathbf{r}')d\mathbf{r}'}{\|\mathbf{r}' - \mathbf{r}\|}, \qquad (2\text{-}13)$$

where Z_A is the charge on nucleus A, located at $\mathbf{R}_A$, $\phi(\mathbf{r})$ is the atomic orbital, and $p_{\mu\nu}$ is the element of density matrix.

Politzer and his research group have made several contributions to the analysis of reactivity in terms of MEP. Extrema of $V(\mathbf{r})$ appear to be particularly useful. As an example, the hydrogen-bond-accepting ability (basicity) of molecules has been shown to be proportional to the value of the MEP minima [Murray and Politzer 1991, 1992].

Murray *et al.* [1994] have developed a quantitative strategy for using the electrostatic potential to analyze molecular interactions in which there are no significant polarizations or charge transfers. These authors proposed a number of statistically-based interaction indices derived from the MEP. The result is a so-called "general interaction properties function" (GIPF) [Murray *et al.*, 1994]. The basic idea is that a property, such as log P, can be expressed in terms of electrostatic and structural parameters:

$$\text{Property} = f(s, \Pi, \sigma^2_{tot}, \nu), \tag{2-14}$$

where s is the surface area of a molecule. Other parameters are defined as:

$$\Pi = \frac{1}{m}\sum_{i=1}^{m} \left|V(\mathbf{r}_i) - \bar{V}\right|, \tag{2-15}$$

$$\sigma_+^2 = \frac{1}{m_+}\sum_{i=1}^{m_+} [V^+(\mathbf{r}_i) - \bar{V}^+]^2,\ V^+(\mathbf{r}_i) > 0\ . \tag{2-16}$$

$$\sigma_-^2 = \frac{1}{m_-}\sum_{i=1}^{m_-} [V^-(\mathbf{r}_i) - \bar{V}^-]^2,\ V^-(\mathbf{r}_i) < 0\ . \tag{2-17}$$

$$\sigma_t^2 = \sigma_+^2 + \sigma_-^2\ , \tag{2-18}$$

$$\nu = \frac{\sigma_+^2\sigma_-^2}{[\sigma_t^2]^2}\ , \tag{2-19}$$

where $V(\mathbf{r}_i)$ is the MEP on the point $\mathbf{r}_i$. The property $\bar{V}$ is the mean MEP over the entire molecular surface, whereas $\bar{V}^+$ and $\bar{V}^-$ are the averages restricted to the regions on the surface where the MEP takes only positive (or zero) and negative values, respectively:

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m} V(\mathbf{r}_i), \tag{2-20}$$

$$\bar{V}^+ = \frac{1}{m_+}\sum_{i=1}^{m_+} V^+(\mathbf{r}_i), V^+(\mathbf{r}_i) > 0 \ , \tag{2-21}$$

$$\bar{V}^- = \frac{1}{m_-}\sum_{i=1}^{m_-} V^-(\mathbf{r}_i), V^-(\mathbf{r}_i) < 0 \ . \tag{2-22}$$

In eqs. (2-15) to (2-22) the molecular surface is represented by a discrete grid of m points. Over m_- points, $V(\mathbf{r}_i)$ takes negative values, whereas the MEP is positive (or zero) over m_+ points ($m = m_+ + m_-$).

The parameter Π (eq. (2-15)) is the average absolute deviation from the mean of the surface electrostatic potential. It is an effective measure of local polarity (or charge separation), which may be quite significant even in a molecule having a zero dipole moment [Brinck *et al.* 1992a]. The total variance, σ^2_t (eq. (2-18)), is a measure of the spread (or dispersion) of the surface potential. The "balance" parameter ν (eq. (2-19)) measures the symmetry of the distribution of positive and negative MEP.

In the GIPF approach, in addition to the parameters s, Π, σ^2_{tot} and ν, several other descriptors of the MEP distribution and their combinations are also used, such as σ^2_+, σ^2_-, $s\Pi$, $s\sigma^2_{tot}$, and $\nu\sigma^2_t$, in order to obtain good results. Politzer and co-workers used their general interaction properties function (GIPF) to analyze correlations of octanol/water and acetonitrile/NaCl-saturated-water partition coefficients, P_{ow} and P_{aw}, for benzene, toluene, and nine nitroaromatic compounds [Murray *et al.*, 1993 c]. (Regarding the technical details, the 11 geometrical structures were optimized at *ab initio* STO-3G level, and the

statistical descriptors Π, σ^2_t, and ν, along with the surface area s, calculated at STO-5G *ab initio* level on molecular surfaces defined by the 0.001 electron/bohr^3 contour of electron density [Francl *et al.* 1984, Bader *et al.* 1987].) Their best correlations with two parameters take the form:

$$\log P_{ow} = \alpha s - \beta s \Pi - \gamma , \qquad \alpha, \beta > 0 , \qquad (2\text{-}23)$$

$$\log P_{aw} = \alpha s - \beta \Pi - \gamma , \qquad \alpha, \beta > 0 , \qquad (2\text{-}24)$$

where s is the molecular surface area. The respective correlation coefficients are 0.980 and 0.971. These equations indicate that an increase in solute size favors partitioning in the organic phase, octanol in the case of eq. (2-23) and acetonitrile in eq. (2-24). In contrast, an increase in Π favors partitioning into water, which is the more polar solvent in either case. Note, nevertheless, that the correlations obtained are not outstanding and leave room for improvement.

Another more extended research of water-octanol partition coefficients performed recently by Brinck *et al.* [1993] included 70 organic molecules of various types and sizes. Four *ad hoc* correlation schemes were explored:

$$\log P_{ow} = a s + b\sigma^2_t + g \quad , \qquad (2\text{-}25)$$

$$\log P_{ow} = a s + b\sigma^2_t + g \Pi + d , \qquad (2\text{-}26)$$

$$\log P_{ow} = a s + b\sigma^2_t + g s \Pi + d , \qquad (2\text{-}27)$$

$$\log P_{ow} = a s + b\sigma^2_- + g s \Pi + d . \qquad (2\text{-}28)$$

Nevertheless, some conclusions can be extracted. The signs and values of coefficients a, b, and g in eq. (2-25) and eq. (2-26) show that partitioning into octanol is favored by a large surface area, while high σ^2_t and Π values correlate with partitioning into water. In eq. (2-27), Π appears multiplied by the surface area. The effect of this is to make the term size dependent. Therefore, eq. (2-27) describes better than eq. (2-26) the fact that the strength of the interaction with bulk water depends on the exposed molecular surface

area. In eq. (2-28) the term σ^2_- gives greater emphasis to the negative portions of the molecular surface, consistent with the conclusions of Famini *et al.* [1992] and Kamlet *et al.* [1988] that the dominating factors in determining P_{ow} are size and hydrogen bond accepting ability. Eq. (2-28) has (marginally) the better correlation coefficient and standard deviation.

The parameters used in Politzer's method are statistical descriptors of a quantum-mechanical MEP on a model molecular surface, however, they are still used in empirical formulas. Du and Arteca [1995 b] presented some novel ideas to extend (and possibly improve) this approach to computing log$P$ by introducing simpler structure-based parameters with a clear physical meaning.

2.2 Quantum-Mechanical Methods to Compute the Solvation Free Energy

Most chemical experiments are done in a solution, whereas traditionally most quantum chemical calculations have been done in the gas phase. The properties of molecules and transition structures in the gas phase or a solution can differ considerably. For example, electrostatic effects are often much less important for species placed in a solvent with a high dielectric constant than they are in the gas phase. For this reason, a number of techniques have been developed in the last 15 years for the quantum-mechanical study of solvated systems. (For on overview on methodologies, see, *e.g.*, Ventura *et al.* [1987] and references therein.) In this chapter, I review some promising developments in this area.

2.2.1 Discrete and Continuum Quantum-Mechanical Models

The standard, discrete quantum-mechanical models are fully capable of describing the basic features of the solute-solvent interaction, including hydrogen bonding, mutual polarization, and charge transfer. They are, however, limited to small solute molecules and a restricted number of solvent molecules. Moreover, obtaining minimum energy configurations for solvent-solute clusters is a difficult and computationally-demanding

task. For this reason, all-atom quantum-mechanical calculations are mostly used for the evaluation of local specific effects, such as strong hydrogen bonding and hydrogen transfer mediated by water molecules. When studying hydration shells, discrete models can explain the interactions between biomolecules and water molecules. However, it is difficult to use these models to extract configurationally-averaged properties of the hydration shell.

Continuum models are more suitable for the description of bulk solvent effects and large solvated systems. The continuum models consider the solvent as an infinite continuous dielectric medium possessing the macroscopic characteristics of the pure liquid (*e.g.*, its dielectric constant and mean polarizability). The solute is placed in a cavity inside the continuum medium, and solute-solvent interactions are treated either classically or quantum-mechanically. The solution process thus consists of inserting a solute molecule into a suitable cavity (spending an amount of "energy of cavitation" for its creation) and then "switching on" the interactions with the surrounding solvent. Schematically, this model is illustrated in Fig. 2-2 (A). The overall change of the Gibbs free energy of solvation, ΔG_{solv}, in the continuum model is evaluated as a sum [Frecer *et al.*, 1991]:

$$\Delta G_{solv} = \Delta G_{elst} + \Delta G_{rep} + \Delta G_{disp} + \Delta G_{cav} + \Delta G_{phas} , \qquad (2\text{-}29)$$

where ΔG_{elst} is the electrostatic contribution to the solvent effect, ΔG_{rep} the repulsion-energy contribution, ΔG_{disp} the dispersion-energy contribution, ΔG_{cav} the cavitation Gibbs free energy, and ΔG_{phas} is related to the configurational entropy changes that occur during the solution process.

In recent years, much effort has been devoted to developing methods for calculating the electrostatic term ΔG_{elst}, which represents the main contribution to ΔG_{solv}. Relatively less attention has been given to the description of the remaining contributions.

2.2.2 Self-Consistent Reaction Field (SCRF) Method for Evaluating ΔG_{elst}

The continuum quantum-mechanical approach is also known as the self-consistent reaction field (SCRF) method. In this approach, the solvent is modeled as a continuous dielectric medium that can be polarized by the solute charge distribution, generating a reaction field which in turn affects the solute charge distribution and so forth. The standard formulation of the SCRF method was developed by Miertuš, Scrocco, and Tomasi (MST) within the *ab initio* framework [Miertuš *et al.* 1981, Miertuš and Tomasi 1982, Bonaccorsi *et al.* 1984], and is commonly known by the acronym "SCRF-MST method."

For an unperturbed solute molecule in a vacuum, the Schrödinger equation in the Born-Oppenheimer approximation is:

$$H_0 \Psi_0 = E_0 \Psi_0 \ , \tag{2-30}$$

where H_0 is the solute's electronic Hamiltonian, and Ψ_0 and E_0 are the electronic eigenfunction and eigenvalue, respectively. As usual, this differential equation can be solved approximately in the framework of the Hartree-Fock or self-consistent field (SCF) approach [Szabo and Ostlund 1989]. The Hartree-Fock equation of a solute in a solvent can be approximately solved within the framework of the Rayleigh-Schrödinger perturbation theory [Szabo and Ostlund 1989]:

$$(H_0 + V_\sigma) \Psi_R = E_R \Psi_R \ , \tag{2-31}$$

assuming that the wavefunction Ψ_R for the solute-solvent system can be derived from the wavefunction Ψ_0 of the solute as a Taylor power-series expansion in the parameters of the perturbation V_σ. The perturbation operator V_σ can be evaluated as follows:

$$V_\sigma(\mathbf{r}) = \int_S \frac{\sigma(\mathbf{s})}{||\mathbf{r}-\mathbf{s}||} d\mathbf{s} = \sum_i^M \frac{q_i}{||\mathbf{r}-\mathbf{r}_i||}, \tag{2-32}$$

where $\sigma(\mathbf{s})$ is the solvent charge 2D-density on the cavity's surface element **s**. This charge density is determined by the Laplace equation for the polarization of the dielectric continuum at the cavity boundary S [Miertuš *et al.* 1981]:

$$\sigma(\mathbf{s}) = -\frac{\varepsilon - 1}{4\pi\varepsilon}\frac{\partial}{\partial \mathbf{n}}\left[V_\rho(\mathbf{s}) + V_\sigma(\mathbf{s})\right]_S \quad , \tag{2-33}$$

where $V_\rho(\mathbf{s})$ and $V_\sigma(\mathbf{s})$ are the electrostatic potential contributions for the free solute and a cavity surface element, respectively. The parameter ε in eq. (2-33) is the dielectric constant of the continuum. Some molecular surface models can be used to build the solute-solvent interface (*e.g.*, a van der Waals surface). The electrostatic solute-solvent interaction is then calculated from the solute Hamiltonian H_0 and the perturbation potential V_σ:

$$\Delta E_{elst} = [\langle\Psi_R|\, H_0 + V_\sigma|\, \Psi_R\rangle + E^{nucl}] - [\,\langle\Psi_0|\, H_0|\Psi_0\rangle + E_0^{nucl}\,] \,, \tag{2-34}$$

where E^{nucl} and E_0^{nucl} are the nuclear repulsive energies with and without the solvent, respectively. [Note that the wave functions Ψ_0 and Ψ_R correspond to optimized geometries of the solute. These structures can differ when the solute is introduced in the cavity, and for this reason, $E^{nucl} \neq E_0^{nucl}$.] Finally, the Gibbs free energy contribution ΔG_{elst} can be evaluated in terms of the wave function Ψ_R and electron density $r(\mathbf{r})$ of the system as follows [Bonaccorsi *et al.* 1990]:

$$\Delta G_{elst} = \Delta E_{elst} + (1/2)\ [\langle\Psi_R|\, V_\sigma|\Psi_R\rangle - \int_S \rho^{nucl}(\mathbf{r})V_\sigma(\mathbf{r})d\mathbf{r}] \,, \tag{2-35}$$

where the last term is the expectation value of $V_\sigma(\mathbf{r})$ over the cavity. Note that the perturbation operator V_σ in eq. (2-31) is built from $V_\rho(s)$ and $V_\sigma(s)$ (see eq. (2-32) and eq. (2-33)), which in turn are obtained from the wave function Ψ_R (eq. (2-31)). Therefore,

this equation has to be solved iteratively by numerical integration over the boundary surface between the solute cavity and the continuum.

A different SCRF algorithm suggested by Wong *et al.* [1991] is included in the quantum-chemical software package Gaussian 92. Energies for solvated systems can be computed with a second-order Møller-Plesset perturbation theory (MP2) or a configuration interaction with double excitations (QCI). The electrostatic effect of the solvent is represented as an additional term to the molecular Hamiltonian in the gas phase:

$$H_{ef} = H_0 + H_1 . \tag{2-36}$$

The perturbation term H_1 describes the coupling between the molecular dipole vector operator **m** and the reaction field vector $\boldsymbol{F}$:

$$H_1 = \mathbf{m} \bullet \boldsymbol{F} . \tag{2-37}$$

The reaction field $\boldsymbol{F}$ is a function of the molecular dipole moment (**m**), the dielectric constant of the medium (ε), and the cavity radius (a_0):

$$\mathbf{F} = \frac{2(\varepsilon - 1)}{2\varepsilon + 1} a_0^3 \, \mathbf{m} . \tag{2-38}$$

Note that systems having zero dipole moment will not exhibit solvent effects in this model, and therefore SCRF calculations performed on them will give the same results as for the gas phase. This is an inherent limitation of the SCRF approach of Wong *et al.* [1991].

Using these methodologies, the partition coefficient can be calculated as the difference of solvation Gibbs free energies of solute in phases "1" and "2", ΔG^1_{solv} and ΔG^2_{solv}, [Miertuš and Moravek, 1990]:

$$\ln P_{2,1} = - \Delta G_{2,1} / RT \,, \tag{2-39}$$

where $\Delta G_{2,1} = \Delta G^{2}_{solv} - \Delta G^{1}_{solv}$ is the free energy of transfer of the solute *from phase "1" to phase "2"* (*i.e.*, $\Delta \overline{G}^{\circ}_{tr}$ in eqs. (1-29) and (1-30)). The usual $\log P_{1,2}$ requires the calculation of individual solvation free energies in each solvent. (In the case of the familiar P_{ow}, *n*-octanol is phase "2" and water is phase "1.") Because a completely *ab initio* calculation of log *P* is difficult, eq. (2-39) could be used as a source for an empirical correlation in a series of related molecules:

$$\log P_{2,1} = a \Delta G_{2,1} + b \;. \tag{2-40}$$

2.2.3 Double-Layer Polarizable Quantum Continuum Model

While the advances in *ab initio* chemistry have been remarkable, the advances in semiempirical methods (both molecular orbital theory and molecular mechanics) have had an impact on a broader range of molecular phenomena. Semiempirical and semiclassical methods are still the only choice for large biomolecules.

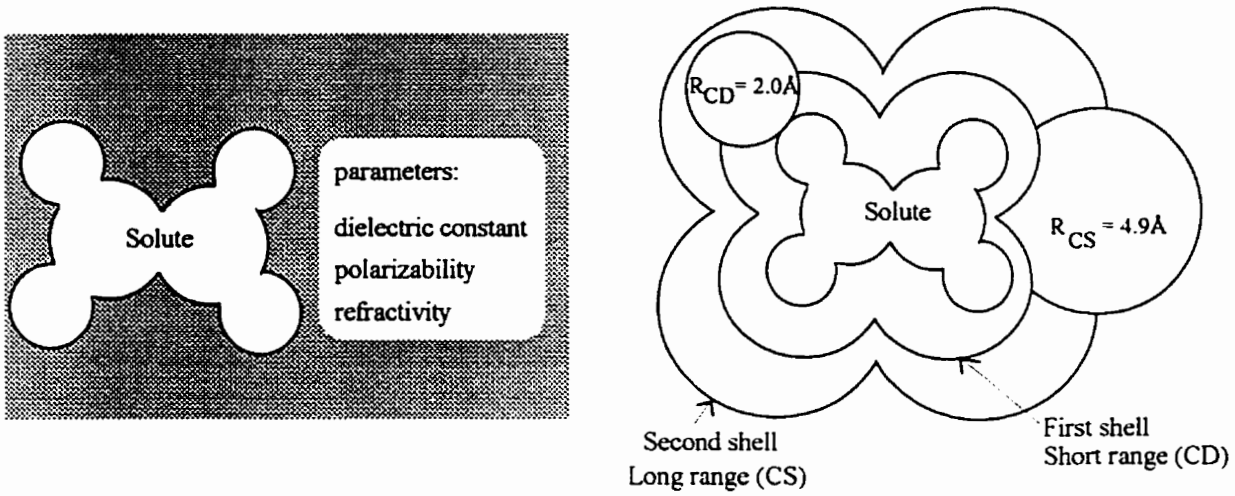
Cramer *et al.* [1991, 1992 a, b, c] have presented a semiempirical quantum mechanical SCRF algorithm, in which two solvation shells are used. The method is based on the AM1 semiempirical method [Dewar *et al.* 1985]. For short-range interactions, a solvent radius of 2.0 Å is used, and for long-range interactions, a solvent radius of 4.9 Å is used to mimic a second layer of water molecules. (see Fig. 2-2 (B).)

In this latter approach, the solvation free energy is expressed as the sum:

$$\Delta G_{solv} = \Delta G_{enp} + \Delta G_{cds} \,, \tag{2-41}$$

where ΔG_{enp} stands for the electronic, nuclear repulsive, and polarization energies. The term ΔG_{cds} accounts for the free energy of forming a cavity in the solvent and for the

changes in the dispersion interactions and the solvent structure that accompany the solvation process.



Polarizable Continuum Model

Double-layer polarizable Continuum Model

A **B**

Figure 2-2: Two polarizable continuum models used in the quantum-mechanical calculation of solvation free energies: (A) Polarizable continuum, (B) Double-layer polarizable continuum.

2.3 Monte Carlo Simulation

The Monte Carlo (MC) method is a tool of computational mathematics and is concerned with "experiments" involving random numbers [Northrup and McCammon 1980, Friesner and Levy 1984, Heermann 1990]. The MC approach allows us to obtain an approximate thermodynamic description of "realistic" systems that cannot be treated analytically. A number of recent techniques employ this methodology to compute solvation free energies. Here, I review some of the important notions and applications.

2.3.1 Monte Carlo Method

To some extent, MC computations are always related to the numerical estimation of a multi-dimensional integral. In this case, the "integration" takes place over a random sampling of points instead of over a regular array of points or a continuum.

In a naïve (though inefficient) implementation of the MC approach for an N-particle system, each particle is put at a random position in the $6N$-dimensional phase space. The resulting configuration has a statistical Boltzmann weight exp(-E/kT), where E is the configurational energy:

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} V(r_{ij}) \ , \quad (i \neq j), \tag{2-42}$$

where $V(r_{ij})$ is the interaction potential between particles i and j, separated by a distance r_{ij}. The statistical thermodynamic average $\bar{F}$ of a molecular property F can then be computed in the canonical ensemble [McQuarrie 1976]:

$$\bar{F} = \frac{\int F \exp(-E/kT) d^{3N}p\, d^{3N}q}{\int \exp(-E/kT) d^{3N}p\, d^{3N}q} \ , \tag{2-43}$$

where $d^{3N}p d^{3N}q$ is a volume element in the $6N$-dimensional phase space, k is the Boltzmann constant, and T the absolute temperature. In practice, the integrals in eq. (2-43) are computed as a discrete sum over the sampled configurations. The denominator of eq. (2-43) is the classical partition function:

$$Q = \int \exp(-E/kT)\, d^{3N}p\, d^{3N}q \ . \tag{2-44}$$

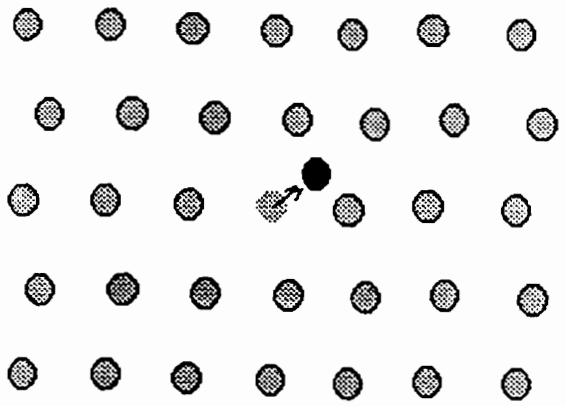
When the interaction potential between particles is velocity-independent, the momentum integrals cancel out, and we have only an integration over the $3N$-dimensional configuration space,

$$\bar{F} = \frac{\int F \exp(-E/kT) d^{3N}q}{\int \exp(-E/kT) d^{3N}q}, \tag{2-45}$$

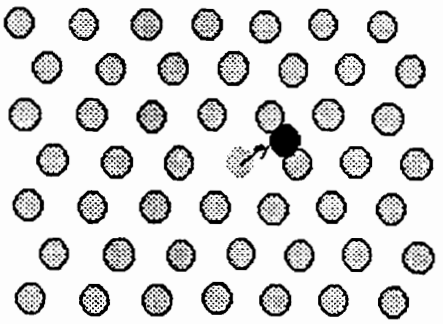
where $d^{3N}q$ is the volume element of the $3N$-dimensional configurational space. The denominator of eq. (2-45) is the *configurational integral* Z_N. In a more rigorous context, these MC averages should be viewed as computed along a Markov chain of sampled configurations [Heermann 1990].

2.3.2 Metropolis Monte Carlo Algorithm

The simplest MC approach puts N particles at random positions in an N-dimensional cube, and then calculates the energy of the system according to eq. (2-45). This method is not practical for close-packed configurations because there are too many sampled configurations with high energies and low statistical weights. This is illustrated in Figure 2-3. For the high density phase, there is a large possibility of producing a high-energy configuration, *i.e.*, a configuration with a small exp(-*E*/*kT*) weight that contributes little to the statistical averages.



For the liquid at lower density, after moving a particle randomly, there is a higher probability that the weight exp(-E/kT) of the resulting configuration is not too small.



For the liquid at higher density, after moving a particle randomly, there is a higher probability that the weight exp(-E/kT) of the resulting configuration is too small.

Figure 2-3. Compared random (Monte Carlo) "moves" in phases of low and high density.

Instead of sampling configurational space in a completely random fashion, the Metropolis MC approach [Metropolis *et al.* 1953] models the *transition probability* between configurations. With this transition probability, one can decide whether or not to change from one random configuration to another (a so-called MC "move"). In Metropolis MC, the N particles are placed in an initial configuration, for example, in a regular lattice. Then, one generates a new configuration by changing the Cartesian coordinates (X_i, Y_i, Z_i) of each particle as follows:

$$\begin{aligned} X_i &\rightarrow X_i + \alpha\xi_1 \\ Y_i &\rightarrow Y_i + \alpha\xi_2 \\ Z_i &\rightarrow Z_i + \alpha\xi_3 , \end{aligned} \tag{2-46}$$

where α is the maximum allowed displacement, and ξ_1, ξ_2 and ξ_3 are random numbers between -1 to 1. With the new configuration, one calculates the energy change ΔE. If $\Delta E<0$, (*i.e.*, the move would bring the system to a lower energy state), the move is allowed and each particle is put in its new position. If $\Delta E>0$, the move is allowed with the probability $\exp(-\Delta E/kT)$. This is implemented by generating a random number ξ_4 between 0 and 1. If $\xi_4 < \exp(-\Delta E/kT)$, then the new configuration is accepted; otherwise, it is rejected. The approach is then repeated, starting from the accepted configuration. The sequence of accepted configurations defines a Markov chain, which is then used to calculate thermodynamic averages as explained in §2.3.1.

2.3.3 Test Particle Approach

This approach was originally proposed by Widom in the 1960's [Widom 1963, Romano and Singer 1979] and recently adapted to the study of solvated systems [Forsman and Jönsson 1994]. Consider an equilibrated system composed of N solvent molecules with a total interaction energy U_0. Let us introduce a perturbation by inserting *one* solute molecule whose interaction energy with the solvent molecules is U_p. Then the configurational integral Z_N for the system may be written as:

$$Z_N = \int \exp(-(U_0 + U_p)/kT)\, dr_N = Z_{N,0} < \exp(-U_p/kT) >_0 , \qquad (2\text{-}47)$$

where $Z_{N,0}$ is the configurational integral of a pure solvent, and $<\cdots>_o$ indicates a configurational average computed with the unperturbed reference system. [All integrals are evaluated with the Monte Carlo method.] The excess Helmholtz free energy, A_{ex}, associated with the perturbation is given by [Forsman and Jönsson 1994]:

$$A_{ex} = -kT \ln < \exp(-U_p/kT) >_0 , \qquad (2\text{-}48)$$

and the excess energy, U_{ex}, is:

$$U_{ex} = < U_p \exp(-U_p/kT) >_0 / < \exp(-U_p/kT) >_0 + [< U_0 \exp(-U_p/kT) >_0 / < \exp(-U_p/kT) >_0 - <U_0>_0] . \qquad (2\text{-}49)$$

The first term in the right-hand side of eq. (2-49) is the average energy due to the interaction between the solute and the water molecules. The second term reflects the energy change in the system as the water molecules reorganize themselves around the solute.

The mean force for solute-solvent interactions can then be calculated directly by taking the derivative of the excess free energy in eq. (2-48). When *two* solute molecules are randomly inserted, the module of the mean force $F(r^*)$ is separated into direct (F_d) and indirect (F_{ind}) contributions:

$$F(r^*) = F_d(r^*) + F_{ind}(r^*) , \qquad (2\text{-}50)$$

where r^* is the distance separating the centres of mass of the solutes. The term $F_d(r^*)$ is due to solute-solute interactions and the term $F_{ind}(R)$ corresponds to the interactions between the solutes and all solvent molecules. The force $F(r^*)$ is a *quantitative measure of solute-solvent affinities and can be used in characterizations of hydrophobicity.*

Recently, Forsman and Jönsson [1994] used the above MC approach to study hydrophobic interactions between solute molecules in a bulk solvent and in the presence of a boundary. The hydrophobic force and free energy of solvation for two particles interacting by Lennard-Jones and hard-sphere potentials were studied as a function of their separation. The entropy, energy, and free energy of a single hydrophobic particle were also calculated.

The bulk system (BS) is a model for an infinitely dilute aqueous solution. In this approach, two nonpolar solute molecules are inserted in water. This method is appropriate when investigating the hydrophobic interaction of small particles, but it becomes increasingly difficult for large molecules.

The anisotropic system (AS) is a model that consists of water molecules enclosed between two infinitely large hydrophobic walls. Forsman and Jönsson [1994] studied two models of walls: a "hard" repulsive surface and a geometrical array simulating a surface of silica.

The results obtained by these approaches are very approximate, but they provide valuable insight to the spatial organization of solvent-solute clusters. These results represent the state of the art in the purely theoretical modeling of hydrophobic interactions.

2.4 Molecular Dynamics Simulation

Molecular Dynamics (MD) methods are as old as the Metropolis Monte Carlo algorithm. The first MD simulations dealt with simple fluids models [Alder and Wainwright 1957, 1959]. Molecular dynamics uses molecular mechanics to compute a wide variety of dynamic and thermodynamic properties of molecular systems. In the following sections, I discuss briefly the use of MD simulations for solute-solvent systems.

2.4.1 Molecular Mechanics

Molecular mechanics represents the potential energy hypersurfaces using parametrized classical-mechanics force fields [Burkert and Allinger 1982]. Only nuclear contributions are explicitly included in molecular mechanics. All information on electron interactions is included (implicitly) in the force-field parameters (*e.g.*, force constants). A typical potential energy function includes terms for bond stretching, bond angle deformation, hindered rotations about single bonds, and nonbonded interactions between atoms separated by three or more bonds. Bond stretching and bending are usually modeled with a simple harmonic potential [Wilson *et al.* 1955]. Torsional (dihedral) rotations are generally modeled as a truncated Fourier series, whereas nonbonded interactions are often represented as Lennard-Jones (van der Waals) and Coulomb (electrostatic) potentials [Lybrand 1990]. A very simple molecular-mechanics force field can then be represented by the following potential energy function:

$$\begin{aligned} V(r_1, r_2, \ldots, r_N) = {} & 1/2 \sum_b K_b (R_b - R_{b0})^2 + 1/2 \sum_a K_a (\theta_a - \theta_{a0})^2 + \\ & + \sum_d K_d [1 + \cos(n_d f_d - g_d)] \\ & + \sum_{i,j} \{ b_{ij} [(\sigma_{ij}/r_{ij})^{12} - 2 (\sigma_{ij}/r_{ij})^6] + q_i q_j / er_{ij} \} , \end{aligned} \qquad (2\text{-}51)$$

where K_a, K_b, and K_d are force constants associated with bending stretching and torsions, respectively. The parameters R_{b0}, θ_{a0}, and g_d are equilibrium values for b-th bond length, a-th bond angle, and the d-th torsional angle, respectively. The parameters σ_{ij}, b_{ij}, and q_i are the Lennard-Jones radius, the depth of the potential well, and the atomic partial charge, respectively. Anharmonic potentials may be used in place of harmonic potentials. In addition, some force fields include cross-terms coupling bond lengths with bond angles.

These relatively simple potential energy functions can be parameterized to represent the properties and behavior of solvated biomolecules. Force constants for bond length, bond angle, and torsional angle terms may be determined from spectroscopic methods or from quantum mechanical calculations for small reference molecules. Lennard-Jones parameters may be derived from scattering, crystal packing, or liquid

structure data as well as from quantum mechanical calculations. Partial charges can be determined from various population analyses of the charge distribution [Bachrach 1994]. Various force fields are available for different purposes. Some familiar ones are MM2 and MM3 [Burkert and Allinger 1982, Bowen and Allinger 1991], AMBER [Weiner *et al.* 1984], CHARMM [Brooks *et al.* 1983, Smith and Karplus 1992], GROMOS [Hermans *et al.* 1984], and OPLS [Jorgensen and Tirado-Rives 1988].

2.4.2 Molecular Dynamics

Unlike Monte Carlo, which is a *probabilistic* (stochastic) procedure, molecular dynamics (MD) is a *deterministic* one. In MD, particles move according to classical Newtonian mechanics or quantum mechanics. In the case of studying large molecular systems (*e.g.*, solution phenomena), one normally uses the classical approach with a molecular mechanics force field. A MD simulation is performed by integrating Newton's equations of motion:

$$\mathbf{F}_i(t) = m_i \mathbf{a}_i(t) = m_i \partial^2 \mathbf{r}_i(t) / \partial t^2, \quad i = 1,2,\ldots N \ (N \equiv \text{number of atoms}), \tag{2-52}$$

where $\mathbf{F}_i(t)$ is the force acting on atom i at time t; $\mathbf{a}_i(t)$ and $\mathbf{r}_i(t)$ are the instantaneous acceleration and position of atom i at time t, respectively. The force on atom i is computed from the potential energy function eq. (2-51):

$$\mathbf{F}_i(t) = - \partial / \partial \mathbf{r}_i \{V(r_1, r_2, \ldots, r_N)\} . \tag{2-53}$$

A variety of algorithms have been used to integrate eqs. (2-52) and (2-53) using a discrete time step Δt. The Verlet algorithm is a common strategy used to make eq. (2-52) discrete. It computes position vectors using the forces and previous positions of atoms [Haile 1992]:

$$\mathbf{r}_i(t+\Delta t) \approx 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + \mathbf{F}_i(t)(\Delta t)^2 / m_i + O[(\Delta t)^4] . \tag{2-54}$$

The velocity vectors are computed with a central-distance formula

$$\mathbf{v}_i(t) = [\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t)]/2\Delta t + O[(\Delta t)^3] . \quad (2\text{-}55)$$

The set of coordinates and velocities $\{\mathbf{r}_i(t),\mathbf{v}_i(t)\}$ defines the *molecular dynamics trajectory*. From the velocities, $\{\mathbf{v}_i(t)\}$, the "instantaneous temperature" $T(t)$ of the system is computed:

$$T(t) = (k/3N) \Sigma_i m_i \|\mathbf{v}_i\|^2, \quad (2\text{-}56)$$

The simulation above samples a *microcanonical* ensemble (with a constant number of particles, constant volume, and constant total energy). This approach is appropriate for isolated, conservative systems. It is possible to perform dynamics simulation in other conditions (*e.g.*, constant temperature or pressure). Constant-temperature trajectories can be constructed by "weakly coupling" the molecular system to a simulated heat bath [Berendsen *et al.* 1984, van Gunsteren and Berendsen 1990]. Dissipative systems can be better simulated by stochastic or Langevin dynamics [Heermann 1990], which brings friction terms and Brownian motion into the Newton's equation of motion.

Thermodynamic average properties can be determined from sufficiently long trajectories. If a MD trajectory is measured from an initial time t_0 to a final time t, the average of a mechanical property $A(t)$ is:

$$< A > \approx \frac{1}{t - t_0} \int_{t_0}^{t} A(t) dt . \quad (2\text{-}57)$$

In practice, the integral eq. (2-57) is made discrete by using a time step of the order Δt $\sim 10^{-15}$ *s*. If $A(t)$ is taken as the energy $E(t)$, we can calculate the mean internal energy $<E>$ and the partition function. From these properties, one can determine all other relevant

thermodynamic properties [McQuarrie 1976] and in this manner, it is possible, in theory, to estimate hydration free energies for a given solute.

2.5 Hybrid Algorithms

No single method can elucidate all aspects of solvation phenomena and hydration shell structure. From *ab initio* quantum chemistry to molecular mechanics, every approach has its own pros and cons. In recent years, a growing trend has been to develop hybrid methods that profit from the advantages in each technique.

2.5.1 Combination of Monte Carlo and Molecular Dynamics Simulation

An intrinsic weakness of common MD algorithms is their difficulty in producing reliable time averages for many properties [Heermann 1990, McDonald and Still 1994]. This problem is caused by an inadequate ensemble sampling and relatively short trajectories. Thus, the results derived may depend on the initial conditions or on the length of the trajectory. Recent reports suggest that many previously reported simulations were too short to give sufficiently accurate free energies (*e.g.*, ± 0.5 kcal/mol) for practical applications [McDonald and Still 1994].

A complete configurational sampling is difficult with MD when the system has multiple conformations separated by large energy barriers. The problem is that free energies can be computed only when the local configurational space about each significantly populated conformer is sampled with the correct statistical weight. Using standard simulation methods, barrier crossing is a rare event and thus the sampling is not exhaustive. Limited by a large energy barrier, a standard MD approach could spend all of its time just sampling the local space of the *starting* conformation.

A mixed Monte Carlo and Stochastic Dynamics algorithm has been developed by Gaunieri and Still [1994] in order to improve the conformational search. This new method is based on the observations that:

(i) Dynamical methods (*e.g.*, stochastic dynamics, SD) do a good job of sampling phase space in systems whose populated states are not separated by large energy barriers.
(ii) MC methods can sample wide regions of the configurational space and produce canonical ensemble averages even for high barriers, provided that a sufficiently long number of steps (MC "moves") are used.
(iii) A mixed simulation algorithm can be devised that alternates SD and MC steps. This approach samples both local and remote regions of the configurational space.

2.5.2 Combination of Quantum Mechanics and Molecular Mechanics

Direct quantum-mechanical approaches to solvation phenomena (*e.g.*, a discrete model involving a large number of solvent molecules) are very difficult. Similarly, the quantum-mechanical continuum model for a solvated macromolecule is hard to implement. For this reason, molecular mechanics has been used to extract thermodynamic properties of solution systems. A more reliable alternative is to combine both quantum-mechanical and molecular-mechanical potentials (QM/MM) in dynamics simulations [Field *et al.* 1990].

A. Partitioning of the System

The method of Field *et al.* [1990] divides the system of interest into a small "quantum mechanical region" (QM), a large "molecular mechanical region (MM)", and a boundary region. The QM and MM regions contain all atoms that are explicitly treated in the calculation, while the boundary region is included so as to account for the neglected surroundings (see Fig. 2-4).

Atoms in the QM region are represented as nuclei and electrons, within the Born-Oppenheimer approximation. The equilibrium nuclear positions in the QM region are determined from the quantum-mechanical potential energy. This region contains all the atoms involved in the reaction process of interest (*e.g.*, the ligand-receptor binding).

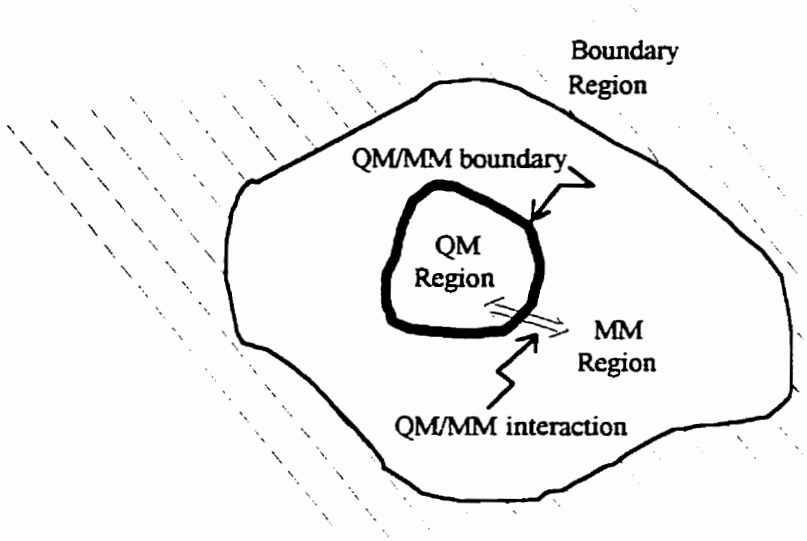


Figure 2-4: Partitioning of the system into quantum mechanical (QM), molecular mechanical (MM), and boundary regions. [Adapted from Field *et al.* 1990.]

The MM region contains the remaining atoms in the system. They are represented as only nuclei, and their interactions are determined from an empirical force field. They constitute the immediate environment for the "QM atoms." The "MM atoms" are included because their interactions and dynamics will influence the behaviour of the QM region.

B. Hamiltonian for the "Mixed" QM/MM System

The Schrödinger equation of the entire system is

$$H_{\text{eff}}\,\Psi(\mathbf{r}, \mathbf{R}_Q, \mathbf{R}_M) = E(\mathbf{R}_Q, \mathbf{R}_M)\,\Psi(\mathbf{r}, \mathbf{R}_Q, \mathbf{R}_M)\ , \qquad (2\text{-}58)$$

where H_{eff} is the effective Hamiltonian of the system. The solvent-solute wavefunction Ψ depends on the electron coordinates, $\mathbf{r}$. As well, it depends parametrically on the positions of the quantum-mechanical nuclei, $\mathbf{R}_Q$, and the molecular-mechanical atoms $\mathbf{R}_M$. The effective Hamiltonian for the partitioned system is written as the sum of four terms:

$$H_{\text{eff}} = H_{\text{QM}} + H_{\text{MM}} + H_{\text{QM/MM}} + H_{\text{Boundary}} \,, \tag{2-59}$$

where H_{QM} and H_{MM} describe the QM and the MM regions, respectively. The remaining terms describe boundary regions (*i.e.*, the QM/MM boundary and the continuum boundary, respectively).

The potential energy for a conformation of QM nuclei and MM atoms is given by:

$$\begin{aligned} E &= \langle\Psi|H_{\text{eff}}|\Psi\rangle/\langle\Psi|\Psi\rangle \\ &= \langle\Psi|\, H_{\text{QM}} + H_{\text{QM/MM}} + H_{\text{Boundary(QM)}}|\Psi\rangle/\langle\Psi|\Psi\rangle \\ &\quad + E_{\text{MM}} + E_{\text{Boundary(MM)}} \,, \end{aligned} \tag{2-60}$$

where the boundary term has been separated into QM and MM parts. Using standard coordinate notations, the terms in Eq. (2-60) are as follows [Field *et al.* 1990]:

$$H_{\text{QM}} = -1/2 \sum_i \nabla_i^2 + \sum_{ij} 1/r_{ij} - \sum_{ia} Z_a/r_{ia} - \sum_{ab} Z_a Z_b/R_{ab} \,, \tag{2-61}$$

$$H_{\text{MM}} = E_{\text{MM}} \,, \tag{2-62}$$

$$\begin{aligned} H_{\text{QM/MM}} &= \sum_i q_M/r_{iM} + \sum_{aM} Z_a q_M/R_{aM} + \\ &\quad + \sum_{aM} \{A_{aM}/(R_{aM})^{12} - B_{aM}/(R_{aM})^6\}, \end{aligned} \tag{2-63}$$

where lower case letters (a,b) identify nuclei in the QM region and the letter "M" identifies nuclei in the MM region. The first term in Eq. (2-63) involves distances between electrons in the QM region and atoms in the MM region. This term must be included in the HF-SCF procedure.

The boundary region is a standard feature of many QM and MM calculations. Because one is restricted to deal with finite size systems, these boundary terms can mimic the behavior of the excluded portion of the system. Two methods are commonly used in

MM calculations; they are the periodic boundary and the stochastic boundary approaches [Field *et al.* 1990].

As discussed in §1.1.1, an assessment of hydrophobicity can be made in terms of the force between two solute molecules in the presence of a solvent. In the hybrid QM/MM approach, the forces on the QM nuclei, $\mathbf{F}_Q$, and on the MM atoms, $\mathbf{F}_M$, are obtained by differentiating eq. (2-60):

$$\mathbf{F}_Q = -\partial E/\partial \mathbf{R}_Q \quad \text{and} \quad \mathbf{F}_M = -\partial E/\partial \mathbf{R}_M \ . \tag{2-64}$$

Liu & Shi [1994] have recently applied this methodology for the first time to the study of solvation phenomena. The authors determined the free energy profile of the nucleophilic addition between formaldehyde and OH^- ions, in aqueous solution. The reaction path in the solution was determined by the semiempirical quantum method AM1 [Dewar *et al.* 1985].

The material covered in this Chapter illustrates the wide spectrum of the state-of-the-art techniques for computer simulation of solutions. Although we are still far from producing reliable *ab initio* predictions of log *P*, the progress is fast. In Chapters 3, 4, and 5 of this thesis, I contribute a number of theoretical developments to the modeling of hydrophobicity, including its prediction from a small number of selected structural data.

2.6 Some Heuristic Measures of Hydrophobicity

Partition coefficients determined by the methods discussed in §1.3 have been used in various applications, including studies of equilibria, emulsions, and design of ion-selective electrodes [Leo *et al.* 1971]. The ρ–σ–π analysis [Hansch and Fujita 1964] illustrated the important role of partition coefficients in QSAR for drug design. During the last three decades, the methodology of drug design has developed tremendously while at the same time approaches to assessing molecular lipophilicity have diversified. In this

section, I discuss briefly some recent developments in the use of hydrophobicity measures in modern drug design.

2.6.1 Hydrophobic Moment

The concept of "hydrophobic moment", which goes beyond the simple log P characterization of hydrophobicity, has been introduced by Eisenberg *et al.* [1982]. This property provides a measure of the asymmetry in the molecular hydrophobicity (the so-called *amphiphilicity*) [Eisenberg and McLachlan 1986]. According to this idea, the hydrophobicity of a molecule can be characterized by two indexes. One (log P_{ow} or a related constant) gives its overall *magnitude*. The other is a measure of the extent of the "*hydrophobic polarization*" throughout the molecule. The amphiphilicity reflects the fact that molecules are made up of a number of polar and nonpolar moieties contributing differently to the overall hydrophobicity.

The use of hydrophobic moments provides interesting insights into several structural features. It has been applied to the classification of the hydrophobicity of amino acid residues and to the establishment of the amphiphilicity of regular protein secondary structural elements [Eisenberg and McLachlan 1986]. These authors have found that the hydrophobic moments of neighboring segments of secondary structures tend to oppose each other in correctly folded proteins, but not in incorrectly folded ones. As well, hydrophobic moments can be used to classify peptide helices.

The above applications of hydrophobic moments are formulated in terms of amino acid contributions. Similarly, hydrophobic moments can be expressed in terms of atomic contributions. According to Eisenberg and McLachlan [1986], the hydrophobic moment is a vector determined from the following sum over atoms:

$$\mathbf{m}_a = \sum_i s_i \, \Delta\sigma_i \, \mathbf{r}_i - \langle s \, \Delta\sigma \rangle \sum_i \mathbf{r}_i \ , \qquad (2\text{-}65)$$

where $\mathbf{r}_i$ is the position vector of nucleus i, s_i is the accessible surface area of atom i, $\Delta\sigma_i$ is an atomic solvation parameter, and the brackets indicate the mean value over all atoms. The second term in eq. (2-65) makes m_α invariant with respect to the choice of origin for the coordinates. When the sum in eq. (2-65) is restricted to the atoms in a single side chain, a residue hydrophobic moment is defined.

For the sake of illustration, I have reproduced the values of these moments for amino acid residues in Table 2-1 [Eisenberg and McLachlan 1986]. From these values, it is apparent that the residues with greatest amphiphilicity are Arg, Lys and Glu. In contrast, the most hydrophobic residues, Trp, Phe, Leu and Ile, all have small amphiphilicities. The direction of the hydrophobic moment is expressed in Table 2-1 by the cosine of the angle (cos q) it forms with a vector defined from the a-carbon to the center of the side-chain. For the highly amphiphilic residues, the direction of the moment is nearly antiparallel to this latter reference vector ($\cos q \approx -1$).

Table 2-1. Amino-acid hydrophobic moments [kcal Å/mol].*

Amino-Acid Residue	m_s	cos q
Gly	(0)	–
Ala	0	–
Val	0.48	0.84
Leu	1.0	0.89
Ile	1.2	0.99
Pro	0.18	0.22
Cys	0.17	0.76
Met	1.9	0.94
Thr	1.5	0.09
Ser	0.73	-0.67
Phe	1.1	0.92
Trp	1.6	0.67
Tyr	1.8	-0.93
Asn	1.3	-0.86
Gln	1.9	-1.0
Asp	1.9	-0.98
Glu	3.0	-0.89
His	0.99	-0.75
Lys	5.7	-0.99
Arg	10.0	-0.96

* Adapted from Eisenberg and McLachlan [1986].

Hydrophobic moments can also be used in the estimation of solvation free energy for protein folding and binding. However, the application of this approach is limited by the availability of the experimental parameters $\Delta\sigma_i$.

2.6.2 Complementary Hydropathicity Map

Kellogg and Abraham [1992] have developed a program (HINT, Hydrophobic INTeractions) that uses hydrophobicity and structural information to construct a "hydropathicity map" of a receptor site. In turn, this is used to design the receptor ligands. The HINT model is based on the notion that hydrophobic molecules are attracted to nonpolar solvents, while hydrophilic molecules are attracted to polar solvents such as water. By extension, it is assumed that molecules will be attracted to analogous regions in the biological receptor. The HINT model uses atomic fragmental contributions to partition coefficients in order to map the hydrophobicity of a binding site. As a rule, positive hydrophobicity atomic contributions represent hydrophobic atoms and negative contributions represent hydrophilic groups, polar atoms, or charged species.

The contribution to hydrophobicity of one atom changes depending on whether it is present on the surface exposed to the solvent or "buried" inside. This difference can be modeled by using the solvent accessible surface areas [Lee and Richards 1971, Richards 1977, Connolly 1983] to scale the atomic contributions. The program HINT models the dependence of the hydrophobic effect on interatomic distances as a linear combination of two functions: an exponential decay for the "coupling" between hydrophobic atomic contributions, and a Lennard-Jones (6-12) potential function to describe nonbonded interactions. If b_{ij} is the interaction between two atoms (i, j), its explicit form is:

$$b_{ij} = s_i f_i s_j f_j R_{ij} + e_{ij} \quad , \tag{2-66}$$

where s_i is the solvent accessible surface area, and f_i is a hydrophobic atomic constant (see below). The functions R and e depend on the distance r between atoms i and j:

$$R_{ij} = T_{ij}\, e^{-r} \quad , \tag{2-67}$$

$$e_{ij} = s\, e_{ij} [\, 2(r_{vdw}/r)^6 - (r_{vdw}/r)^{12}] \, , \tag{2-68}$$

where e_{ij} and r_{vdw} are the standard Lennard-Jones parameters [Hirschfelder *et. al* 1954], and s is the *total* molecular surface area. The "sign-flip" function T_{ij} is used as an adjustable parameter for correcting unfavorable polar-polar interactions or for taking into account the occurrence of hydrogen bonding.

2.6.3 3D Molecular Lipophilicity Potential Profiles

Furet *et al.* [1988] and Audry *et al.* [1989a,b] have developed the concept of "molecular lipophilicity potential" as a tool for the "visualization" of the three-dimensional hydrophobic characteristics of a compound. By considering that hydrophobicity is somehow distributed all over a molecule, this approach can describe the details of the lipophilic and hydrophilic regions of a molecular surface.

In introducing a molecular lipophilicity potential, we consider a molecule M surrounded by organic solvent molecules of low polarity and assume that its overall lipophilicity — measured, for example, by log P — can be decomposed into fragmental or atomic contributions. These discrete contributions to hydrophobicity are represented by the parameters f_i (cf. eq. (2-66)):

$$\log P = \sum_i f_i \quad . \tag{2-69}$$

Intuitively, the arrangement of the solvent molecules around M is expected to vary from a random distribution at far distances to a more ordered state as one gets closer to M, depending on the lipophilic/lipophobic tendencies of the molecular fragments. According to this picture, the distribution of the solvent molecules around M will depend on the hydrophobic f_i constants and the distance of the solvent to each fragment. Audry *et al.* [1989a,b] have expressed this idea in a "molecular lipophilicity potential" (MLP), defined as follows:

$$\text{MLP} = \sum_i f_i /(1 + d_i) \ , \tag{2-70}$$

where d_i is the distance (in Å) between a given point outside the molecular surface and the fragment "i."

The above formula indicates that solvent molecules around M experience a hydrophobic "force field," whose value depends on interactions with lipophilic ($f_i > 0$) and hydrophilic ($f_i < 0$) groups of M. The magnitude of this interaction is maximal when d_i is equal to zero and progressively diminishes farther from the solute. Provided a set of f_i values, eq. (2-70) characterizes the hydrophobicity around a molecule, much as the electronic properties of this molecule can be characterized by the electrostatic force field created by its charge distribution. In this context, the MLP can be regarded as an extension of the concept of hydrophobic moment as discussed before.

Note that the MLP is only a heuristic notion, conceived to give a 3D extension to the simple log P representation of hydrophobicity. Eq. (2-70) correctly conveys chemical intuition and current ideas concerning hydrophobicity. Nevertheless, eq. (2-70) is not based on any rigorous theoretical framework. Other mathematical expressions could also be used. A number of functions were compared by Croizet *et al.* [1990], but it was found that none offered a particular advantage over eq. (2-70). A number of alternatives are also discussed by Heiden *et al.* [1993]. It has to be pointed out that atomic lipophilicity parameters f_i's used in the MLP are obtained from the PLS regression with experimental data log P_{ow}. Atomic lipophilicity parameters f_i's are best for the predictions of log P_{ow} in the statistical point of view, but they are not for MLP, because the signs of f_i's may lose their physical meaning in the regression calculations. I will discuss this issue in more detail in Chapter 3.

In summary, the definition of the MLP function is the first attempt to represent hydrophobicity as a function varying continuously over the different parts of a molecule. The MLP takes into account the effect of the atomic environment on the hydrophobicity of a fragment. By combining a representation of geometric and hydrophobic properties,

the MLP becomes a useful tool in molecular modeling and drug design. It can be applied to analyze the complementarity of a ligand with the active site in terms of shape and hydrophobicity. For instance, one can compare the MLP created by the ligand on a surface resembling the active site to the MLP created by the macromolecule on its own active site. Similar approaches have been used in the literature for the more rigorously defined electrostatic potential. In this thesis, I am interested in *comparing electrostatics and hydrophobicity as tools for molecular design.*

Finally, it is worth stressing that the evaluation of the MLP is simple and requires little computer time. To generate a profile, one needs only the 3D nuclear coordinates and the atomic hydrophobicity parameters f_i as input. In Chapter 3, I present some new results on the relation between the 3D MLP profiles and the electrostatic potential profiles.

2.6.4 Atomic Hydrophobic Parameters f_i

As shown in eq. (2-70), atomic hydrophobic parameters f_i are the building blocks for the construction of the molecular lipophilicity potential. In the next Chapter, I discuss briefly some theoretical methods that can be used to compute the free energy of a solution, and thus eventually derive parameters such as f_i. However, as of today, no theoretical method is advanced enough to succeed in this task and, therefore, one has to use experimental data to derive the f_i values.

Starting from eq. (1-69) in conjunction with the log P data for about 500 representative organic molecules (and using least-square techniques), Ghose and Crippen [1986] have derived f_i values for atoms in the most common functional groups. They have classified the atoms in 90 different "types" or classes. These classes take into account the number and nature of the atoms directly connected to the one under consideration. Improved parameter sets can be found in Viswanadhan *et al.* [1989]. Furet *et al.* [1988] and Audry *et al.* [1989 a, b] used the parameters by Ghose and Crippen [1986] and Viswanadhan *et al.* [1989] to compute the lipophilicity potential.

Eisenberg and McLachlan [1986] took a similar approach for evaluating atomic hydrophobicity constants from the solvation energy of proteins. These authors use the free energy of transfer of a given amino acid residue from the interior of a protein to the aqueous phase as a measure of its hydrophobicity. In turn, the solvation free energy is written as a sum of atomic contributions proportional to the solvent-accessible surface areas s_i of the atoms in the residue R:

$$\Delta G_R = \sum_i \Delta\sigma_i s_i \ . \qquad (2\text{-}71)$$

By fitting ΔG_R values of small amino acid analogues to surface area contributions, Eisenberg and McLachlan [1986] obtained values for the proportionality constants $\Delta\sigma_i$ for five atom types: carbon, neutral oxygen and nitrogen, charged oxygen (O^-), charged nitrogen (N^-), and sulphur. These values can then be used for the computation of hydrophobic moments in eq. (2-65).

2.6.5 Group Contributions to the Hydration Thermodynamic Properties

In the studies of solution theory, a different approach and data set have been developed during the last three decades: group contributions to the hydration thermodynamic properties [Cabani and Gianni 1979, Cabani *et al.* 1971, 1981].

With the progress of experiments, a large number of data for standard thermodynamic functions of hydration ΔG°_h, ΔH°_h, ΔS°_h, ΔC°_h, and partial molar properties $\overline{C}^\circ_{p,2}$ and $\overline{V}^\circ_2$ (of non-charged organic compounds) in water are available. These data represent a subtantial reservoir of information on water-organic solutes interactions and, based on this information, people want to know how the thermodynamic properties of water are related to the molecular structure of solute molecules.

Three methods are suggested by different authors [Cabani *et al.* 1971, 1981]. In the most commonly used, the contribution of a repetitive unit to each molar thermodynamic property is calculated as a difference between the property values for two

consecutive members of a homologous series. In the second method, the molecules are subdivided into groups, each of which is assumed to contribute a constant amount to the thermodynamic quantity. These contributions are calculated using a least squares method. Finally, in a third procedure, the hydrocarbons are selected as reference molecules and the effects of substituting some hydrocarbon surface area (or volume) with a like surface area (or volume) of hydrophilic nature are evaluated [Edward and Farrell 1975, Terasawa *et al.* 1975, Cabani *et al.* 1978, Cabani and Gianni 1979].

Chapter 3: Heuristic Lipophilicity Potential for Computer-Aided Rational Drug Design

Summary

In this chapter I suggest a heuristic molecular lipophilicity potential (HMLP), a structure-based technique requiring no empirical indices of atomic lipophilicity and where the input data used are molecular geometries and molecular surfaces. The HMLP is a modified electrostatic potential, combined with the averaged influences from the molecular environment. Quantum mechanics is used to calculate the electron density function $\rho(\mathbf{r})$ and the electrostatic potential $V(\mathbf{r})$ on the molecular surface, and from this information a lipophilicity potential $L(\mathbf{r})$ is generated. The HMLP is a unified lipophilicity and hydrophilicity potential. The electrostatic interactions of dipole and multipole moments, hydrogen bonds, and charged atoms in a molecule are included in the hydrophilicity in this model. Therefore, HMLP is also a unified electrostatic and lipophilic potential. The HMLP is used to study hydrogen bonds and water-octanol partition coefficients in several examples. The calculated results show that HMLP gives qualitatively and quantitatively correct, as well as chemically reasonable, results in cases where comparisons are available and these comparisons indicate that the HMLP has advantages over the empirical lipophilicity potential in many aspects. The HMLP is a three-dimensional and easily visualizable representation of molecular lipophilicity, and is recommended as a potential tool in computer aided three-dimensional drug design.

3.1 Introduction

Lipophilic or hydrophobic effect is one of the most important properties of organic and biological molecules. Molecular lipophilicity plays an important role in the study of molecular biological activities and the interaction between ligand and protein. For computer aided rational drug design, molecular lipophilicity is one of the key factors [Hansch 1971, Loew *et al.* 1993, Klebe *et al.* 1994, Jain *et al.* 1994, Kellogg and Abraham 1992]. In recent years, with the help of the great progress in computational chemistry, various computational methods have become available for studies of lipophilicity, including Monte Carlo simulation [Forsman and Jönsson 1994, Guillot *et al.* 1991, Tanaka and Nakanishi 1991], molecular dynamics simulation [Haile 1992, Smith and Haymet 1993], quantum *ab initio* and semiempirical SCRF (self consistent reaction field) methods of continuum medium model [Miertus and Tomasi 1982, Miertus and Moravek 1990, Wong *et al.* 1991, Bonaccorsi *et al.* 1990, Cramer and Truhlar 1992], and combined methods of quantum mechanics and molecular mechanics [Field *et al.* 1990].

Information about lipophilicity has been accumulated during the last three decades, concerning the physical nature of hydrophobic hydration (HH) and hydrophobic interaction (HI), the structure of hydration shells, the thermodynamic properties (enthalpic and entropic changes), and the lipophilic force between organic solute molecules in aqueous solution. Reflecting these developments, elaborate methods have been used to represent and describe molecular lipophilicity: one-dimensional scalar descriptor partition coefficients between water and organic solvent (most often, octanol is used) [Leo *et al.* 1971], one-dimensional vector descriptor lipophilicity moment [Eisenberg *et al.* 1982, Eisenberg and McLachlan 1986], two-dimensional lipophilicity maps [Heiden *et al.* 1993, Náray-Szabó and Nagy 1989, Náray-Szabó 1989, 1986], and three-dimensional lipophilicity potential [Croizet *et al.* 1990, Furet *et al.* 1988]. Audry *et al.* [1986, 1989] have suggested a formula for the calculation of lipophilicity potential. In their computer program developed for the calculations of lipophilicity potential, the formula takes the simple form [Croizet *et al.* 1990],

$$L(\mathbf{r}) = \sum_i \frac{f_i}{1 + \left\| \mathbf{r} - \mathbf{r}_i \right\|^{\gamma}}, \qquad (3\text{-}1)$$

where $\mathbf{r}_i$ is the position of nucleus i, and summation is over all constituent atoms. If the point $\mathbf{r}$ is on atom i, $\|\mathbf{r}\text{-}\mathbf{r}_i\| = 0$. In this situation, the denominator of eq. (3-1) is 1. This means that f_i is the dominant factor in the space surrounding atom i. Lipophilicity potential $L(\mathbf{r})$ defined by eq. (3-1) gives us a picture: for an organic solute molecule, its lipophilic surface area exerts the lipophilic force into the surrounding space to attract non-polar molecules, and repulse water molecules; whereas, its hydrophilic surface area exerts hydrophilic force to attract polar molecules. It is clear that lipophilicity potential defined by eq. (3-1) is not based on a rigorous theoretical model and is not a true physical potential. There may be other empirical MLP formulas, such as one using Gaussian type distance-dependence. A number of possible functions were compared by Croizet *et al.* [1990]. Alternative formulas are also discussed by Heiden *et al.* [1993]. Originally in eq. (3-1) the exponent γ is 1, and the atomic lipophilic contributions f_i decay with the distance $\|\mathbf{r}\text{-}\mathbf{r}_i\|$. Actually, one can think f_i decays with a higher power, $\|\mathbf{r}\text{-}\mathbf{r}_i\|^{\gamma}$. Later I shall discuss the effects of γ. All earlier MLP models have been empirical, meaning they depend on an empirical parameter set of fragmental or atomic lipophilicity indices [Ghose and Crippen 1986, Viswanadhan *et al.* 1989]. A review can be found in Chapter 2, §2.6.3.

3.1.1 Role of Molecular Lipophilicity in Drug Design

In rational ligand design, it is becoming accepted that consideration should be given to a combination of all three types of molecular interactions: steric, electrostatic, and hydrophobic factors [Bone and Villar 1995]. Each of these factors plays its part in deciding the optimum arrangement of a ligand in a binding site. The steric factor is readily assessed by a number of methods, for example, the intersection volume of a set of related molecules [Tokarski *et al.* 1994, Meyer and Richards 1991, Masek *et al.* 1993] or 'sterimol' parameters [Verloop *et al.* 1976]. The importance of the molecular electrostatic potential (MEP) in long-range ligand-receptor interactions has long been recognized [Weinstein 1975, 1981]. MEP is readily evaluated and visualized by computing its

distribution at points on the van der Waals surface using classical or quantum mechanics. Specifically the positions, magnitudes, and number of its maximum and minimum, have often been used in rationalizing the relative activities of ligands for a given receptor [Hopfinger 1983].

As illustrated in Chapter 1, §1.4.2 and §1.4.3, molecular lipophilicity plays an important role in the diffusion and binding of drug molecules to their biological target. Great efforts have been made to add the lipophilicity factor to rational drug design during the last three decades. However, due to the complex dependence of molecular lipophilicity on the chemical and physical nature, so far there is no theoretical method for the measurement of molecular lipophilicity.

3.1.2 Lipophilic Potential Energy Field in CoMFA

Comparative molecular field analysis (CoMFA) has been introduced in QSAR and drug design since 1988 [Cramer *et al.* 1988, 1996], however, the pioneering works on CoMFA date back to the 1960's. CoMFA has become one of the most powerful tools in drug design and has pioneered a new paradigm of three-dimensional QSAR where the shapes and properties of molecules are related to specific molecular features (substitutes, *etc.*) and their spatial relationships. Thus molecular modification to improve biological activity based on QSAR can be rooted in the actual chemistry of the involved molecules [Waller and Kellogg 1996]. In the applications of CoMFA, there are a variety of ways to supplement more information to the model by modifying the energy field set and, in this way, CoMFA is very successful in computer-aided drug design.

The standard potential energy field, in its native form, is a steric and electrostatic potential field. The probe atom of standard CoMFA is an sp^3 hybridized carbon atom with an effective radius of 1.53 Å and +1.0 charge. The probe atom to ligand atom distance-dependence of the potential functions are the standard 6-12 Lennard-Jones potential and *r*-square term of the Coulombic potential, resulting in steep changes as the probe nears the surface of a molecule. Both steric and electrostatic energies have to be truncated at some

arbitrary level to eliminate points within the van der Waals shell.

While the steric and electrostatic properties of molecules are the major physicochemical properties related to biological activity, they are purely enthalpic interactions. In many cases additional properties of molecules should be introduced on a three-dimensional basis. Molecular lipophilicity, the entropic property, is one of these types of properties and should be included in the CoMFA framework. In Chapter 2, §2.6.3, I reviewed the empirical molecular lipophilicity potential (EMLP) developed by Audry *et al.*[1986, 1989] and Furet *et al.* [1988], and which depends on an empirical parameter set of atomic lipophilic indices. Kellogg and Abraham have contributed significantly to including the EMLP field in the CoMFA approach [Kellogg and Abraham 1992]. So far, all attempts at introducing 3-dimensional lipophilicity potential into the CoMFA are based on the EMLP. A non-empirical lipophilicity potential is keenly needed in computer-aided drug design.

3.2 Heuristic Molecular Lipophilicity Potential

In Chapter 1, §1.2, I mentioned that MEP is the best physical quantity used in the study of molecular interactions [Tomasi 1981, Tasi and Pálinkó 1995]. MEP has been used successfully by many authors in the study of electrostatic interactions such as hydrogen bonds, dipolar moment interaction, and in the prediction of biological active sites [Politzer 1981]. In fact, all types of molecular interactions originate from electrostatic interactions. The goal of my thesis research is to establish a unified lipophilicity and hydrophilicity potential model based on MEP.

3.2.1 Unified Lipophilicity and Hydrophilicity Measurement System

Usually, molecular lipophilicity and hydrophilicity are two different properties, having different physical and chemical natures. However, it is extremely useful to unify these two properties in one measurement system. In this system, lipophilicity and hydrophilicity are two ends of one phenomenon and such a measurement system is very beneficial to both theoretical and experimental chemistry. In Fig. 3-1, I suggest a unified

lipophilicity and hydrophilicity measurement system in which I follow the conventions of empirical MLP [Croizet *et al.* 1990, Audry *et al.* 1986, Audry *et al.* 1989 a, b] and atomic lipophilicity indices [Ghose and Crippen 1986, Viswanadhan *et al.* 1989], where the positive values are used for lipophilicity, and the negative values are used for hydrophilicity.

In the unified measurement system, the lipophilicity effect has its original meaning—an entropy-dominated effect caused by the reorientation or reconstruction of water molecules around non-polar parts of a solute molecule. However, hydrophilicity effects include the interactions of dipole and multipole moments, charged atoms in a molecule, and hydrogen bonding. In other words, hydrophilicity includes most types of electrostatic interactions. Therefore, a unified lipophilicity and hydrophilicity measurement system is also a unified lipophilic and electrostatic measurement system.

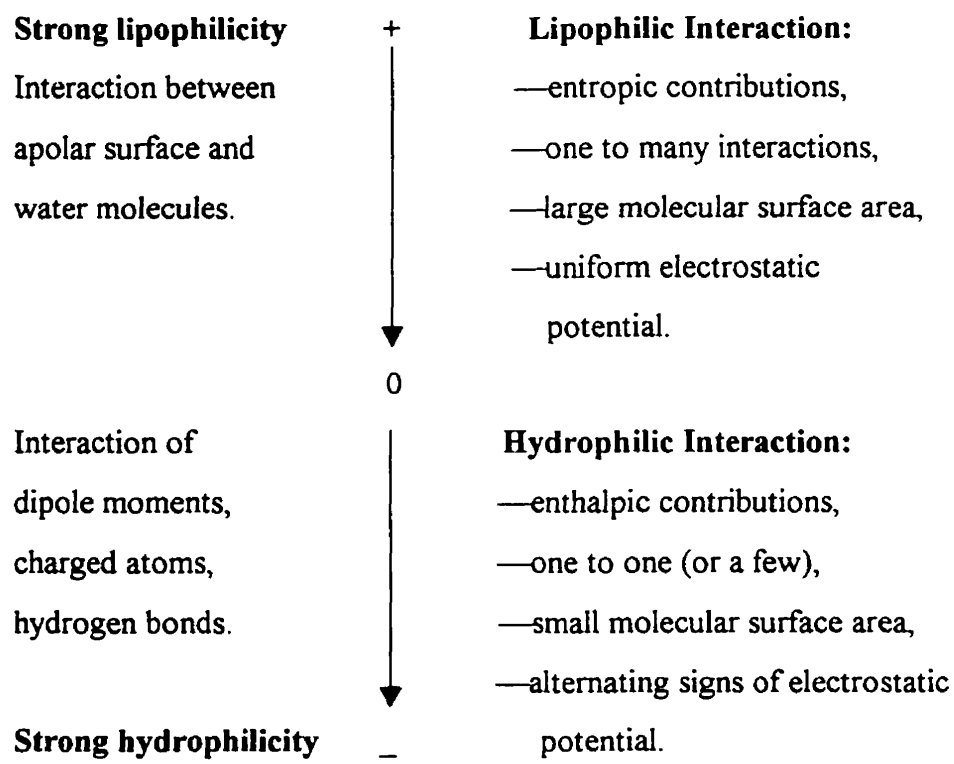


Figure 3-1. A measurement system for unified lipophilicity and hydrophilicity potential.

In Fig. 3-1, while the positive values are used for the lipophilicity, the negative values are used for hydrophilicity. This is not only the convention used in the empirical MLP and atomic lipophilic indices, but it also has a theoretical background. The logarithm of partition coefficient between water and octanol, $\log P_{ow}$, has been used as an overall measure of molecular lipophilicity. The partition coefficient is connected with transfer free energy, and is assumed to be the sum of the contributions from all constituent fragments or atoms,

$$\log P_{ow} = \log \frac{C_{organic}}{C_{water}} = \frac{-\Delta G^{\circ}_{tr}}{2.303RT} = \sum_i n_i f_i . \qquad (3\text{-}2)$$

It is obvious that if a compound is a "water-lover", there is a higher concentration in the water phase than in the organic phase, and log P_{ow} should be negative. Otherwise, if a compound is an "oil-lover", there is a higher concentration in the organic phase than in the water phase, and $\log P_{ow}$ should be positive. This means that the hydrophilic group has a negative contribution ($f_i<0$) to log P_{ow}. On the other hand, the lipophilic group has a positive contribution ($f_i>0$) to $\log P_{ow}$. These conventions are consistent with the Chinese traditions of *YIN* and *YANG*. Lipophilicity, meaning dry, corresponds to *YANG*, and is positive; on the other hand, hydrophilicity, meaning wet, coresponds to *YIN*, and is negative. It should be pointed out that in the data set of empirical atomic lipophilic indices [Viswanadhan *et al.* 1989], some hydrophilic atoms, such as oxygen in OH (phenol, enol, and carboxyl), get positive values (f_O=0.5212). This is unreasonable. The reason may be that it is from the least square regression calculation. The signs of atomic empirical lipophilic indices are assigned by the partial least square (PLS) regression calculations. Sometimes the signs assigned by PLS lose their physical meaning. The empirical atomic lipophilic indices are good for the predictions of log P_{ow} of new compounds in the viewpoint of statistics, however, they are not good for molecular lipophilicity potentials described by eq. (3-1).

3.2.2 Distributions of Charge and MEP on Molecular Surface

Chemists often think that a molecule consists of charged atoms. A simple picture of molecular interactions is that of positively charged atoms in one molecule attracting the negatively charged atoms and repulsing the positively charged atoms in other molecules. In some cases, for example, in the qualitative studies of hydrogen-bonding, dipole interactions, nucleophilic and electrophilic attacks, this is a good approximation. In these cases, the atoms under consideration are strongly charged. However, this model is not always true. Particularly, in the study of molecular lipophilicity, this simple approximation may present the wrong picture.

Here I show an example of charge and MEP distribution on the pentanoic acid molecule in Table 3-1. The data in Table 3-1 are calculated by Gaussian 92 at the level RHF/6-31G*. Geometry is optimized at the RHF/STO-3G level, and the molecular surface is the van der Waals fused sphere surface. In atomic units, the charge is the electron charge, e=1.6021×10^{-19} coulomb, the length is the bohr a_0=5.2917×10^{-9} cm or 0.52917 Å, and the energy is the hartree, 627.525 kcal/mol or 2625.6 kJ/mol. In Table 3-1, S_{total} represents the exposed atomic total area on the molecular surface. S^+ and S^- are the surface areas of positive and negative MEP, respectively, and b^+ and b^- are atomic positive and negative MEP-descriptors defined by the following equations [Du and Arteca 1997],

$$b_i^+ = \sum_k V^+(\mathbf{r}_k)\Delta s_k \ , \tag{3-3}$$

$$b_i^- = \sum_k V^-(\mathbf{r}_k)\Delta s_k \ , \tag{3-4}$$

$$b_{total} = b_i^+ + b_i^- \ . \tag{3-5}$$

Table 3-1. Atomic charges, surface areas, and surface-MEP-descriptors of pentanoic acid. Calculated by Gaussian 92 at the level RHF/6-31G*, in atomic units.

Atom	q_i	S^+	S^-	S_{total}	b^+	b^-	b_i
C_1(COOH)	0.3099	34.48	29.73	64.21	0.4490	-0.6028	-0.1538
=O	-0.2722	0	35.54	35.54	0	-2.3205	-2.3205
O (OH)	-0.2951	3.064	26.65	29.72	0.0374	-0.9528	-0.9154
H (OH)	0.2013	30.34	0	30.34	1.668	0	1.6682
C_2	-0.1262	43.78	4.196	47.98	0.8951	-0.0346	0.8605
H	0.0738	13.73	0	13.73	0.4571	0	0.4571
H	0.0738	13.73	0	13.73	0.4572	0	0.4572
C_3	-0.0948	26.82	16.78	43.60	0.3817	-0.3170	0.0647
H	0.0596	12.38	0.6037	12.98	0.1868	-0.0013	0.1855
H	0.0596	12.38	0.6037	12.98	0.1870	-0.0013	0.1857
C_4	-0.0941	47.98	0	47.98	0.6223	0	0.6223
H	0.0527	13.73	0	13.73	0.3154	0	0.3154
H	0.0527	13.73	0	13.73	0.3156	0	0.3156
C_5	-0.1776	69.68	16.97	86.65	0.4647	-0.0174	0.4473
H	0.0592	13.13	0	13.13	0.2446	0	0.2446
H	0.0587	12.98	0	12.98	0.2326	0	0.2326
H	0.0587	12.98	0	12.98	0.2332	0	0.2332

Atomic charge q_i is the sum of nuclear charge Z_i and electronic charge $q_i^{(e)}$ on the atom i, $q_i = Z_i + q_i^{(e)}$. In quantum chemistry, electronic charges are usually obtained based on Mulliken population analysis. As shown in the definition equation of MEP, eq. (1-5), there are two different contributions to $V(\mathbf{r})$: the contributions of nuclear charges and of electron density $\rho(\mathbf{r})$. There is the possibility that on the surface of a negatively charged carbon, the $V(\mathbf{r})$ is positive. As shown in Table 3-1, except C_1, which is in the carboxyl group -COOH, all other carbons have negative net atomic charges q_i. However, except for

C_1, all other carbons have positive MEP-descriptors b_{total}. All hydrogen atoms on the hydrocarbon chain have positive b_{total}, too.

This example shows that for the lipophilic hydrocarbon chain, net atomic charges and MEP-surface-descriptors tell us different stories. In a hydrocarbon chain, both carbon and hydrogen atoms are weakly charged and have a weak interaction with water molecules. If one thinks that carbon and hydrogen atoms are negatively and positively charged, alternating in a lipophilic hydrocarbon chain, the structure of the water molecules surrounding this lipophilic surface is the same as surrounding the hydrophilic surface: water molecules interact with the surface through the positive end and negative end of the molecule alternately oriented, as shown in Fig. 3-2 (a). However, based on the surface-MEP descriptors, the b_{total}'s of both carbon and hydrogen atoms in a hydrocarbon chain are positive, and water molecules arrange themselves tangentially to the lipophilic surface, as shown in Fig. 3-2 (b). If one uses the MEP-equivalent formal charges, one will find that both carbons and hydrogens in the lipophilic hydrocarbon chain have positive formal charges. Therefore, the quantum quantity MEP is more reliable than classical quantity atomic net charges in the study of molecular interactions.

3.2.3 Heuristic Molecular Lipophilicity Potential

As mentioned in Chapter 1, MEP is the best physical property for the description of molecular interactions. However, so far, it has not been successfully used to describe lipophilicity, as in the studies of hydrogen-bonds [Murray *et al.* 1991 a, b, Gao 1994, Mishra and Kumar 1995] and nucleophilic and electrophilic attacks, though a number of promising efforts have been made by some authors [Náray-Szabó and Nagy 1989, Náray-Szabó 1989, Náray-Szabó 1986]. The reason for this is that molecular lipophilicity is an entropy-dominated phenomenon, dealing with the interactions of huge numbers of water molecules, and cannot be illustrated based only on the MEP distributions on individual atoms, unlike hydrogen-bonding, where the maximum and minimum of MEP are sufficient for a qualitative description. However, the description of lipophilicity needs a large microscopic volume element.

The lipophilic effect is mainly dominated by a negative entropic change of water molecules. Usually, chemists think that interactions of dipole and multipole moments, charged atoms, dispersions, polarization and hydrogen-bonding are electrostatic interactions, and that lipophilic interactions have a non-electrostatic origin that is entropy driven [Isaelachvili 1992]. In the macroscopic point of view, this is true, however, in the microscopic point of view, the lipophilic interaction also originates from electrostatic interactions.

The studies of rare gases and hydrocarbons in water show that although ΔH of the solution is negative, such compounds are insoluble [Leo *et al.* 1971, Israelachvili 1992]. The reason is the large negative change of ΔS in this process [Dogonadze *et al.* 1985, Frank and Evans 1945, Israelachvili 1992]. According to Frank and Evans [1945], when organic compounds are placed in water, the water molecules arrange themselves around the non-polar parts in what was termed "iceberg" structures. The arrangement of water molecules in the hydration shell is more ordered than in the bulk of solvent, in a manner analogous to that of ice. However, the density in the hydration shell is higher than in the bulk, which is not true for ice [Leo *et al.* 1971]. Frank and Evans made their conclusion 50 years ago, however, all recent studies support their conclusion [Forsman and Jönsson 1991, Guillot *et al.* 1991]. The investigative results of Monte Carlo simulations and molecular dynamic studies provide a good insight into the structure of the hydration shell [Forsman and Jönsson 1991, Guillot *et al.* 1991, Tanaka and Nakanishi 1991, Haile 1992, Smith and Haymet 1993]. These studies show that in the hydration shell, water molecules are arranged in a more ordered manner than in the bulk of the solvent.

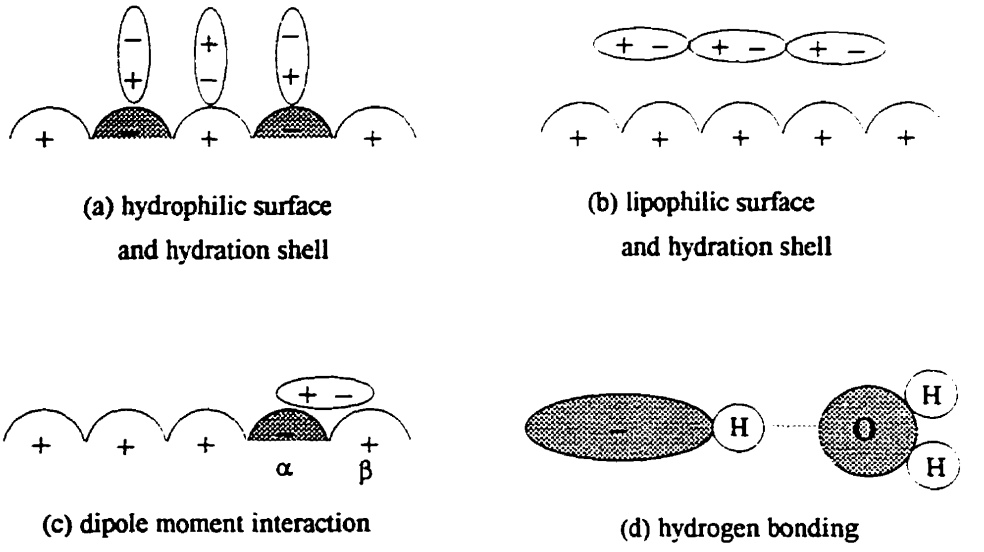


Figure 3-2. (a) Hydrophilic surface and hydration shell, (b) Lipophilic surface and hydration shell, (c) Local dipole moment, and (d) Hydrogen-bonding. Signs "+" and "-" are for MEP, not for charge.

Fig. 3-2 shows four types of interactions between various molecular surfaces and water molecules, where signs "+" and "-" stand for positive and negative MEP, respectively. In Fig. 3-2 (a), the hydrophilic surface consists of atoms with alternating MEP, such as may be expected on silica and fiber surfaces. In the hydration shell around the hydrophilic surface, water molecules bind to the surface in an energy-favorable way: the positive ends and negative ends of water molecules stick on the surface alternatively. This arrangement is similar to the structure of water molecules in the bulk of aqueous solution. Fig. 3-2 (b) shows the lipophilic surface, which consists of atoms with uniform MEP of the same sign [Du and Arteca 1996]. In the hydration shell surrounding the lipophilic surface, water molecules are placed tangentially to the lipophilic surface. This is more favorable for energy than having water molecules arrange themselves parallel and perpendicular to the surface, resulting in strong repulsive interactions between water molecules. However this structure is unfavorable with respect to entropy. This model of lipophilic surface and the structure of the hydration shell is supported by Monte Carlo simulation conducted by Guillot *et al.* [1991]. They find that in the hydration shell around

a lipophilic surface, "*water molecules are arranged, on average, tangentially to the (lipophilic) solute molecule*". Another molecular dynamics simulation conducted by Lee and Rossky [1994], using a tetrahedral ST2 water model, shows that "*a typical water molecule at the (lipophilic) surfaces has one potentially hydrogen-bonding group oriented toward the hydrophobic surface*". On every face of a cubic ST2 water model, there should be two hydrogen-bonding elements (donor or acceptor). Therefore water molecules avoid their hydrogen-bonding elements facing the lipophilic surface. Fig. 3-2 (c) shows a local dipole moment on the molecular surface where the atom α is bordering another atom β with the opposite MEP. As shown in Fig. 3-2 (d), hydrogen bonding is observed in circumstances where a hydrogen atom covalently bonds to an electronegative atom, such as oxygen, nitrogen or a halogen. It is clear that at a point **r** on the molecular surface, the lipophilicity potential is decided not only by the atom the point **r** belongs to, but also by the molecular environment.

A complete theoretical derivation of the free energy change from first principles, explicitly including a huge number of water molecules, is not easy. No rigorous theoretical method is available for this task. However, a heuristic lipophilicity potential model is likely to be sufficient for some tasks in molecular modeling and chemical design. Here I suggest a unified lipophilicity and hydrophilicity potential model, as follows:

$$\underset{\mathbf{r}\in S_\alpha}{L(\mathbf{r})} = V(\mathbf{r})\sum_{i\neq\alpha} M_i(\mathbf{r};\mathbf{R}_i, b_i), \tag{3-6}$$

where $V(\mathbf{r})$ is MEP at the point **r**, and **r** is on the surface S_α of atom α. In the sum, $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$ is the screening function on position **r** from atom *i*. In eq. (3-6) the summation is over all constituent atoms except atom α. In the screening function $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$, $\mathbf{R}_i$ is the nuclear position of atom *i*, and b_i is the atomic surface-MEP descriptor of atom *i* [Du and Arteca 1997],

$$b_i = \sum_{k\in S_i} V(\mathbf{r}_k)\Delta s_k\ , \tag{3-7}$$

where Δs_k is the area element on the surface of atom i. Summation is over all exposed surfaces of atom i.

Hydrophobic effects are complex phenomena. The term usually refers to both hydrophobic hydration (HH) and hydrophobic interaction (HI) [Forsman and Jönsson 1994, Israelachvili 1992, Head-Gordon 1995]. HH concerns the thermodynamic and structural changes that are associated with the solvation of a non-polar solute in water [Guillot *et al.* 1991, Tanaka and Nakanishi 1991], and is conveyed by thermodynamic properties: free energy, enthalpy, and entropy of hydration. HI refers to the interactions between two organic molecules dissolved in an aqueous solution [Smith and Haymet 1992, Israelachvili 1992]. HI is more useful in chemistry and HH is the basis for the understanding of the nature of HI and for making qualitative and quantitative predictions of HI. Heuristic molecular lipophilicity potential (HMLP) is a tool for the study of HH, not directly for the study of hydrophobic force law, however, it can be used in the research of HI indirectly.

3.2.4 Screening Function in HMLP

The screening function is the center of HMLP. Atom-based screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ can take a number of forms. In Chapter 4, I will do a detailed selection of screening functions and optimization of parameters used in the screening functions. Here I just discuss the properties and physical meaning of screening function using a power distance-dependent function,

$$M_i(\mathbf{r}; \mathbf{R}_i, b_i) = \frac{r_0^\gamma}{b_0} \frac{b_i}{\|\mathbf{R}_i - \mathbf{r}\|^\gamma} = \varsigma \frac{b_i}{\|\mathbf{R}_i - \mathbf{r}\|^\gamma} . \qquad (3\text{-}8)$$

In eq. (3-8), r_0, b_0, and γ are parameters. The unit of b_0 is the same as b_i (energy•area); r_0 has a unit of length. Therefore, $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ is a dimensionless function. In eq. (3-8), $\zeta=(r_0)^\gamma/b_0$ is a simple scaling factor. In later calculations, I take $\zeta=1$. Exponent γ in eq. (3-

8) is the parameter that decides how strong the influence is and how rapidly the influence decays with distance. In this section, I test the power distance-dependent screening function, eq. (3-8), and optimize parameter γ based on the experimental partition coefficients $\log P_{ow}$. A more careful and complete optimization will be done in Chapter 4.

The heuristic MLP defined by eq. (3-6) and the properties of screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ can be interpreted as follows:

(1) Lipophilicity potential $L(\mathbf{r})$ is an average, or modified, electrostatic potential. $\Sigma_{i\neq a} M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ is the modifying factor representing the influence from all surrounding atoms at point $\mathbf{r}$.

(2) If $\Sigma_{i\neq a} M_i(\mathbf{r}; \mathbf{R}_i, b)$ has the same sign as $V(\mathbf{r})$, then the lipophilicity potential $L(\mathbf{r})$ is positive, and point $\mathbf{r}$ is lipophilic. Whereas, if $\Sigma_{i\neq a} M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ has a sign opposite to that of $V(\mathbf{r})$, then the lipophilicity potential $L(\mathbf{r})$ is negative, and point $\mathbf{r}$ is hydrophilic.

(3) The influences from all other atoms decay with the distance $||\mathbf{R}_i-\mathbf{r}||$. In eq. (3-8), γ is the parameter that decides how strong the influence is and how rapidly the influence decays with distance.

(4) Atomic surface-MEP descriptor b_i defined by eq. (3-7) represents the MEP distribution on the surface of atom i. The effect of b_i is similar to the empirical lipophilic parameters f_i in eq. (3-1), however, b_i's are theoretical structural parameters, not empirical parameters.

(5) It is best to think of $\Sigma_{i\neq a} M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ as a unitless modifying factor. Therefore, the unit of lipophilicity potential $L(\mathbf{r})$ is the same as that of the electrostatic potential $V(\mathbf{r})$. However, there is no direct, simple connection between the values of $L(\mathbf{r})$ and the free energy of the solution system.

As discussed in Chapter 1, §1.2.2, the electrostatic potential $V(\mathbf{r}_A)$ is a measure of the interaction ability of atom A in the solute molecule with a water molecule at reference position $\mathbf{r}_A$, and $V(\mathbf{r}_B)$ is the measure of the interaction ability of atom B in the solute molecule with another water molecule at reference position $\mathbf{r}_B$. If $V(\mathbf{r}_A)$ and $V(\mathbf{r}_B)$ have the same sign, the interaction between two water molecules is repulsive (lipophilic);

otherwise, the interaction between two water molecules is attractive (hydrophilic). The effect of atom B on the interaction ability of atom A with a water molecule decays with distance $\mathbf{R}_{AB}$. In the model of HMLP, there are no water molecules involved explicitly, however, through the screening function, the effects of water molecules are considered in this model implicitly.

3.2.5 Limitations of HMLP

The heuristic lipophilicity potential developed in this research is a structure-based potential. In the calculation of the heuristic MLP based on eqs. (3-6) to (3-8), the input data are molecular geometries and molecular surfaces. This technique can be implemented for realistic contour surfaces determined from *ab initio* electron densities [Mezey 1990] and can also be used for various empirical molecular surfaces. In the case of a van der Waals's surface, it is a fused-sphere surface. HMLP is rooted in *ab initio* quantum chemistry. Quantum chemical methods are used to calculate the electron density function $\rho(\mathbf{r})$ and the electrostatic potential $V(\mathbf{r})$ following the definition eq. (1-5). Therefore the applications of HMLP in practical drug design and QSAR studies are limited by the ability of the *ab initio* quantum chemical approach. Maybe someone suspects that the HMLP may be too cumbersome to be taken up as a widely used tool in practical studies.

For the shape analysis of the heuristic lipophilicity potential, shape group methods are applicable [Mezey 1993, 1990, 1986]. Whereas these methods are designed for small molecules, a new technique for the calculation of electron density functions $\rho(\mathbf{r})$ of large biomolecules was developed by Walker and Mezey [1994, 1993]. This technique, the molecular electron density Lego assembler (MEDLA), a so-called "computational microscope" [Borman 1995], has been used to construct *ab initio* quality electron densities at the 6-31G** level for proteins containing more than 1,000 atoms. The macromolecular density matrix method ADMA, the Adjustable Density Matrix Assembler [Mezey 1995 a, b], is a method that no longer needs an extensive numerical density data base, and appears advantageous for MEP applications. Therefore, there is no insurmountable difficulty for the application of HMLP in drug design and QSAR studies.

In the HMLP method there are no empirical parameters of atomic lipophilicity indices used , therefore it is a non-empirical MLP. However, it is not a rigorous theoretical model because not all aspects of this model are derived from first principles. Therefore, I regard it as a heuristic model.

3.3 Simple Examples and Tests of HMLP

In this section, I show simple examples and the tests applied to the calculation of results for several small molecules: ethanol (C_2H_5OH), *n*-propylamine ($C_3H_7NH_2$), and *n*-propanoic acid (C_2H_5COOH). In the following calculations, the screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ takes the form of eq. (3-8).

3.3.1 Atomic and Molecular Lipophilicity Indices

Based on the definition of HMLP, eq. (3-6), the atomic lipophilic index l_a can be defined as follows:

$$l_a = \sum_{i \in S_a} L(\mathbf{r}_i)\Delta s_i \ , \tag{3-9}$$

where the summation is over all the exposed area, S_a, of atom a. If $l_a > 0$, then atom a is lipophilic, whereas, if $l_a < 0$, then atom a is hydrophilic. The molecular lipophilic index (L_M) and the hydrophilic index (H_M) are the sum of the corresponding values for all lipophilic atoms and hydrophilic atoms, respectively,

$$L_M = \sum_{\substack{a \\ (l_a>0)}} l_a \ , \tag{3-10}$$

$$H_M = \sum_{\substack{a \\ (l_a<0)}} l_a \ . \tag{3-11}$$

In some cases, for convenience, one also can define the lipophilic and hydrophilic indices for molecular fragments or functional groups. These indices are very useful in the description of molecular local lipophilicity and for checking the reasonableness and validity

of HMLP.

Table 3-2. Heuristic and empirical lipophilicity indices of ethanol.

Atoms		Heuristic Indices l_a (hartree bohr^2)	Empirical Indices $l_a^{(cm)}$ (no unit)
(-CH_2-)	C_1	0.07855	2.839
	H	0.00667	0.9255
	H	0.00668	0.9248
(-CH_3)	C_2	0.02297	11.50
	H	0.00318	2.068
	H	0.00057	1.777
	H	0.00058	1.798
(-OH)	O	-0.1722	0.1106
	H	-0.0707	-1.367
	L_M	0.1192	21.94
	H_M	-0.2429	-1.367

Table 3-2 shows the atomic lipophilic indices l_a of ethanol. Molecular geometry is optimized at the Hartree-Fock level with the STO-3G basis set, and a fused sphere van der Waals's surface [Du and Arteca 1996] is used (R_C=2.00 Å, R_H=1.17 Å, R_O=1.39 Å). Electron density is calculated by the Gaussian 92 program at the RHF/6-31G* level. A surface grid point density is chosen as 25 point/ Å^2 [Connolly 1983 a, b, 1985]. In this example, the exponent in eq. (3-8) is taken as γ=2.50. In Table 3-2, the empirical lipophilic indices of ethanol are calculated using the same equations (3-9), (3-10), and (3-11), based on the empirical lipophilicity potential eq. (3-1) [Croizet *et al.* 1990] and using empirical atomic lipophilic indices [Viswanadhan *et al.* 1989]. As expected, the calculations of HMLP show that the hydrocarbon part takes positive values (lipophilic), and the hydroxyl

group takes negative values (hydrophilic). Oxygen is the most hydrophilic atom, and C_1 is the most lipophilic atom in ethanol. However, in the empirical calculations, the atomic lipophilic index of oxygen is (surprisingly) positive (lipophilic), and the molecular lipophilic index L_M is much higher than the hydrophilic index H_M. Based on chemical intuition, the empirical lipophilicity potential result for ethanol is not satisfactory.

3.3.2 Effects of Point Density and Basis Sets

Table 3-3 gives the comparison using different basis sets and surface point densities in the calculation of ethanol. All other conventions are the same as in Table 3-2. As shown in Table 3-3, there is not much difference using point densities 10, 25, and 50 point/Å^2. All 4 types of basis sets give qualitatively the same results, however, when the poor basis set (STO-3G) is used, the numerical results are very different from the others. The basis set in this type of calculation should include polarization functions.

Table 3-3. Comparison of different basis sets and surface point densities using ethanol as an example. Results in atomic unit: hartree bohr^2

Point Density	Basis Set	Total Points	Index l_O	Index l_H	Index L_M	Index H_M
10	6-31G*	738	-0.1729	-0.0717	0.1125	-0.2446
25	6-31G*	1730	-0.1722	-0.0707	0.1192	-0.2429
50	6-31G*	2483	-0.1737	-0.0718	0.1179	-0.2456
25	4-31G*	1730	-0.1689	-0.0714	0.1144	-0.2402
25	3-21G*	1730	-0.1940	-0.0877	0.1255	-0.2825
25	4-31G	1730	-0.2253	-0.0896	0.1566	-0.3149
25	STO-3G	1730	-0.0964	-0.0426	0.0634	-0.1396

3.3.3 Effects of Exponent and Atomic Radii

The exponent γ in eq. (3-8) is an important parameter. Table 3-4 summarizes the results of calculations for *n*-propylamine using different values of γ. A point density of 25,

and the basis set 6-31G* are used. Atomic radii are: R_C=2.00 Å, R_H=1.17 Å, and R_N=1.46 Å. Molecular geometry is optimized at the STO-3G level. In the amino group -NH_2, there are two positively charged hydrogen atoms and one negatively charged nitrogen atom, therefore, the lipophilic indices of the two hydrogen atoms are very sensitive to the exponent γ. As shown in Table 3-4, when $\gamma \leq 2.00$, l_H is positive, for $2.40 \leq \gamma \leq 3.00$, l_H is negative. Beyond 2.40, l_H remains essentially constant, however, l_N decreases remarkably. In further calculations we have used the value γ=2.50. However, this is not a rigorous optimization, and γ=2.50 may not be the best value. In Chapter 4 a more careful optimization will be presented.

Table 3-4. Test calculations of the exponent γ using *n*-propylamine.
In atomic unit: hartree $bohr^2$.

Exponent γ	l_N (-NH_2)	l_H (-NH_2)	L_M	H_M
1.00	-1.3227	0.1663	1.8444	-1.3227
1.50	-0.6073	0.0384	0.8151	-0.6073
2.00	-0.2909	0.01326	0.3762	-0.2909
2.40	-0.1658	-0.00490	0.2195	-0.1756
2.50	-0.1445	-0.00544	0.1930	-0.1554
2.60	-0.1262	-0.00571	0.1698	-0.1376
2.70	-0.1102	-0.00578	0.1497	-0.1218
2.80	-0.0963	-0.00569	0.1320	-0.1078
3.00	-0.0740	-0.00525	0.1031	-0.0845

A molecular surface can be regarded as the molecular interaction interface [Mezey 1990]. The generation of a molecular surface is a key step in this approach. Table 3-5 lists the results of calculations for *n*-propylamine using different atomic radii. Exponent γ is taken as 2.50 while all other conditions are the same as in Table 3-4. From Table 3-5, we know that for a fused-sphere van der Waals surface, atomic radii affect the atomic

lipophilic indices to a certain degree. For simplicity, in this research we use van der Waals's fused-sphere surfaces [Connolly 1983 a, b, 1985]. Atomic radii are optimized based on MEP criteria by Du and Arteca [1996]. These results are very close to the results of Ooi *et al.* [1987]. However, our optimizations didn't include all atomic types. It should be pointed out that theoretical electron isodensity surfaces [Mezey 1990] might be more reliable in this type of research. In Chapter 6 I will discuss this topic in more detail.

Table 3-5. Comparison of using different atomic radii, *n*-propylamine taken as an example. $\gamma = 2.50$. In atomic unit: hartree bohr^2.

Atomic Radii (Å)	l_N (-NH_2)	l_H (-NH_2)	L_M	H_M
R_C=2.00, R_N=1.46, R_H=1.17	-0.1445	-0.00544	0.1930	-0.1554
R_C=1.75, R_N=1.46, R_H=1.17	-0.1870	0.00637	0.3266	-0.1870
R_C=1.75, R_N=1.55, R_H=1.17	-0.1914	-0.00262	0.3047	-0.1967
R_C=2.00, R_N=1.55, R_H=1.17	-0.1452	-0.01008	0.1891	-0.1653

3.3.4 Lipophilicity of Functional Groups and Hydrogen Bonds

Table 3-6 lists the atomic lipophilic indices l_a's and the indices of functional groups (-NH_2, -OH, and -COOH) of *n*-propanol, *n*-propylamine, and *n*-propanoic acid, calculated by HMLP and EMLP. The indices of all three hydrophilic functional groups and atoms H, O, and N in these groups have negative values using HMLP. However, the indices of the hydrophilic group -COOH and atoms O in -OH of hydroxyl and carboxyl groups are positive by EMLP. This is unreasonable. The order of the indices of the three functional groups is $l_{\text{-NH2}} > l_{\text{COOH}} > l_{\text{OH}}$ if the carboxyl carbon is included in the -COOH group, and $l_{\text{-NH2}} > l_{\text{OH}} > l_{\text{COOH}}$ if the carboxyl carbon is not included in the -COOH group.

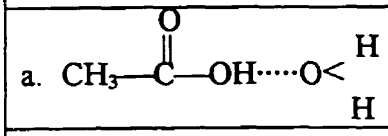
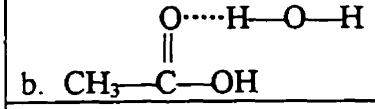
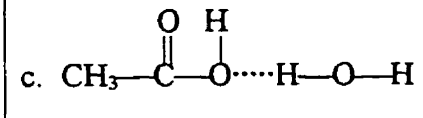
Table 3-6. Atomic lipophilic indices of functional groups (-NH_2, -OH, and -COOH)

Functional groups	Heuristic indices (hartree $bohr^2$)	Empirical indices
-NH_2 ($C_3H_7NH_2$)	l_N = -0.1445	$l_N^{(em)}$ = -2.8713
	l_H = -0.0054	$l_H^{(em)}$ = -3.6734
	l_{-NH_2} = -0.1499	$l_{-NH_2}^{(em)}$ = -6.5447
-OH (C_3H_7OH)	l_O = -0.1693	$l_O^{(em)}$ = 0.4957
	l_H = -0.0717	$l_H^{(em)}$ = -0.8826
	l_{-OH} = -0.2410	$l_{-OH}^{(em)}$ = -0.3869
-COOH (C_2H_5COOH)	$l_{=O}$ = -0.0915	$l_{=O}^{(em)}$ = -0.5734
	l_O = -0.0653	$l_O^{(em)}$ = 0.9178
	l_H = -0.1074	$l_H^{(em)}$ = -1.3931
	l_C = 0.0569	$l_C^{(em)}$ = 1.4863
	l_{-COOH} = -0.2073	$l_{-COOH}^{(em)}$ = 0.4376

Politzer and his research group have successfully studied hydrogen bonds using MEP in an extended region [Gao 1994]. The data listed in Table 3-6 can be used to study hydrogen bonds. I find that heuristic indices l_H's in the three types of functional groups are -0.0054 (NH_2), -0.0717 (OH), and -0.1074 (COOH), respectively, in reasonable accord with their hydrogen-bonding donor strength.

For carboxylic acid, the hydrogen-bond energies of three possible types of hydrogen bonds and lipophilic indices are listed in Table 3-7. The hydrogen-bond energies are taken from the results of Gao [1994]. Our results of lipophilic indices are in good agreement with the hydrogen-bond energies. Compared with other hydrogen bond indices, such as MEP [Murray *et al.* 1991 b], MFP [Mishra and Kumar 1995] and Mulliken population, one advantage of lipophilicity indices l_a' is that both hydrogen bond donors and acceptors use the same indices l_a.

Table 3-7. Hydrogen-bonding energies and lipophilic indices of carboxylic acid

	Hydrogen Bonds	6-31G(d)* (kcal/mol)	HMLP indices (atomic unit)
a. 	Donor	-8.5	l_H=-0.1052
b. 	Acceptor	-5.5	$l_{=O}$=-0.0930
c. 	Acceptor	-2.4	l_O=-0.0614

* [Gao 1994]

3.4 Partition Coefficients and HMLP

Partition coefficients logP_{ow} are experimental data. For a long time they have been used as the overall measure of molecular lipophilicity. In this section, I use various indices of HMLP to calculate partition coefficients of several families of compounds as a test of HMLP. Also, I make a comparison between heuristic MLP and empirical MLP [Furet *et al.* 1988, Audry *et al.* 1989 a, b, Audry *et al.* 1986].

3.4.1 Calculation Results

Table 3-8 lists the heuristic and empirical molecular lipophilic and hydrophilic indices of 41 molecules, including linear hydrocarbons, aliphatic amines, alcohols, and acids. The partition coefficient data, logP_{ow}, are from Leo *et al.* [1971].

Table 3-8. Molecular lipophilic and hydrophilic indices of 41 molecules.
In atomic unit: hartree bohr^2.

Hydrocarbons	L_M	H_M	$L_M^{(em)}$	$H_M^{(em)}$	Expr. logP_{ow}	Calc. logP_{ow}
CH_4	0.0128	0.0000	65.55	0.0000	1.09	1.07
C_2H_6	0.0417	0.0000	94.56	0.0000	1.81	2.03
C_3H_8	0.0474	0.0000	144.0	0.0000	2.36	2.22
C_4H_{10}	0.0643	0.0000	202.4	0.0000	2.89	2.78
C_5H_{12}	0.0803	0.0000	263.0	0.0000	3.39	3.32
C_6H_{14}	0.1028	0.0000	336.3	0.0000	3.90	4.06
C_7H_{16}	0.1155	0.0000	404.8	0.0000	5.18	5.13
C_8H_{18}	0.1348	0.0000	487.5	0.0000	____	____
C_9H_{20}	0.1492	0.0000	564.6	0.0000	____	____
$C_{10}H_{22}$	0.1684	0.0000	657.4	0.0000	____	____

(Continued)

Alcohol	L_M	H_M	$L_M^{(em)}$	$H_M^{(em)}$	Expr. logP_{ow}	Calc. logP_{ow}
CH_3OH	0.1358	-0.2350	1.316	-3.856	-0.817	-0.798
C_2H_5OH	0.1192	-0.2429	21.95	-1.367	-0.315	-0.387
C_3H_7OH	0.1577	-0.2410	42.62	-0.883	0.342	0.281
C_4H_9OH	0.1792	-0.2422	68.07	-0.4017	0.817	0.876
$C_5H_{11}OH$	0.2105	-0.2421	97.23	0.0000	1.457	1.562
$C_6H_{13}OH$	0.2338	-0.2421	129.9	0.0000	1.982	2.080
$C_7H_{15}OH$	0.2582	-0.2409	165.5	0.0000	2.542	2.504
$C_8H_{17}OH$	0.2789	-0.2408	203.6	0.0000	3.062	2.954
$C_9H_{19}OH$	0.2999	-0.2407	244.1	0.0000	____	____
$C_{10}H_{21}OH$	0.3172	-0.2405	286.0	0.0000	____	____

(Continued)

Amine	L_M	H_M	$L_M^{(em)}$	$H_M^{(em)}$	Expr. logP_{ow}	Calc. logP_{ow}
CH_3NH_2	0.06829	-0.1768	0.000	-27.53	-0.564	-0.607
$C_2H_5NH_2$	0.1394	-0.1497	5.470	-14.28	-0.253	-0.560
$C_3H_7NH_2$	0.1930	-0.1554	20.48	-11.50	0.010	0.317
$C_4H_9NH_2$	0.2434	-0.1500	40.65	-8.916	0.779	0.796
$C_5H_{11}NH_2$	0.2804	-0.1500	65.82	-7.336	1.080	1.27
$C_6H_{13}NH_2$	0.3160	-0.1499	94.83	-6.009	1.526	1.73
$C_7H_{15}NH_2$	0.3457	-0.1499	127.2	-4.859	2.183	2.12
$C_8H_{17}NH_2$	0.3738	-0.1499	162.3	-3.727	2.789	2.48
$C_9H_{19}NH_2$	0.3983	-0.1499	200.0	-2.715	____	____
$C_{10}H_{21}NH_2$	0.4217	-0.1499	240.0	-1.736	____	____

(Continued)

Acid	L_M	H_M	$L_M^{(em)}$	$H_M^{(em)}$	Expr. logP_{ow}	Calc. logP_{ow}
HCOOH	0.1227	-0.2434	0.0000	-29.57	-0.539	-0.657
CH_3COOH	0.2160	-0.2596	17.67	-3.790	-0.310	-0.019
C_2H_5COOH	0.2247	-0.2642	35.44	-1.967	0.253	-0.057
C_3H_7COOH	0.3141	-0.2634	59.69	-1.042	0.766	1.07
C_4H_9COOH	0.3593	-0.2702	87.48	-0.4851	1.411	1.41
$C_5H_{11}COOH$	0.4098	-0.2701	119.4	-0.0523	2.017	2.04
$C_6H_{13}COOH$	0.4428	-0.2701	154.0	0.0000	2.642	2.44
$C_7H_{15}COOH$	0.4773	-0.2700	191.9	0.0000	____	____
$C_8H_{17}COOH$	0.5033	-0.2700	231.5	0.0000	____	____
$C_9H_{19}COOH$	0.5299	-0.2699	274.0	0.0000	____	____
$C_{10}H_{21}COOH$	0.5517	-0.2699	317.9	0.0000	____	____

If one assumes that the relationship between $\log P_{ow}$ and lipophilic indices L_M and hydrophilic indices H_M is linear, then the following approximation can be used:

$$\log P_{ow} = C_0 + C_1 L_M + C_2 H_M . \quad (3\text{-}12)$$

The linear coefficients are determined by least square regression with the experimental log P_{ow} data. I show below the correlations restricted to families of compounds. The results include the correlation coefficient (r), the standard error (σ), the number of compounds in the regression (n), and the standard errors in the linear coefficients for the compounds.

(1) Linear hydrocarbons (n=7, r=0.995, σ=0.152):

$$\log P_{ow} = (0.6444 \pm 0.1197) + (33.28 \pm 1.519) L_M . \quad (3\text{-}13)$$

(2) Aliphatic alcohols (n=8, r=0.997, σ=0.130):

$$\log P_{ow} = (-27.01 \pm 4.85) + (22.22 \pm 0.87) L_M + (-98.71 \pm 20.29) H_M . \quad (3\text{-}14)$$

(3) Aliphatic amines (n=8, r=0.980, σ=0.282):

$$\log P_{ow} = (-7.18 \pm 2.78) + (12.94 \pm 1.45) L_M + (-32.18 \pm 16.29) H_M . \quad (3\text{-}15)$$

(4) Aliphatic carboxylic acids (n=7, r=0.978, σ=0.305):

$$\log P_{ow} = (5.53 \pm 6.81) + (12.34 \pm 2.34) L_M + (31.63 \pm 28.23) H_M . \quad (3\text{-}16)$$

For hydrocarbons, all hydrophilic indices (H_M) are zero, therefore, there is only one parameter L_M in eq. (3-13). For aliphatic alcohols the correlation coefficient (r=0.997) is very good. However, correlation coefficient cannot be used as the only criterion to judge linearity [Cassidy and Janoski 1992]. Next I use linearity plot to test the linearity between $\log P_{ow}$ and two indices L_M and H_M. The results are plotted in Figure 3-3. Based on the plots in Figure 3-3 (a) to (d), it seems that the linearity is not bad. However, linearity plots, Figure 3-3 (e) to (h), reveal that between $\log P_{ow}$ and H_M there is almost no linearity and between $\log P_{ow}$ and L_M, when the number of carbon atoms is more, the linearity is better. The linearity of $\log P_{ow}$ to L_M of amines is better than that of alcohols, cf. Figure 3-3 (f) and (h). Indices L_M's are almost a constant in a family, therefore there is no linear relationship between $\log P_{ow}$ and H_M in a family of compound.

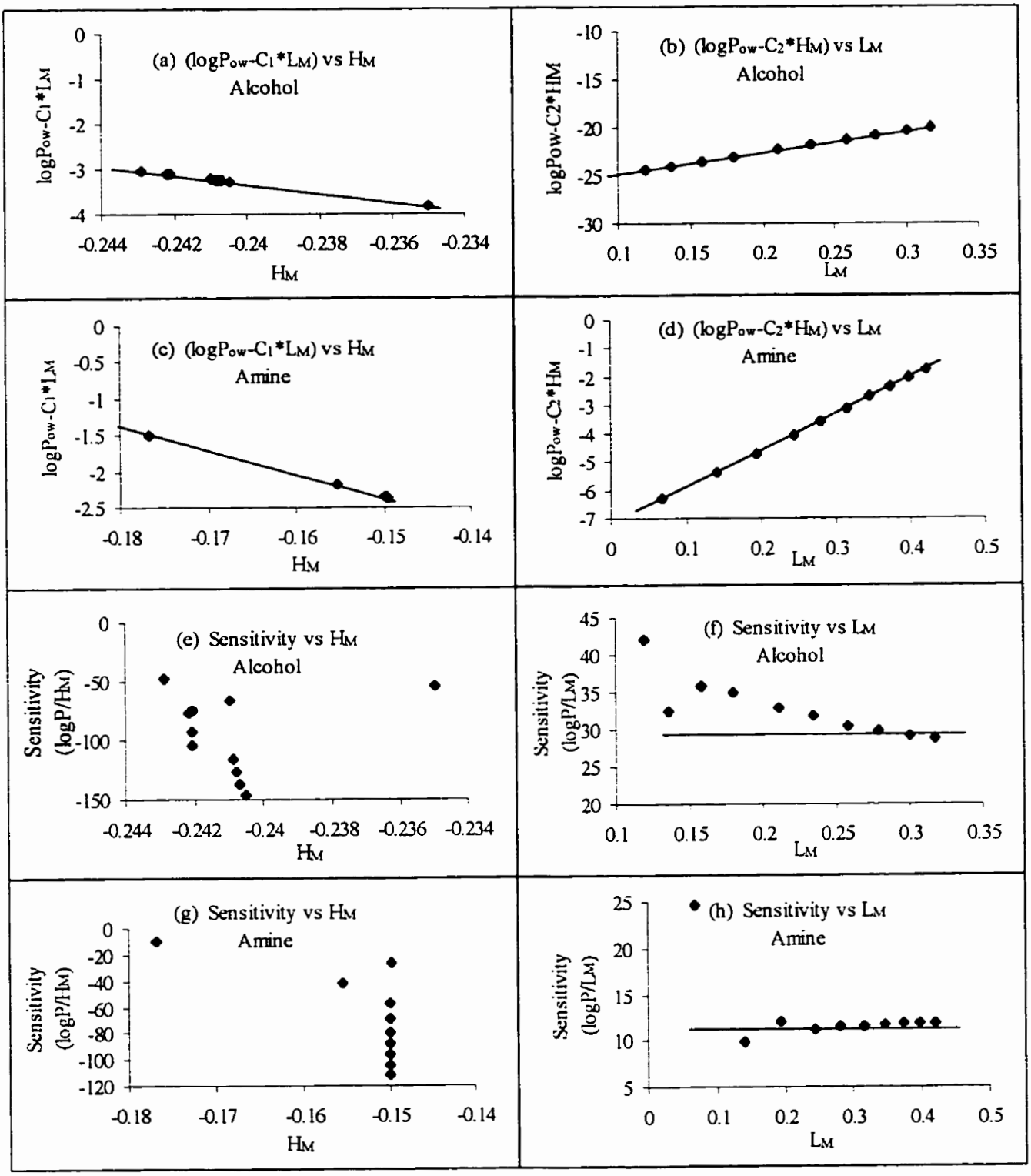


Figure 3-3. Linearity plots of aliphatic alcohols and amines. (a) to (d) are plots logP_{ow} to H_M and L_M. (e) to (h) are linearity plots, sensitivity s= logP_{ow}/H_M or L_M to H_M or L_M. If there is a linear relationship, the linearity plot should be horizontal straight line.

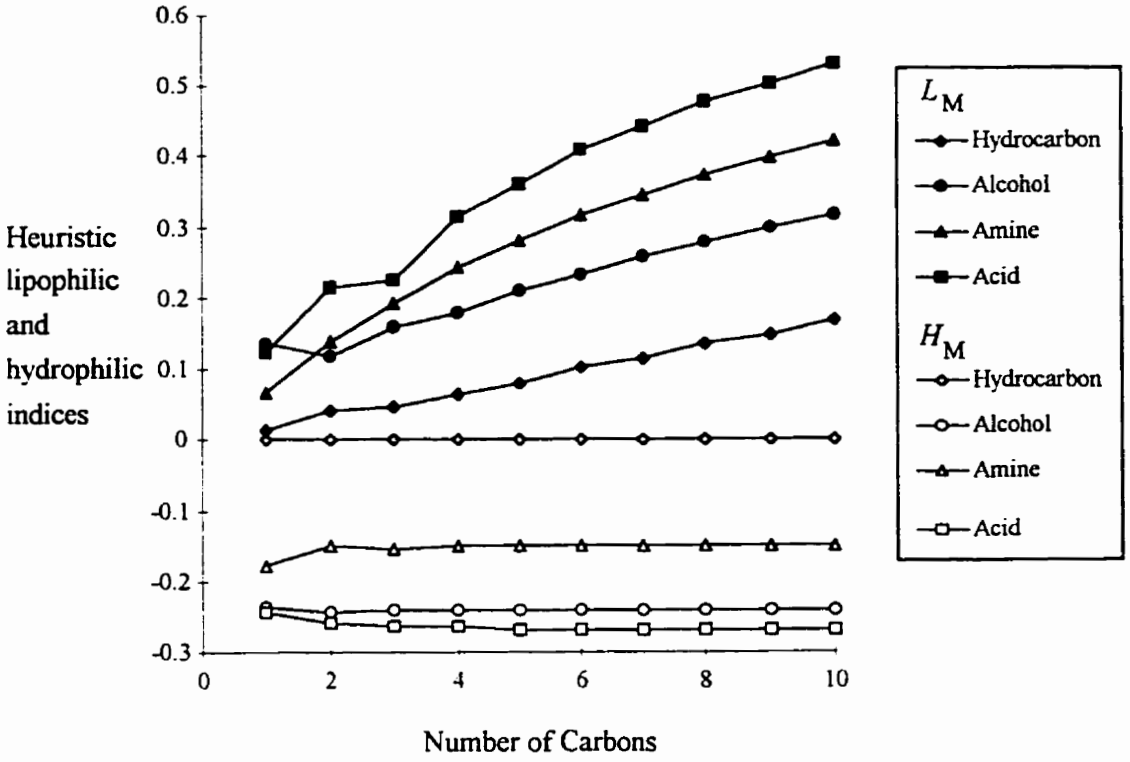


Figure 3-4. Molecular heuristic lipophilic indices L_M (top four curves) and hydrophilic indices H_M (bottom four curves) as a function of the number of carbon atoms in linear hydrocarbons, aliphatic alcohols, amines, and acids.

Molecular lipophilic indices L_M and hydrophilic indices H_M are the functions of molecular structure. Fig. 3-4 shows the behavior of L_M and H_M of HMLP as a function of the number of carbon atoms for four types of compounds (linear hydrocarbons, aliphatic amines, alcohols, and acids). In Fig. 3-4, the top four curves are the theoretical lipophilic indices L_M of four types of compounds, and the bottom four curves are the theoretical hydrophilic indices H_M. Fig. 3-5 shows the empirical indices $L_M^{(em)}$ and $H_M^{(em)}$.

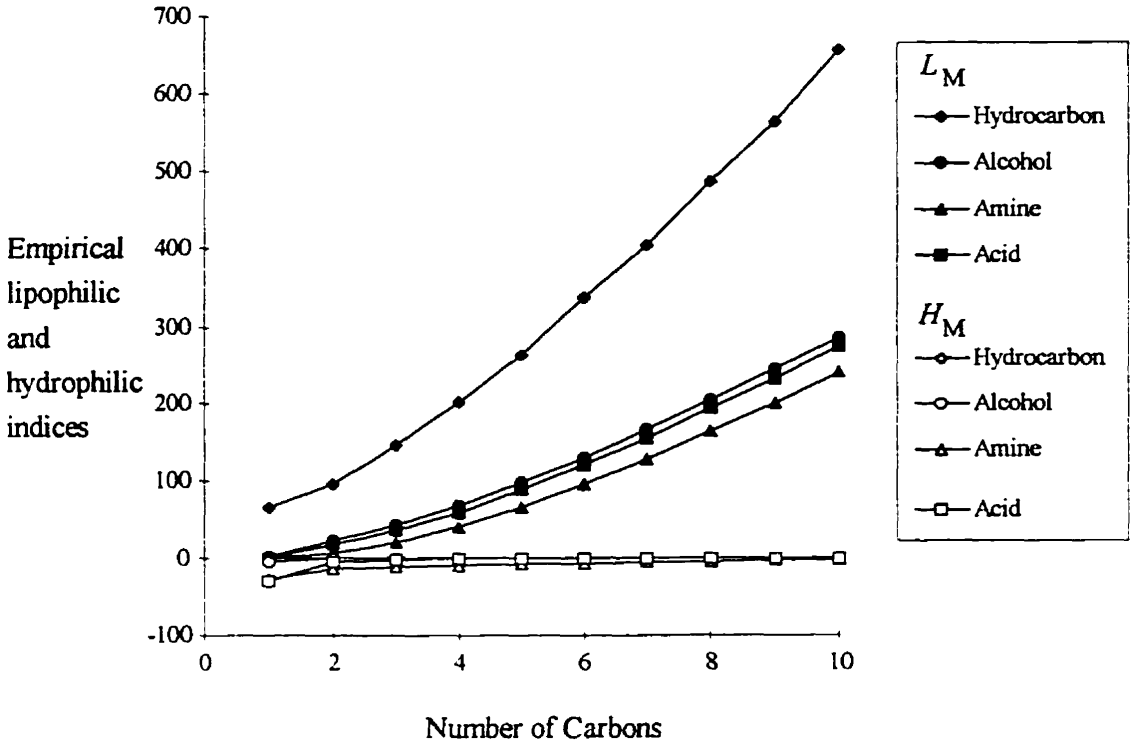


Figure 3-5. Empirically determined molecular lipophilic indices $L_M^{(em)}$ (top four curves) and hydrophilic indices $H_M^{(em)}$ (bottom four curves, some nearly coincident) as a function of the number of carbon atoms in linear hydrocarbons, aliphatic alcohols, amines, and acids.

One can see in Fig. 3-4 that for a family of compounds, the heuristic lipophilic indices L_M increase with an increase in the number of carbon atoms. However, these functions are not exactly linear. The increase becomes less noticeable as the number of carbon atoms increases in the hydrocarbon chain. On the other hand, H_M almost stays constant. The empirical indices $L_M^{(em)}$ increase sharply with increasing number of carbon atoms in the molecule. The hydrophilic index H_M should have a negative value if there is a hydrophilic group in a molecule. Indices H_M of heuristic MLP give reasonable results. Except for hydrocarbons, which have no hydrophilic groups and where all H_M's are zero, all three other types of compounds, alcohols, acids, and amines, have constant negative H_M, contributions from the hydrophilic functional groups. However, empirical MLP give unreasonable results. Only formic acid and methylamine have negative $H_M^{(em)}$. When there

are more than two carbon atoms, indices $H_M^{(cm)}$ of all acids, amines and alcohols are zero. The values of $H_M^{(cm)}$ for the three types of compounds do not follow chemical intuition. These results show that the atomic lipophilic parameters $\{f_i\}$ provided by Ghose and Crippen [1986] and Viswanadhan *et al.* [1989] may need modifications. Also, the exponent γ=1 in the empirical MLP, eq. (3-1), is too small. Correlation between experimental data partition coefficients log P_{ow} and molecular lipophilic and hydrophilic indices L_M and H_M provides a criterion for the optimization of parameter γ in eq. (3-8). It also gives a criterion for the selection of the mathematical form of the screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$. However, there are some limitations to the use of partition coefficients as a criterion of HMLP.

3.4.2 Limitation of Partition Coefficients as a Criterion of HMLP

The quantity log P_{ow} is related to the molar standard free energy $\Delta G°_{tr}$ of transfer from the aqueous phase to the organic phase,

$$-\ln P_{ow} = \Delta G°_{tr} = \Delta H°_{tr} - T\Delta S°_{tr}. \qquad (3\text{-}17)$$

HMLP is the 3-dimensional representation of molecular lipophilicity. It gives detailed distributions of molecular lipophilicity in molecular space or on the surface. On the other hand, partition coefficients are one-dimensional scalar descriptors. The conversion from 3D HMLP to log P_{ow} is not so straightforward.

Many authors [Cabani and Gianni 1979, Cabani *et al.* 1981] point out that values of log P_{ow} are not very sensitive to the change of molecular structures. The reason is that sometimes the changes of two components of transfer free energy, enthalpy and entropy, caused by the substitutes, cancel each other. Partition coefficients are also affected by intramolecular interactions such as hydrogen bonding. Here I don't try to present a general method for the calculation of log P_{ow} from HMLP, or to compete with other methods. My purpose is to check the reasonableness of HMLP based on experimental logP_{ow} data using several families of simple compounds. The calculated partition coefficients using HMLP

are basically good in the above families of simple compounds, and this is a strong support for the validity of using HMLP. However, partition coefficients are not a good criterion for the 3-dimesional HMLP. I will discuss this question in the next section.

3.5 Conclusions and Discussions

A major goal in chemical research is to predict the behavior of new compounds based on their molecular structures. Quantitative correlations of molecular structures of ligands with the binding constants, and subsequently the predictions for novel compounds are the tasks of QSAR (quantitative structure/activity relationship). The heuristic lipophilicity potential, defined by eqs. (3-6)-(3-11), is based on structural information. The input data are molecular geometries and molecular surfaces. Quantum mechanics is used to calculate the electron density function $\rho(\mathbf{r})$ and the electrostatic potential $V(\mathbf{r})$. The examples in this study show that this model gives qualitatively and quantitatively correct, chemically reasonable results, and it works in cases not well described by the empirical lipophilicity potential [Croizet et al. 1990, Audry et al. 1986, Ghose and Crippen 1986]. Here I want to explore the possibility that this new technique could be used in computer aided rational 3D drug design.

3.5.1 Adding More Information in CoMFA

As illustrated in §3.1.1, in the studies of protein-ligand complexes, there are three main types of factors: steric, electrostatic, and lipophilic. In §3.1.2, I described the efforts that have been made by many authors to include the lipophilic potential energy field in CoMFA. Except for the lipophilic field, many authors want to put more fields in CoMFA, such as the H-bonding field [Kim 1993], the molecular orbital field [Waller and Marshall 1993, Waller *et al.* 1995], and the electrotopological field [Kellogg *et al.* 1997, Kier and Hall 1992], as well as information about molecular similarity [Klebe *et al.* 1994]. The creation of each of the above fields comes from a specific chemical and physical sense and is based on certain theoretical or experimental backgrounds. Several serious questions are raised with the addition of various new fields into CoMFA: Is there any contradiction among various fields? What are the most important fields? How much noise is brought

into CoMFA with various fields? I cannot answer these questions. Here, I just try to explore how a new field is developed.

As introduced in §3.1.2, the standard probe atom of CoMFA is a sp^3 carbon having the van der Waals properties and a charge of +1.0 *e*. Then van der Waals's 6-12 potential function and Coulombic law are used to calculate the interaction energies between the molecule under consideration and the probe atom at the lattice intersections. The energy field produced in this way contains information of electrostatic and steric interactions [Cramer *et al.* 1988]. In §3.1.2, I have already introduced the application of EMLP in CoMFA. In the program GRID developed by Kim [1993], a neutral water molecule is used as a probe to produce a H-bonding field. Two hydrogen bond donors and two hydrogen bond acceptors are assigned to the probe. Hydrogen bonding potential energy is calculated at each grid point according to the following equation,

$$E_{hb} = \left(\frac{C}{d^6} - \frac{D}{d^4}\right)\cos(m\theta), \tag{3-18}$$

where C, D and *m* are parameters for a specific hydrogen bond. In the GRID-CoMFA approach, a H_2O probe has also been used in conjunction with a steric probe (CH_3) and an electrostatic probe (H^-) to model steric and electrostatic fields. In the study of acceptor binding affinities of a series of benzodiazepines [Kim 1993] using the GRID-CoMFA model, it was found that the hydrophobic field explained 78% of the variance in the binding data, while the electrostatic field accounted for 18%. However, another example using the HINT field in the re-examination of the classic steroid data set shows that the empirical hydrophobic field does not improve the statistical measures of the model [Kellogg *et al.* 1991]. It is easy to imagine that one can also use other molecules or groups, such as NH_3 and OH, as probes and build other potential energy fields [Pastor *et al.* 1997, Waller and Kellogg 1996].

Actually, creating a new field for CoMFA is rather simple. All one needs to do is to make a set of atomic parameters with a certain physical or chemical meaning, and select

a distance-dependent function for the parameters. Then the field is created by summing the effects of all atoms on each grid point in the cage surrounding the molecule [Waller and Kellogg 1996]. The new field may or may not improve the performance of CoMFA, and different fields may give contradictory results and explanations.

It is obvious that HMLP can be used to make a new field for CoFMA. The heuristic lipophilicity potential is a modified electrostatic potential, taking the averaged influences from the environment. HMLP is very successful in the quantitative descriptions of molecular lipophilicity, hydrophilicity, and the hydrogen bond in the examples in §3.3. Therefore HMLP is a unified lipophilic-electrostatic potential, it contains the information of both MEP and MLP, and has a three-dimensional form. These characteristics appear to make HMLP very useful in the application of the CoFMA technique and in three-dimensional drug design. Unlike all other fields, the HMLP field is a non-empirical field, and there are no empirical atomic parameters used. HMLP is not a technique used to replace 3D QSAR and CoMFA, but rather, I suggest that HMLP may provide new complementary features to CoMFA and 3D QSAR. In Chapter 6, I will discuss this topic further.

3.5.2 Molecular Surface Used in HMLP

Atomic surface-MEP descriptors b_i, defined by eq. (3-7), are theoretical and structure-based atomic parameters, and play an important role in HMLP. In the calculation of b_i, the summation is over all area elements on the surface of atom i. Therefore, the model of HMLP depends on a molecular surface and the method of dividing a molecular surface into atomic pieces. In the case of a van der Waals's fused-sphere surface, a molecular surface and atomic pieces are decided by the atomic radii. As shown in Table 3-5, atomic radii affect the calculation results of the atomic lipophilic indices to a certain degree. There are several factors that affect the final results due to atomic radii.

1). If the radius of an atom is larger than it should be, $V(\mathbf{r})$ may be smaller than it should be and the surface area larger than it should be. Therefore, b_i shows little change

with the scale of atomic radii because the changes of area and $V(\mathbf{r})$ cancel each other.

2). An incorrect atomic radius may cut off part of the atomic surface or take part of the surface from a neighboring atom(s). Therefore, atomic radii affect the atomic lipophilic indices to a certain degree, see eq. (3-6) and (3-9). Particularly in the case that two bordering atoms have strongly opposite MEP.

3). The true border between two atoms may be not formed by fused spheres, but an irregular border. A fused-sphere surface may cause an incorrect division of atomic surfaces. An example can be found in Table 3-1, where the absolute values of b^+ and b^- of C_1 (in COOH) and C_3 (in the second CH_2 next to COOH) are very close.

The van der Waals's fused-sphere surface may not be the best molecular surface for the HMLP. It is hard to say where the atomic border in a molecular surface. In Chapter 6, I will discuss this question again. I think the theoretical electron isodensity surface may be better than the van der Waals's fused-sphere surface, and fuzzy sets and fuzzy logic can be used in the division of atomic surfaces.

3.5.3 Improvement of Screening Function

Lipophilicity potential is a real physical potential of solute molecules in an aqueous solution. It has a complex physical and chemical nature [Israelachvili 1992], and much further research is needed for a complete theoretical treatment of lipophilicity potential from first principles. The heuristic lipophilicity potential developed in this work is just one step toward this goal. The interaction between water molecules and an organic solute molecule at the position around an atom in the molecular surface is affected by all of its neighboring atoms. Atom-based screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ describes the influence from atom i. In the model of heuristic lipophilicity potential, the key is to find a good screening function. Besides the atomic surface MEP-descriptor b_i, one might need to include more properties, such as atomic shape parameters, which describe the interfering effect of atoms on the hydrogen-bonding network of water molecules in aqueous solution. Considering the size of water molecules, the screening function may be an oscillatory distance-decaying function.

Unlike empirical MLP, the screening function of HMLP has a certain physical meaning. However, at the first stage, I regard it as a mathematical function, and optimize it based on physical and chemical facts, such as experimental $\log P_{ow}$ data or solvation free energies through correlation calculations. The experimental technique that is best suited to provide the answers to the calculation results of HMLP is nuclear magnetic resonance (NMR) [Lee and Rossky 1994]. HMLP provides the information of 3D distributions of lipophilicity in a molecular space. NMR gives the details of water structure near certain atomic groups [Piculell 1986, Piculell and Halle 1986, Halle and Piculell 1986], therefore it may provide a good criterion for any further improvement of HMLP.

Just like the empirical MLP [Kellogg and Abraham 1992], heuristic MLP, possibly in combination with a computer graphic technique to show the distribution of MLP on the molecular surface in different colors, provides more detailed information, and gives the express visualization of lipophilicity. This technique is expected to be valuable for the study of complementary lipophilic and electrostatic maps on molecular surfaces in both direct and indirect drug design [Kellogg and Abraham 1992]. In Chapter 6, I will discuss this topic in more detail.

Chapter 4 Optimization of Screening Functions and Parameters

Summary

In this chapter, I test and compare three possible atom-based screening functions used in the heuristic molecular lipophilicity potential (HMLP). Screening function 1 is a power distance-dependent function, $b_i/\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}$, screening function 2 is an exponential distance-dependent function, $b_i exp(-\|\mathbf{R}_i-\mathbf{r}\|/d_0)$, and screening function 3 is a distance-dependent weighted function, $sign(b_i)exp[-\xi(\|\mathbf{R}_i-\mathbf{r}\|/|b_i|)]$. For every screening function, the parameters (γ, d_0, and ξ) are optimized using 41 common organic molecules of 4 types of compounds: aliphatic alcohols, aliphatic carboxylic acids, aliphatic amines, and aliphatic alkanes. The results of calculations show that screening function 3 cannot give chemically reasonable results, however, both power screening function 1 and exponential screening function 2 give chemically satisfactory results. There are two notable differences between screening functions 1 and 2. First, the exponential screening function has larger values in the short distance than the power screening function, therefore more influence from the nearest neighbors are involved using screening function 2 than screening function 1. Second, the power screening function has larger values in the long distance than the exponential screening function, therefore screening function 1 is affected more by atoms at long distance than screening function 2. For screening function 1, the suitable range of parameter γ is $1.0<\gamma<3.0$, $\gamma=2.3$ is recommended, and $\gamma=2.0$ is the nearest integral value. For screening function 2, the suitable range of parameter d_0 is $1.5<d_0<3.0$, and $d_0=2.0$ is recommended. HMLP developed in this research provides a potential tool for the combinatorial chemistry of small molecules and computer-aided three-dimensional drug design.

4.1 Introduction

In Chapter 3, I suggested a model of heuristic molecular lipophilicity potential (HMLP), and presented some examples and simple applications of this model. HMLP is a fully structural approach. There are no empirical atomic or fragment lipophilicity indices used, and the original input data are molecular geometries and molecular surfaces. HMLP is a three-dimensional, unified lipophilicity and hydrophilicity potential model. However, because of the complexity of this task, the formulas and equations used in this model are not all derived from first principles, therefore I say it is a heuristic model. The results obtained suggest that HMLP has the potential to become a useful tool in the areas of molecular modeling and computer-aided rational drug design.

The atom-based screening function, $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$, plays an important role in HMLP. In this study, I focus on the selections of mathematical forms of screening functions and the optimizations of parameters used in screening functions. I will compare three possible screening functions, and optimize the parameters used in these screening functions based on the calculation results of 41 common organic compounds of several families.

4.1.1 The Role of Screening Function in HMLP

For convenience, I rewrite the defining equation of heuristic molecular lipophilicity potential below, which first appeared in Chapter 3, §3.2.3 (eq. (3-6)),

$$L(\mathbf{r}) = V(\mathbf{r}) \sum_{\substack{\mathbf{r}\in S_\alpha \\ i\neq\alpha}} M_i(\mathbf{r};\mathbf{R}_i, b_i), \qquad (3\text{-}6)$$

where $V(\mathbf{r})$ is the molecular electrostatic potential (MEP) at point $\mathbf{r}$, and $\mathbf{r}$ is on the surface S_α of atom α. In summation, $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$ is an atom-based screening function at position $\mathbf{r}$ from atom i. In eq. (3-6) summation is over all constituent atoms, except atom α, on which point $\mathbf{r}$ sits. In the atom-based screening function, $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$, $\mathbf{R}_i$ is the nuclear position of atom i, and b_i is the atomic surface-MEP descriptor of atom i [Du and Arteca 1996 a], which is the integration of MEP on the atomic surface. Atomic surface-

MEP descriptor b_i's are the parameters describing the distributions of MEP on the molecular surface of atoms, however, they are structural and theoretical parameters, instead of empirical parameters. HMLP uses the same conventions as the empirical MLP [Ghose and Crippen 1986, Furet *et al.* 1988, Heiden *et al.* 1993]: the positive values of $L(\mathbf{r})$ represent lipophilicity, and the negative values of $L(\mathbf{r})$ are for hydrophilicity. HMLP is a unified lipophilicity and hydrophilicity potential. Here hydrophilicity includes the interactions of dipole moments, hydrogen bonds, and charged atoms of solute molecules with water molecules. Therefore HMLP is also a unified lipophilic and electrostatic potential.

The basic idea of the heuristic molecular lipophilicity potential defined by eq. (3-6) is that the interactions between organic molecules and water molecules at point **r** on the molecular surface are not only decided by the atom to which point **r** belongs, but also by a large microenvironment. The atom-based screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ conveys this idea, which represents the influence of atom i on point **r**. If the influence on point **r** from all surrounding atoms, $\Sigma_{i\neq\alpha}M_i(\mathbf{r}; \mathbf{R}_i, b_i)$, has the same sign as electrostatic potential $V(\mathbf{r})$, the lipophilicity potential $L(\mathbf{r})$ is positive, and at point **r,** the molecule is lipophilic. Otherwise, if the influence on point **r** from all surrounding atoms, $\Sigma_{i\neq\alpha}M_i(\mathbf{r}; \mathbf{R}_i, b_i)$, has the opposite sign to electrostatic potential $V(\mathbf{r})$, the lipophilicity potential $L(\mathbf{r})$ is negative, and at point **r** the molecule is hydrophilic. The model of HMLP does not include the structure of the hydration shell explicitly, but considers the effects of water molecules implicitly through the screening function.

4.1.2 General Considerations of Screening Functions

It is common knowledge that a lipophilic surface region is a nonpolar area, while a hydrophilic region is the polar area on the molecular surface. This means that the designation of lipophilic or hydrophilic surface area is decided by the molecular environment. In the study of molecular lipophilicity, the effects from surrounding atoms have drawn the attention of many authors. Israelachvili [1992 p. 135] believes that the hydrophilic and hydrophobic interactions are interdependent, unlike various electrostatic and dispersion interactions, which are independent interactions. He presents an example

where the hydrophobic energy per CH_2 group of an alkane chain is greatly changed when a hydrophilic head-group, such as OH, is attached to the end of the chain [Israelachvili 1992 p. 353]. Ghose and Crippen [1986] have developed a data set of empirical atomic lipophilicity indices for carbon, oxygen, nitrogen, sulfur, and halogens. They have classified these elements into 110 "atomic types". The factors taken into consideration for the classification are: 1) the electron distribution around the atom, 2) the approachability of the solvent, 3) the nature of the nearest atoms attached to the atom concerned, and 4) the influences from the next nearest neighbors. Carbon atoms may have as many as 4 directly connected neighbors, therefore the situations are extremely complex. For carbon alone, there are as many as 51 "atomic types", each assigned different values. Hansch and Leo [1979], Rekker [1977], Eisenberg and McLachlan [1986] have conducted similar studies.

Náray-Szabó has studied this problem using a different method [Náray-Szabó 1986]. He emphasizes that besides a geometric fit, electrostatic attraction and matching of nonpolar regions are also necessary to ensure optimal binding of a ligand to a biological acceptor. The electrostatic attraction accounts for ion-pair interactions and hydrogen bonding, while the matching of the nonpolar region represents the hydrophobic interactions. He uses MEP to describe the electrostatic interactions, and uses MEF (molecular electrostatic field) [Dughan *et al.* 1991, Mishra and Kumar 1995, Náray-Szabó 1989 a, b] to describe the hydrophobic interactions. MEF is the gradient of MEP,

$$E(\mathbf{r}) = -\nabla V(\mathbf{r}). \qquad (4\text{-}1)$$

MEF's are vectors. Actually, MEF represents the changes in the magnitude and direction of MEP in a molecular space. The results of calculations of MEF indicate that in the surroundings of the polar part of a molecular surface, MEF's show big changes in both direction and magnitude, on the other hand, in the surroundings of the nonpolar lipophilic part, there are no direction changes, and only small changes in magnitude. Platt and Silverman [1996] developed a new technique for the expansion of multipolar decomposition of electrostatic potential (MDEP). Multipolar expansion of MEP provides

an immediate characterization of the MEP distributions on the molecular space. Both MEF and MDEP are modifications of MEP, and the motivations for introducing MEF and MDEP are to attempt to describe the imbalance of MEP distributions on the molecular space, which is essential for the description of molecular lipophilicity.

In my heuristic molecular lipophilicity potential, all these considerations are included in the screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$. An atom-based screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ represents the influence at point $\mathbf{r}$ from atom i, and plays a key role in the HMLP. It is a function of position $\mathbf{r}$. Nuclear coordinates $\mathbf{R}_i$ and atomic surface-MEP descriptor b_i are parameters in the screening function. There may be more parameters needed for the description of some special situations, however, at the first stage, I hope to keep the mathematical form of the screening function as simple as possible. A complete theoretical derivation of the screening function from first principles, explicitly including a huge number of water molecules, is not easy. On the other hand, in some cases a heuristic lipophilicity potential is enough for the purpose of molecular modeling.

There is no experimental evidence to show what mathematical form the screening function should be. An experiment for the study of hydrophilic force law conducted by Israelachvili and Pashley [Israelachvili 1992 p.132] shows that the hydrophobic force decays exponentially in the range 0-10 nm. Marcelja *et al.* [1977] propose an equation for the calculation of hydrophobic force, which is an exponentially distance-dependent function. The hydrophobic force law is focused on the interaction between two organic molecules in an aqueous solution. However, a screening function is not designed for the hydrophobic force law between two solute molecules, but for the interaction ability of water molecules with the solute molecule in a certain position. Experimental results can provide some hints, however, there is no experimental technique advanced enough to give direct help for the selection of screening functions so far.

4.1.3 Equations of Empirical MLP

The empirical molecular lipophilicity potential (EMLP) takes a number of forms [Audry *et al.* 1989 a, b, 1986, Croizet *et al.* 1990, Heiden *et al.* 1993]. Audry *et al.* [1989 a, b] present an equation for the 3-dimensional representation of EMLP,

$$MLP = \sum_i \frac{f_i}{1+d_i} \tag{4-2}$$

where d_i is the distance (in Å) between a given point outside the molecular surface and the atom i, and f_i's are empirical atomic lipophilicity indices. Fauchère *et al.* [1988] propose another form for the EMLP. They have defined an exponential distance dependence for a fragmental contribution: EMLP ~ $exp(-d)$. Heiden *et al.* [1993] extend the selection of EMLP functions widely. They point out that there is no physical reason for the use of one or another distance dependent function in the empirical lipophilicity potential. They suggest a general form for EMLP [Heiden *et al.* 1993],

$$MLP = \frac{\sum_i g(d_i) f_i}{N} = \frac{\sum_i g(d_i) f_i}{\sum_i g(d_i)}, \tag{4-3}$$

where $g(d_i)$ is a distance dependent function, and $N=\Sigma_i\, g(d_i)$ is the normalization factor. Heiden *et al.* [1993] have discussed the conditions fulfilled by function $g(d_i)$. They use the Fermi function as $g(d_i)$,

$$g(d_i) = \frac{1}{\exp[a(d_i - d_{cut-off})] + 1}, \tag{4-4}$$

where $d_{cut\text{-}off}$ is an assigned cut-off value called the proximity distance.

4.2 Screening Functions in HMLP

All of the above research results of EMLP are valuable for the selection of a screening function for HMLP. Unlike empirical MLP, in which it is difficult to tell the physical meaning of function $g(d_i)$ because of the ambiguity of empirical parameters f_i's, the screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ of heuristic MLP may have a certain physical meaning.

4.2.1 Assumptions for Screening Functions in HMLP

Molecular electrostatic potential $V(\mathbf{r})$ is the measurement of the interaction ability of a solute molecule with a unit test charge at point $\mathbf{r}$. Suppose a water molecule is a dipole consisting of two opposite point charges (q^+_w and q^-_w), then $V(\mathbf{r}_A)$ and $V(\mathbf{r}_B)$ are the measurements of interaction abilities of atom A and atom B in a solute molecule with water molecules at point $\mathbf{r}_A$ and $\mathbf{r}_B$, respectively. If the MEP at atom A is positive, $V^+(\mathbf{r}_A)$, then the interaction between atom A and negative charge q^-_w of a water molecule is attractive,

$$E_{A,W} = V^+(\mathbf{r}_A)\, q^-_w < 0. \tag{4-5}$$

If the MEP at the neighboring atom B of atom A is negative, $V^-(\mathbf{r}_B)$, there is an attractive interaction between atom B and positive charge q^+_w of a water molecule,

$$E_{B,W} = V^-(\mathbf{r}_B)\, q^+_w < 0. \tag{4-6}$$

The interaction between two water molecules binding on atom A and B is attractive, too, $q^+_w\, q^-_w / \mathbf{r}_{AB} < 0$. Therefore, both atom A and B are hydrophilic. If the MEP at atom B is positive, $V^+(\mathbf{r}_B)$, there is an attractive interaction between atom B and the negative charge q^-_w of a water molecule,

$$E_{B,W} = V^+(\mathbf{r}_B)\, q^-_w. \tag{4-7}$$

However, the interaction between two water molecules binding on atoms A and B is repulsive, $q^-_w\ q^-_w\ /\mathbf{r}_{AB} > 0$, and is much stronger than the attractive interactions with atoms A and B. In this case, both atom A and B are lipophilic. I assume that the influence of atom B on atom A is affected by atomic surface-MEP descriptor b_i of atom B and the distance between atom A and B. If one uses the ST2 water model [Stillinger and Rahman 1974], which contains two hydrogen bonding donors and two hydrogen bonding acceptors located along four tetrahedral arms radiating out from the center of the O atom, the interaction between two water molecules is assumed to involve 16 Coulombic terms representing the interactions between four point charges on one molecule with four on the other. There are an additional 16 Coulombic terms between two atoms (q_A and q_B) and 8 point charges of two water molecules.

Because of the complexity of this task, I do not initially pursue the physical meaning of the screening function, but instead think of the screening function as a mathematical function, and select its mathematical form and optimize its parameters based on the calculation results. From the chemical and physical facts, a screening function should satisfy the following four conditions:

1) If at point **r** a molecule is lipophilic, $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ has the same sign as $V(\mathbf{r})$; otherwise, $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ has the opposite sign as $V(\mathbf{r})$,
2) If the absolute value of atomic surface-MEP descriptor b_i is higher, the absolute value of $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ is higher too; otherwise, $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ is smaller,
3) $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ decays with the distance $\|\mathbf{R}_i-\mathbf{r}\|$;
4) $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ is a dimensionless function.

Condition 1 ensures that EMLP is chemically reasonable. Condition 2 conveys the idea that the atomic surface-MEP descriptor b_i plays an important role in the screening function. Condition 3 is based on physical fact. Condition 4 makes the HMLP have the same unit as MEP. The above four conditions are assumptions for the screening function, yet there may be other conditions. For example, $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ may decay with the distance in an oscillatory fashion and may contain a geometric or topological parameter which

represents the tendency of an atom to interfere with the hydrogen bonding network of water molecules. The reasonableness of conditions should be examined by calculation results based on the chemical and physical facts.

4.2.2 Three Possible Screening Functions for HMLP

The atom-based screening function $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ can take a number of forms. Here I suggest three possible screening functions:

$$M_i(\mathbf{r};\mathbf{R}_i,b_i) = \frac{r_0^{\gamma}}{b_0}\frac{b_i}{\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}} = \varsigma\frac{b_i}{\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}}, \tag{4-8}$$

$$M_i(\mathbf{r};\mathbf{R}_i,b_i) = \frac{b_i}{b_0}e^{-\frac{\|\mathbf{R}_i-\mathbf{r}\|}{d_0}}, \tag{4-9}$$

$$M_i(\mathbf{r};\mathbf{R}_i,b_i) = \frac{b_i}{|b_i|}e^{-\frac{b_0}{\lambda_0}\frac{\|\mathbf{R}_i-\mathbf{r}\|}{|b_i|}} = sign(b_i)\,e^{-\xi\frac{\|\mathbf{R}_i-\mathbf{r}\|}{|b_i|}}. \tag{4-10}$$

Screening function (4-8) has been used in Chapter 3. In the above three functions (r_0, b_0, γ), (b_0, d_0) and (b_0, λ_0) are parameters. The unit of b_0 is the same as b_i (energy•area) and r_0 , d_0, and λ_0 have units of length. Therefore, $M_i(\mathbf{r}; \mathbf{R}_i, b_i)$ is a dimensionless function in all of the above three equations. In screening function 1, eq. (4-8), $\zeta=(r_0)^{\gamma}/b_0$ is a simple scaling factor. In my calculations, I take $\zeta=1$. Exponent γ is the parameter that decides how strong the influence is and how rapidly the influence decays with distance. It will be optimized in this study. In screening function 2, eq. (4-9), b_0 is a simple scaling factor, and is assigned value $b_0=1$ in this study. Parameter d_0 in eq. (4-9) plays the same role as γ in eq. (4-8), and will be optimized later. In the screening function 3, eq. (4-10), $\xi=b_0/\lambda_0$ makes the exponent a dimensionless quantity, and affects the behavior of the screening function, like γ and d_0 in eq. (4-8) and (4-9). It will be optimized, too. The positions of atomic MEP-surface descriptor b_i in the three functions are different. In eqs. (4-8) and (4-9), b_i is a factor of distance dependent functions, $1/|\mathbf{R}_i\text{-}\mathbf{r}|^{\gamma}$ and $exp(-|\mathbf{R}_i\text{-}\mathbf{r}|/d_0)$. In eq. (4-10), the sign of b_i is a factor, however, the value of b_i is put in the exponent, like a weighting function. In this research, I will test and compare all three screening functions

and optimize parameters used in them based on the chemical and physical facts and experimental partition coefficient data, $\log P_{ow}$.

4.3 Optimizations of Screening Functions and Parameters

In this study, the *ab initio* quantum chemical program package Gaussian 92 is used to calculate electron density $\rho(\mathbf{r})$ and MEP's on the grid of molecular surfaces at the RHF/6-31G* level. Molecular geometries are optimized using Gaussian 92 at the level RHF/STO-3G. Fused-sphere van der Waals surfaces are used, and atomic radii are optimized based on MEP criteria [Du and Arteca 1996]. Molecular surfaces are generated by program MS [Connolly 1985, 1983 a, b] using point density of 25 points/Å^2.

4.3.1 Optimizations Using Four Simple Compounds

Fig. 4-1 shows the optimizations of parameter γ in screening function 1, eq. (4-8), using (a) ethanol, (b) propionic acid, (c) ethylamine, and (d) propane. In Fig. 4-1, atomic lipophilicity indices, l_a, molecular lipophilic indices, L_M, and molecular hydrophobic indices, H_M, are shown for each molecule according to the definitions of eqs. (3-9), (3-10), and (3-11) in Chapter 3, §3.3.1. If the atomic index $l_a>0$, then the atom a is lipophilic, whereas, if $l_a<0$, then atom a is hydrophilic. The molecular lipophilic index (L_M) is the sum of all lipophilic atoms ($l_a>0$), and the hydrophilic index (H_M) is the sum of all hydrophilic atoms ($l_a<0$), respectively.

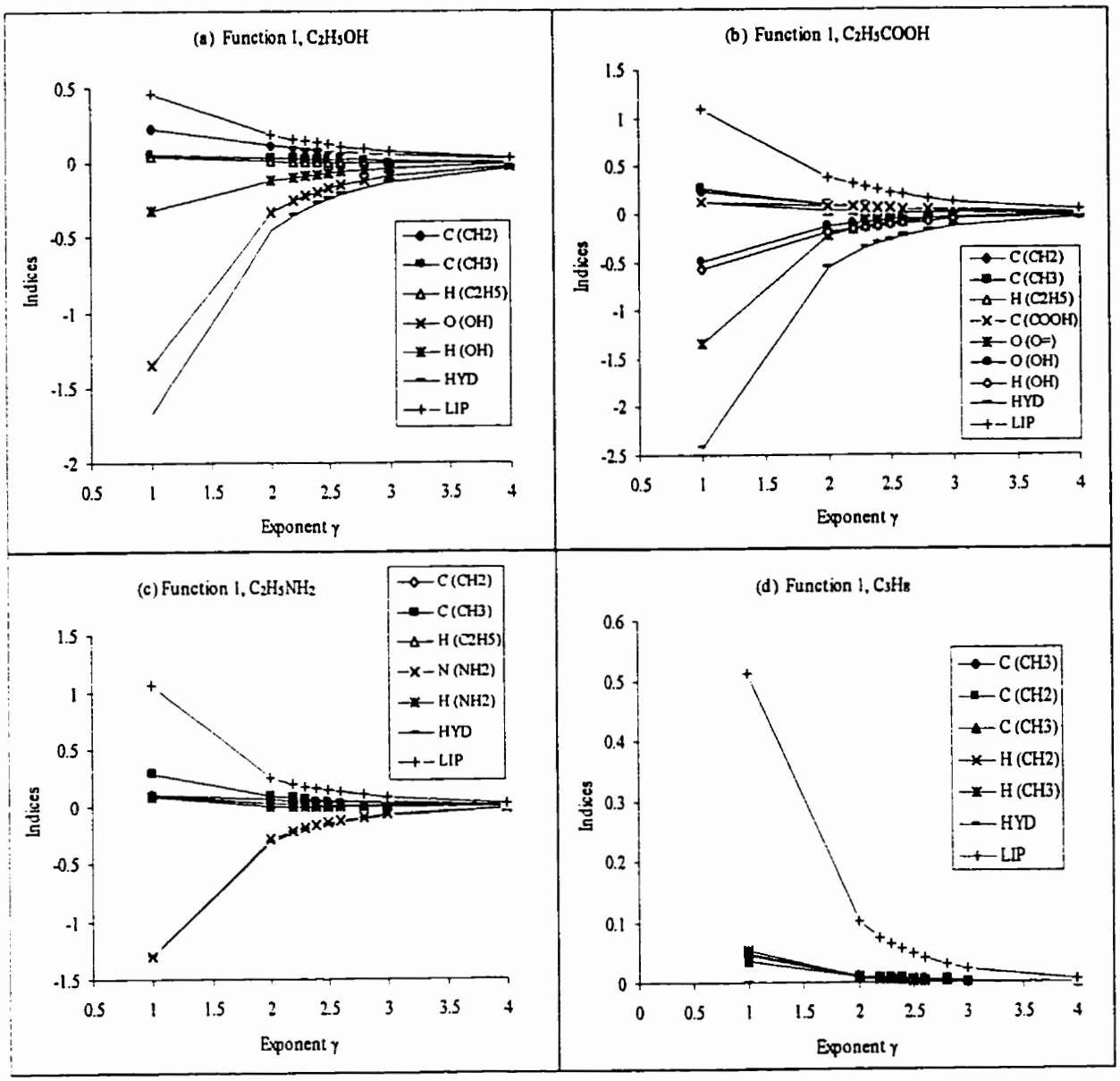


Figure 4-1. Optimization of γ parameter in screening function 1, eq. (4-8), using (a) ethanol, (b) propionic acid, (c) ethylamine, and (d) propane. Molecular lipophilic indices, L_M, molecular hydrophilic indices, H_M, and atomic lipophilicity indices, I_a, of functional groups, carbons, and several of the hydrogen atoms are shown.

In Fig. 4-1, various indices of the 4 molecules (ethanol, propionic acid, ethylamine, and propane) are shown as functions of γ in screening function 1. Various indices of ethanol are shown in Fig. 4-1 (a). Atomic lipophilicity indices I_O and I_H in hydroxyl group are negative. This means that O and H in the hydroxyl group are hydrophilic atoms. All carbons and hydrogens in the hydrocarbon chain are lipophilic, having positive lipophilicity indices. All values of γ give qualitatively reasonable results based on the chemical facts. The absolute values of all indices decrease with increasing γ, however, in the range γ=2.0 - 2.5, decreases in the indices are getting smaller. In Fig. 4-1 (b), propionic acid shows behavior very similar to that of ethanol. A detailed examination show that there is an order of magnitude change in the atomic lipophilicity indices ($I_{=O}$, I_O and I_H) in the carboxyl group with an increase of γ. When $\gamma \leq 2.0$, the order is $I_{=O}<I_H<I_O$; then, when $\gamma>2.0$, the order becomes $I_H<I_{=O}<I_O$; finally, when γ reaches the value γ=4.0, the order is $I_H<I_O<I_{=O}$. The details of these observations can be found in Fig. 4-4 (a). Ethylamine has something special, as shown in Fig. 4-1 (c). Two hydrogens in the amino group $-NH_2$ have positive lipophilicity indices, I_H, in the range $\gamma \leq 1.0$. This is unreasonable from a chemical point of view. However, when $\gamma \geq 2.0$, the indices I_H of the two hydrogens turn to negative, and at the value γ=2.3, the I_H's reach their minimum (I_H=-0.0113). The details of these observations can be found in Fig. 4-4 (b). For propane in Fig. 4-1 (d), there are no negative atomic lipophilicity indices, and the molecular hydrophilic index H_M is 0. This means that all atoms (carbons and hydrogens) are lipophilic. This result is reasonable based on the chemical facts.

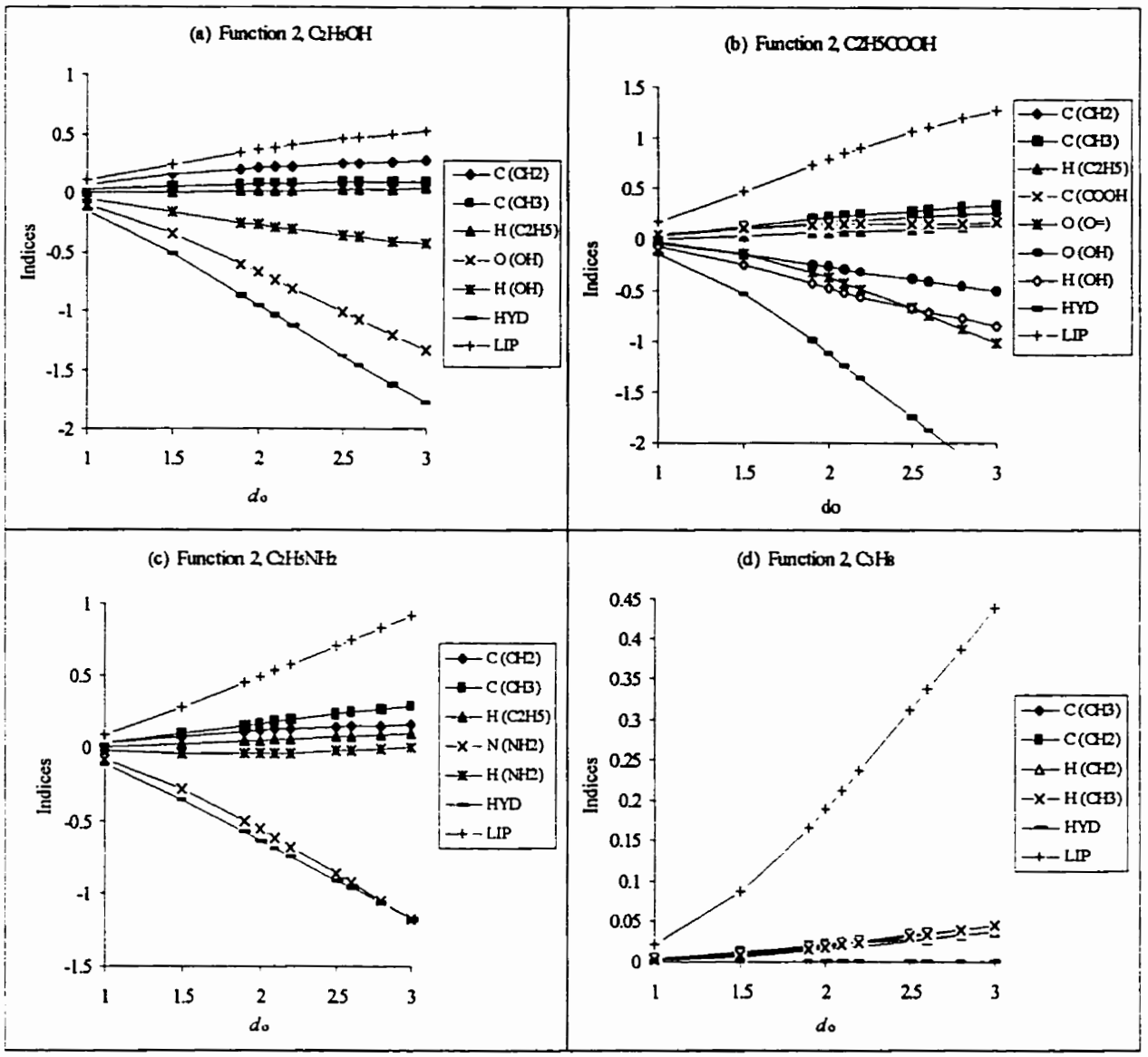


Figure 4-2. Optimizations of parameters d_o in screening function 2, eq. (4-9). (a) ethanol, (b) propionic acid, (c) ethylamine, and (d) propane. Molecular lipophilic indices, L_M, molecular hydrophilic indices, H_M, and atomic lipophilicity indices, l_a, of functional groups, carbons, and several hydrogen atoms are shown.

In Fig. 4-2 (a) the absolute values of various indices of ethanol increase nearly linearly with parameter d_0 of screening function 2. As in Fig. 4-1 (a), O and H of the hydroxyl group are hydrophilic, having negative atomic lipophilicity indices l_O and l_H. All carbons and hydrogens in the hydrocarbon chain are lipophilic, having positive lipophilicity indices. The behavior of propionic acid, see Fig. 4-2 (b), is very similar to that of ethanol. Once again, an order of magnitude change is found for the atomic lipophilicity indices of the carboxyl group. When $d_o \leq 1.0$, the order is $l_H < l_O < l_{=O}$; then, for $d_o > 1.5$, the order becomes $l_H < l_{=O} < l_O$; finally, when d_o reaches the value $d_o > 2.6$, the order is $l_{=O} < l_H < l_O$. Details can be found in Fig. 4-4(c). As in Fig. 4-1 (c) of ethylamine, the atomic lipophilicity indices l_H's of two hydrogens in the amino group NH_2 have a minimum (l_H=-0.0411) at d_o=1.9. The details can be found in Fig. 4-4 (d). For propane in Fig. 4-2 (d), there are no negative atomic lipophilicity indices, and the molecular hydrophilic index H_M is 0, as in Fig. 4-1 (d).

At first glance, the general tendencies of various indices in Fig. 4-3 are much similar to those in Fig. 4-1; however, more careful examination shows that there are some differences between the two figures. The absolute values of various indices in Fig. 4-3 are much smaller than those in Fig. 4-1. For hydrocarbon propane, Fig. 4-3 (d), all indices are almost zero. In Fig. 4-3 (a) of ethanol, atomic lipophilicity indices of l_O and l_H of the hydroxyl group are negative, and atomic lipophilicity indices of two carbons in the hydrocarbon chain have positive values, as in Fig. 4-1 (a). However, the hydrogen indices in the hydrocarbon chain have very small negative values. In Fig. 4-3 (b), in the hydrocarbon chain of propionic acid, the hydrogen and the carbon atoms, connected directly with the carboxyl group, have small negative atomic lipophilicity indices. In Fig. 4-3 (c) of ethylamine, a negative atomic lipophilicity index is found for the carbon atom in the methyl of the hydrocarbon chain. All these phenomena are unreasonable based on the chemical facts.

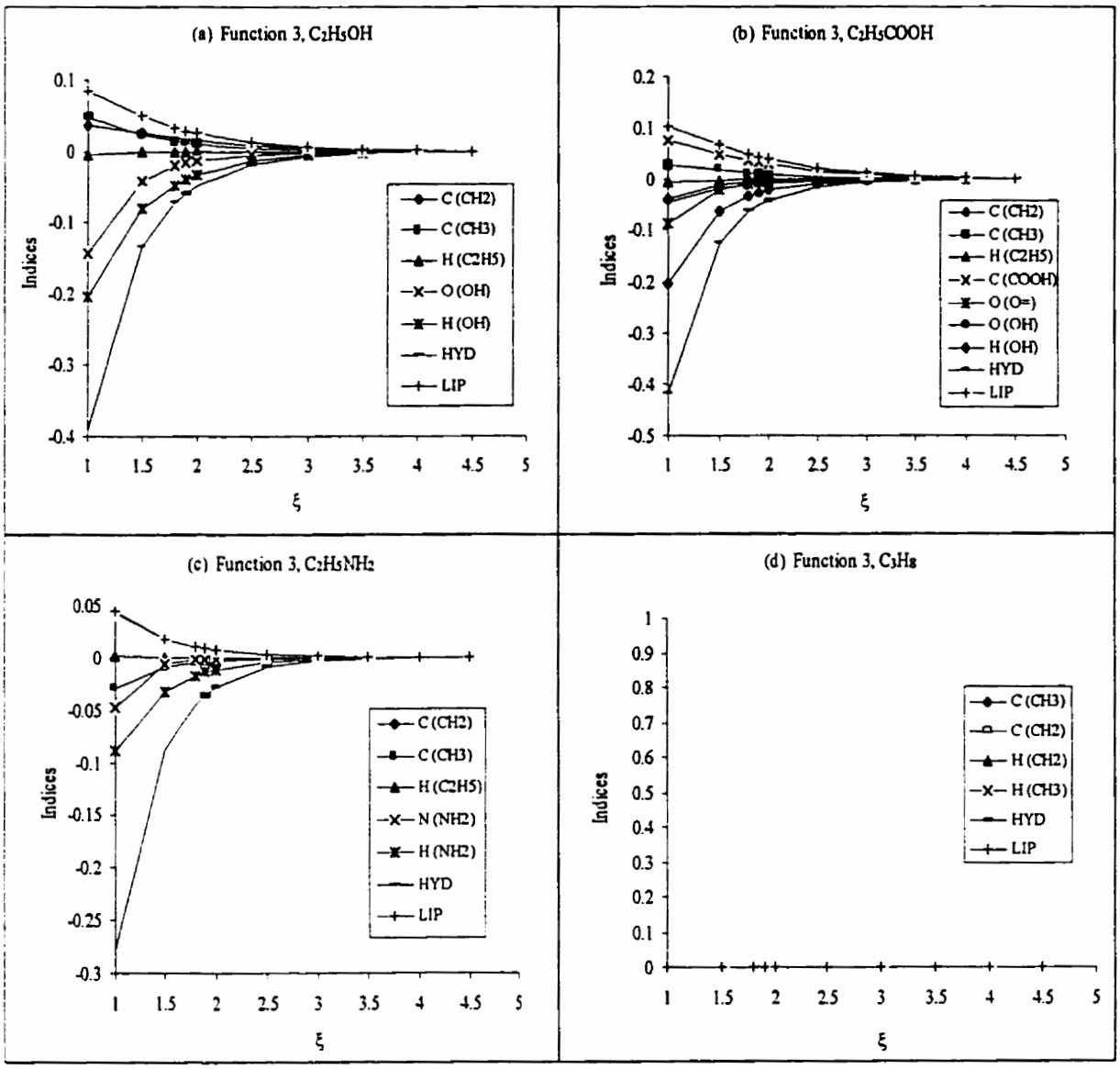


Figure 4-3. Optimizations of ξ parameter in screening function 3, eq. (4-10), (a) ethanol, (b) propionic acid, (c) ethylamine, and (d) propane. Molecular lipophilic indices, L_M, molecular hydrophilic indices, H_M, and atomic lipophilicity indices, l_a, of functional groups, carbons, and several hydrogen atoms are shown.

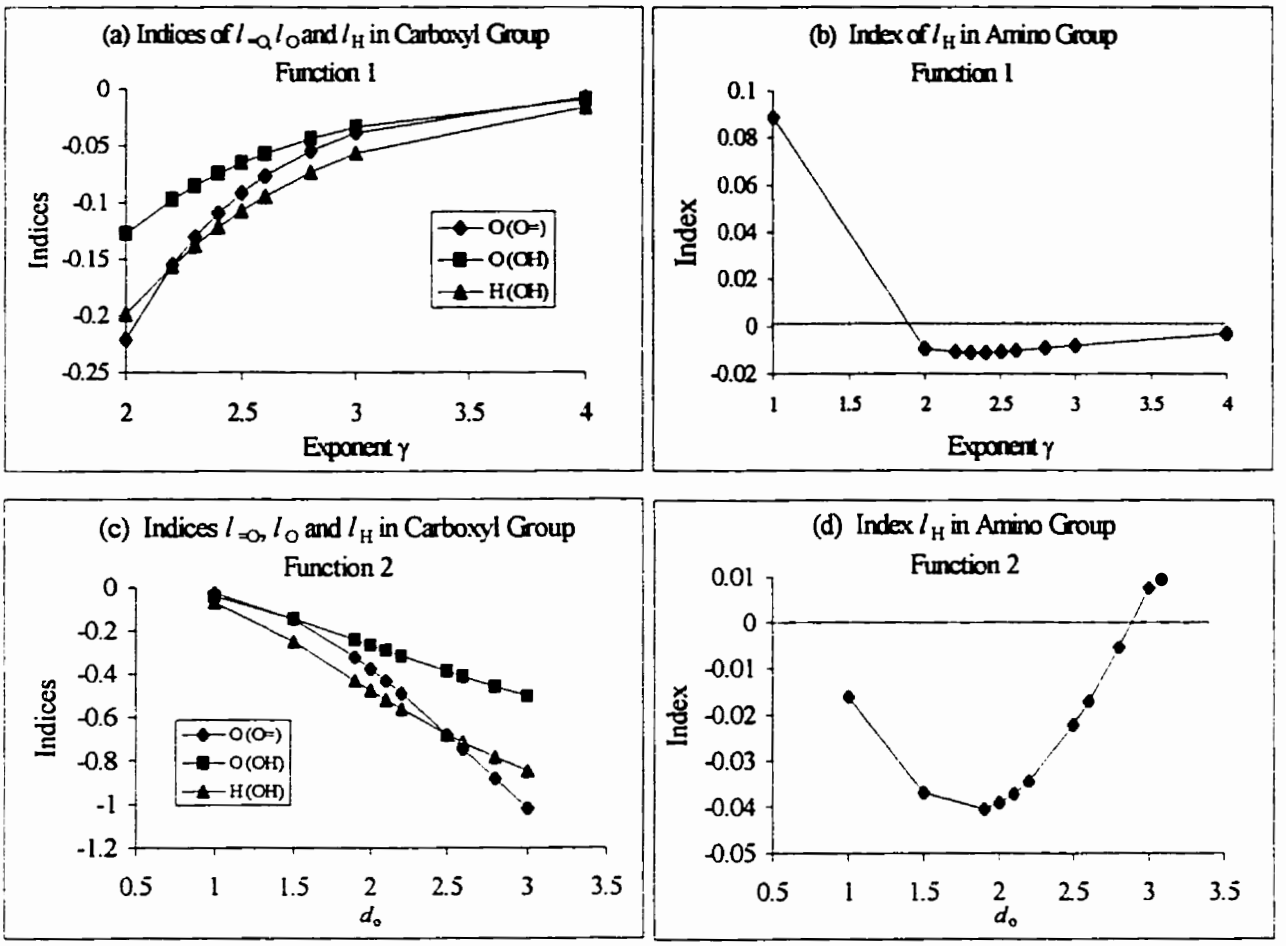


Figure 4-4. Changes in the order of the indices of $l_{=O}$, l_O, and l_H in carboxyl group, COOH, and the minimum of l_H in amino group, NH_2, using screening functions 1 and 2.

4.3.2 Optimizations Using 41 Compounds

In this part of §4.3, I show the optimizations of the three screening functions and parameters, using 4 series of compounds: aliphatic alcohols, aliphatic carboxylic acids, aliphatic amines, and linear hydrocarbons, which contain carbon atoms from 1 to 10. I will show the calculated results of molecular lipophilic indices, L_M, and hydrophilic indices, H_M, as functions of the number of carbon atoms and the parameters γ, d_o, and ξ in the three types of screening functions.

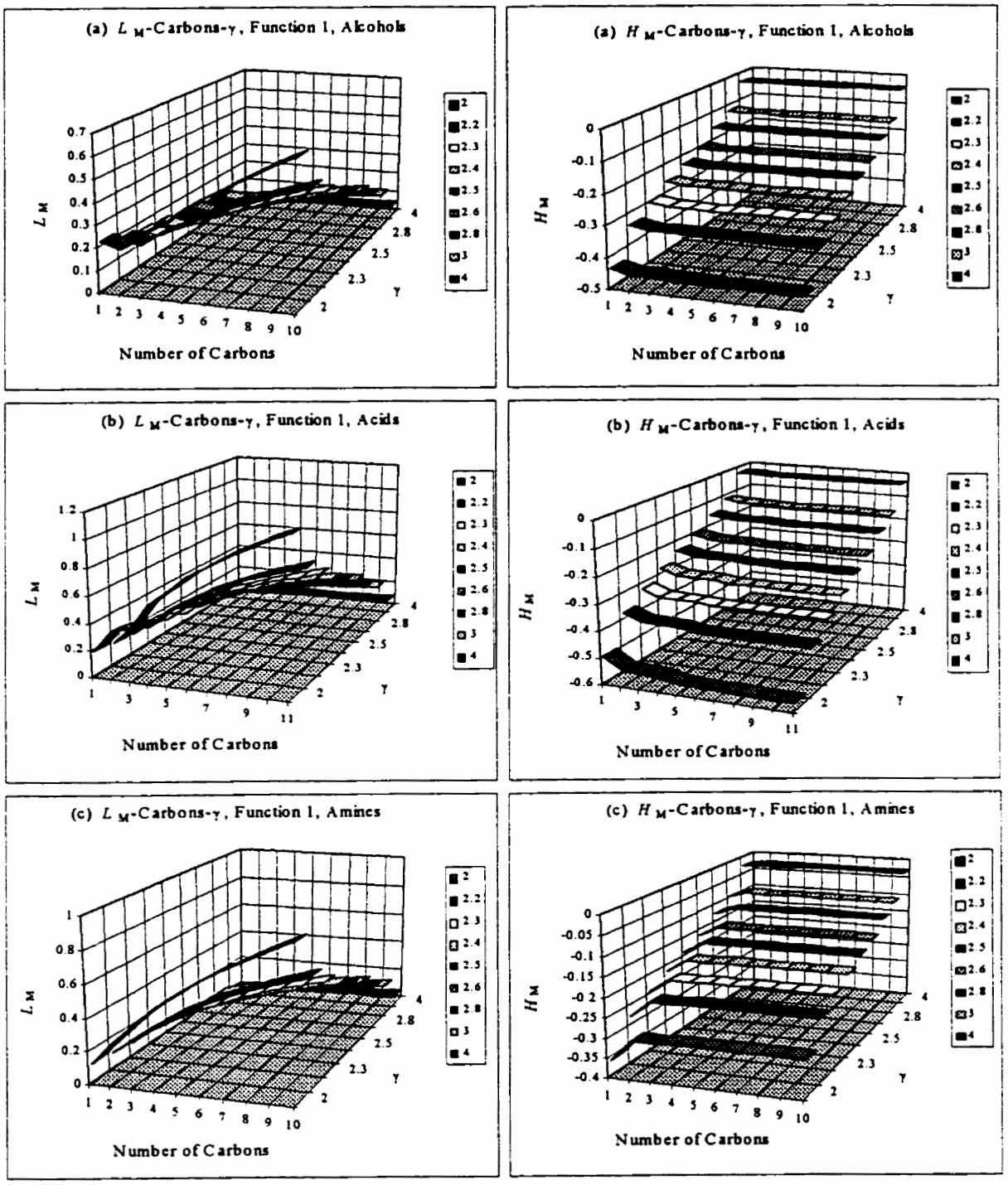


Figure 4-5. Optimization of parameters γ in screening function 1, eq. (4-8), using (a) aliphatic alcohols, (b) aliphatic carboxylic acids, and (c) aliphatic amines. Molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, are shown as functions of the number of carbon atoms and parameter γ.

Fig. 4-5 tells us that the molecular lipophilic indices, L_M, increase with the number of carbon atoms and decrease with increasing magnitude of exponent γ in screening function 1 for all three types of compounds, respectively. The increases and decreases are approximately linear, however, more careful examination shows that the increases in L_M are not exactly linear, but become smaller with increasing number of carbons. Molecular hydrophilic indices, H_M, remain basically constant with increasing number of carbon atoms, and increase with increasing magnitude of exponent γ.

Fig. 4-6 shows that the molecular lipophilic indices, L_M, increase with the number of carbon atoms for all three types of compounds, and the increases are approximately linear, as in Fig. 4-5. However, L_M's increase with the parameter d_0 of screening function 2, unlike γ of screening function 1, where L_M's were observed to decrease. Careful examination shows that the increases of L_M became smaller as the number of carbons increases, as in Fig. 4-5. Molecular hydrophilic indices, H_M's, remain basically constant as the number of carbon atoms increase and decrease with increasing magnitude of d_0.

Fig. 4-7 is completely different from Fig. 4-5 and 4-6. The molecular lipophilic indices, L_M, increase slightly with the number of carbon atoms for all three types of compounds. However, when ξ is smaller (*e.g.*, ξ=1.0), the increases are greater. For the first three carbons, there is a big fluctuation in the increases. The L_M's decrease with increasing ξ. The molecular hydrophilic indices, H_M, are basically constant with the increasing carbon atoms, however, there is a fluctuation in the first three carbons. The H_M's increase with an increase in the parameter ξ.

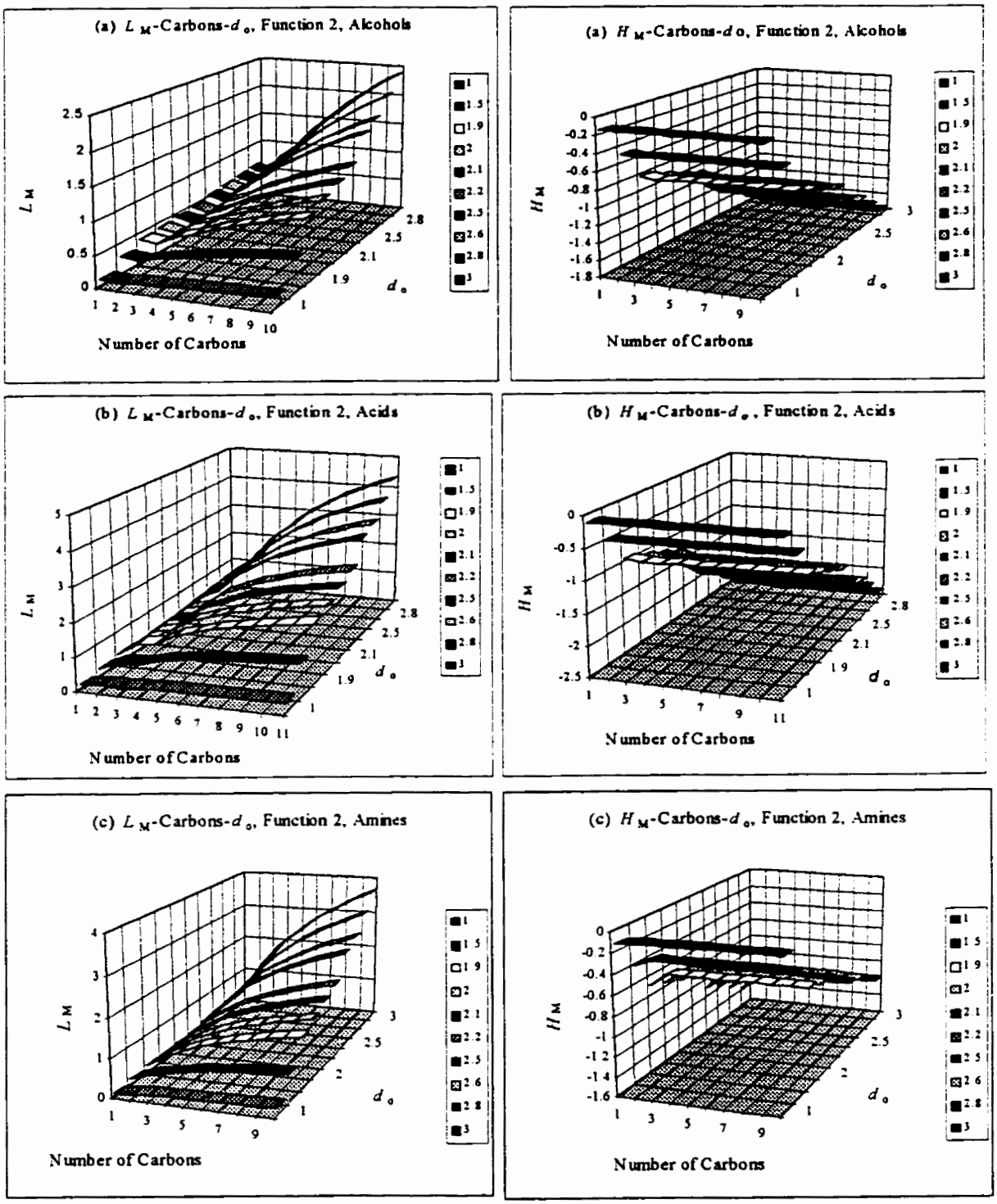


Figure 4-6. Optimization of parameters d_0 in screening function 2, eq. (4-9), using (a) aliphatic alcohols, (b) aliphatic carboxylic acids, and (c) aliphatic amines. Molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, are shown as functions of the number of carbon atoms and parameter d_0.

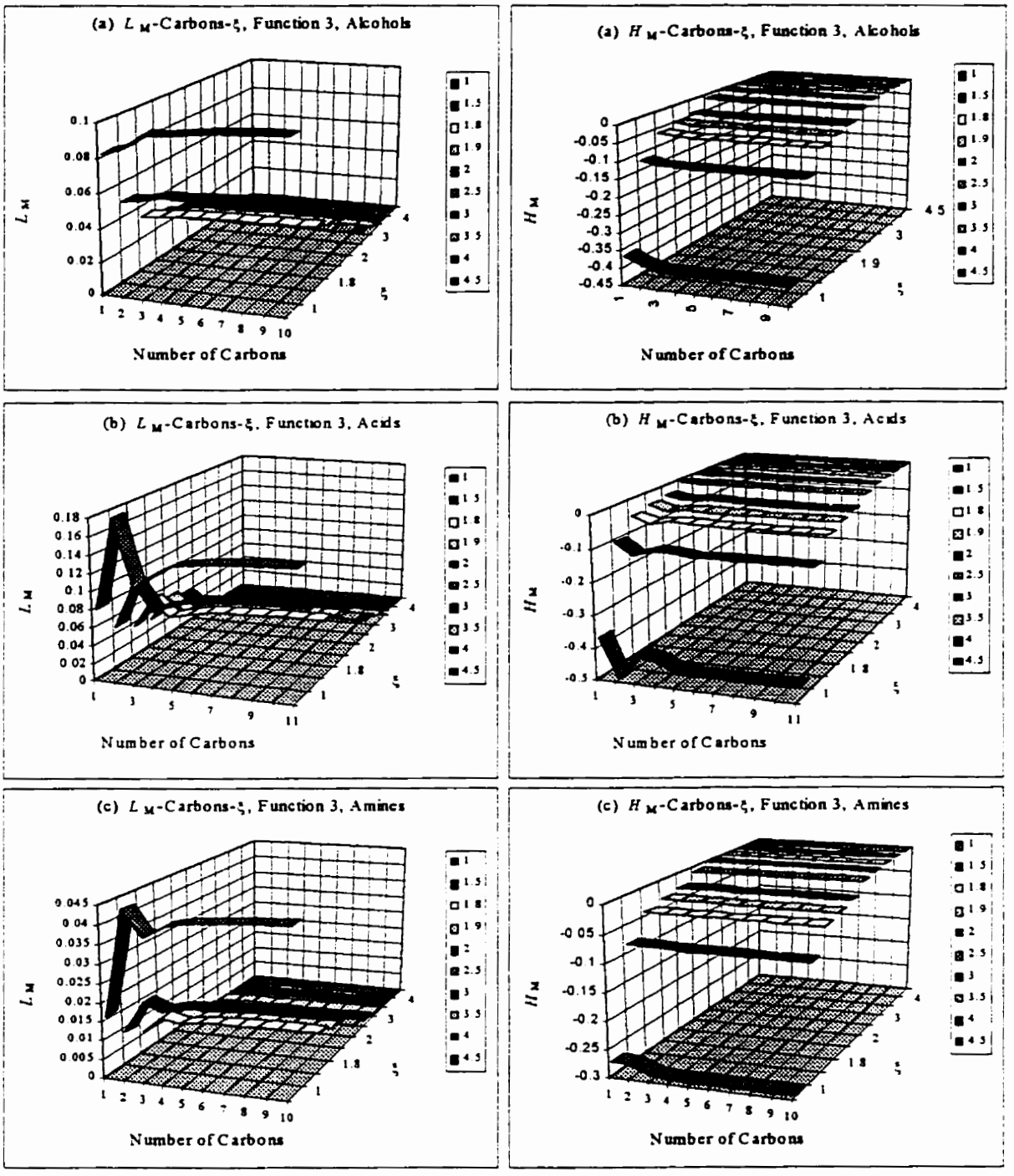


Figure 4-7. Optimization of parameters ξ in screening function 3, eq. (4-10), using (a) aliphatic alcohols, (b) aliphatic carboxylic acids, and (c) aliphatic amines. Molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, are shown as functions of the number of carbon atoms and parameter ξ.

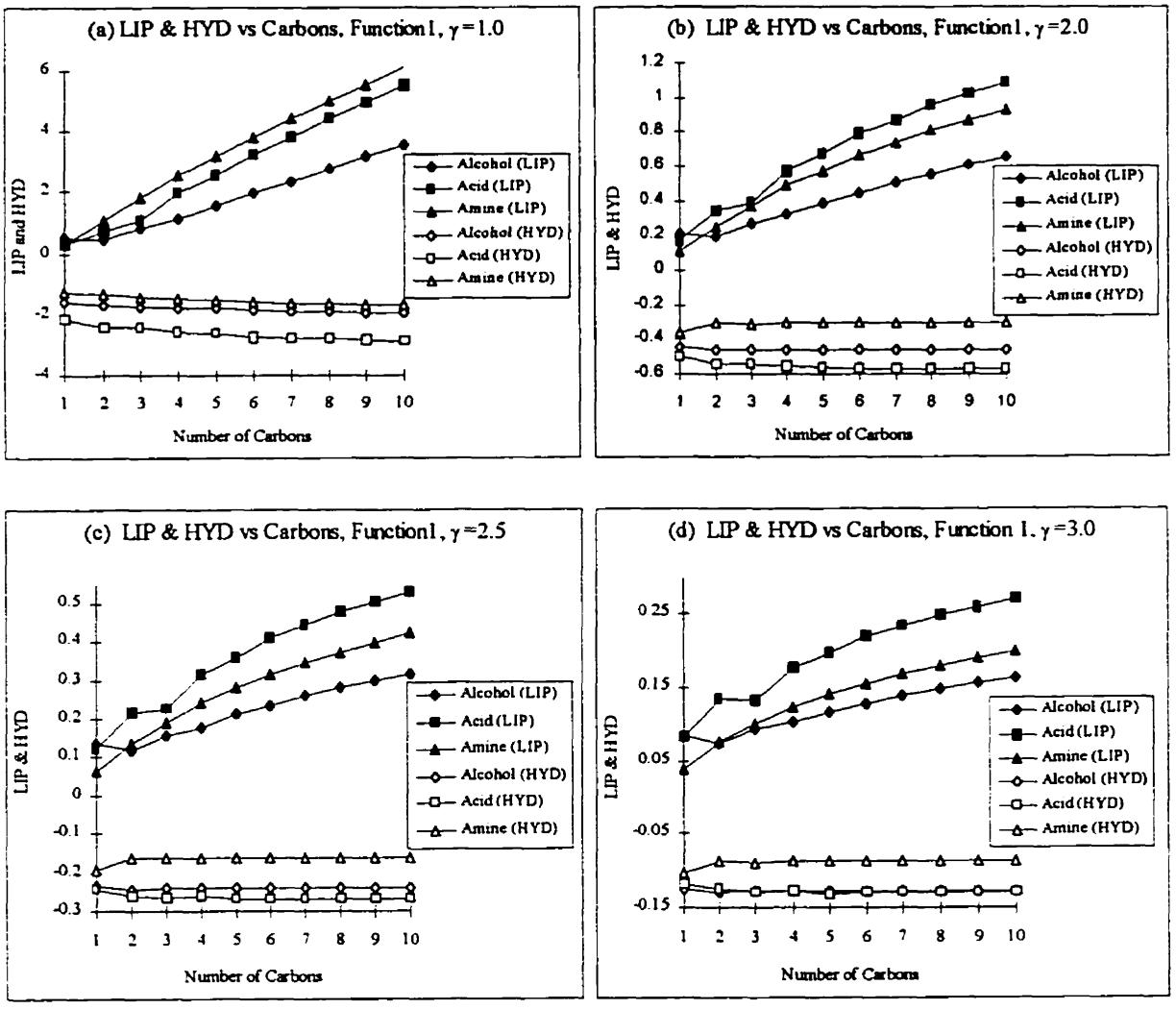


Figure 4-8. Molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, are shown as functions of the number of carbon atoms for aliphatic alcohols, aliphatic carboxylic acids, and aliphatic amines. Parameter γ in screening function 1 takes several values: (a) γ=1.0, (b) γ=2.0, (c) γ=2.5, and (d) γ=3.0.

Fig. 4-8 shows the molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, as functions of the number of carbon atoms for aliphatic alcohols, aliphatic carboxylic acids, and aliphatic amines, using screening function 1. In Fig. 4-8, parameter γ in eq. (4–8) takes several values: (a) γ=1.0, (b) γ=2.0, (c) γ=2.5, and (d) γ=3.0. From

Fig. 4-8, I find that when γ=1.0, the order of the L_M's of the three families of compounds is amines>acids>alcohols, and the order of the H_M's of the three compounds is acids<alcohols<amines. However, when γ becomes larger, *e.g.*, γ=2.0, 2.5, 3.0, the order of L_M's of the three compounds changes to acids>amines>alcohols. The order of H_M's remains the same, however, when γ=3.0, values of H_M for the alcohols and acids almost overlap each other (cf. Fig. 4-8 (d)). For all 4 values of γ, the molecular hydrophilic indices H_M's remain basically constant with increasing number of carbon atoms. The molecular lipophilic indices, L_M, increase with the number of carbon atoms. However, the increases are not exactly linear; the γ is smaller and the linearity is better.

Molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, are shown in Fig. 4-9 as functions of the number of carbon atoms for aliphatic alcohols, aliphatic carboxylic acids, and aliphatic amines using screening function 2. In Fig. 4-9, parameter d_o in eq. (4–9) takes several values: (a) d_0=1.0, (b) d_0=1.5, (c) d_0=2.0, and (d) d_0=3.0. From Fig. 4-9, I find that when d_0=1.0, the order of L_M for the three compounds is acids>amines>alcohols, and the order of H_M for the three compounds is alcohols<acids<amines. However, when d_0 becomes larger, *e.g.*, d_0=1.5, 2.0, and 3.0, one order is changed: the order of H_M for the three compounds becomes acids< alcohols<amines, while the order for L_M remains the same. For the 4 different values of d_0, the molecular hydrophilic indices, H_M, remain basically constant with an increase in the number of carbon atoms. Molecular lipophilic indices, L_M, increase with the number of carbon atoms. However, the increase is not exactly linear; the d_0 is larger and the linearity is better.

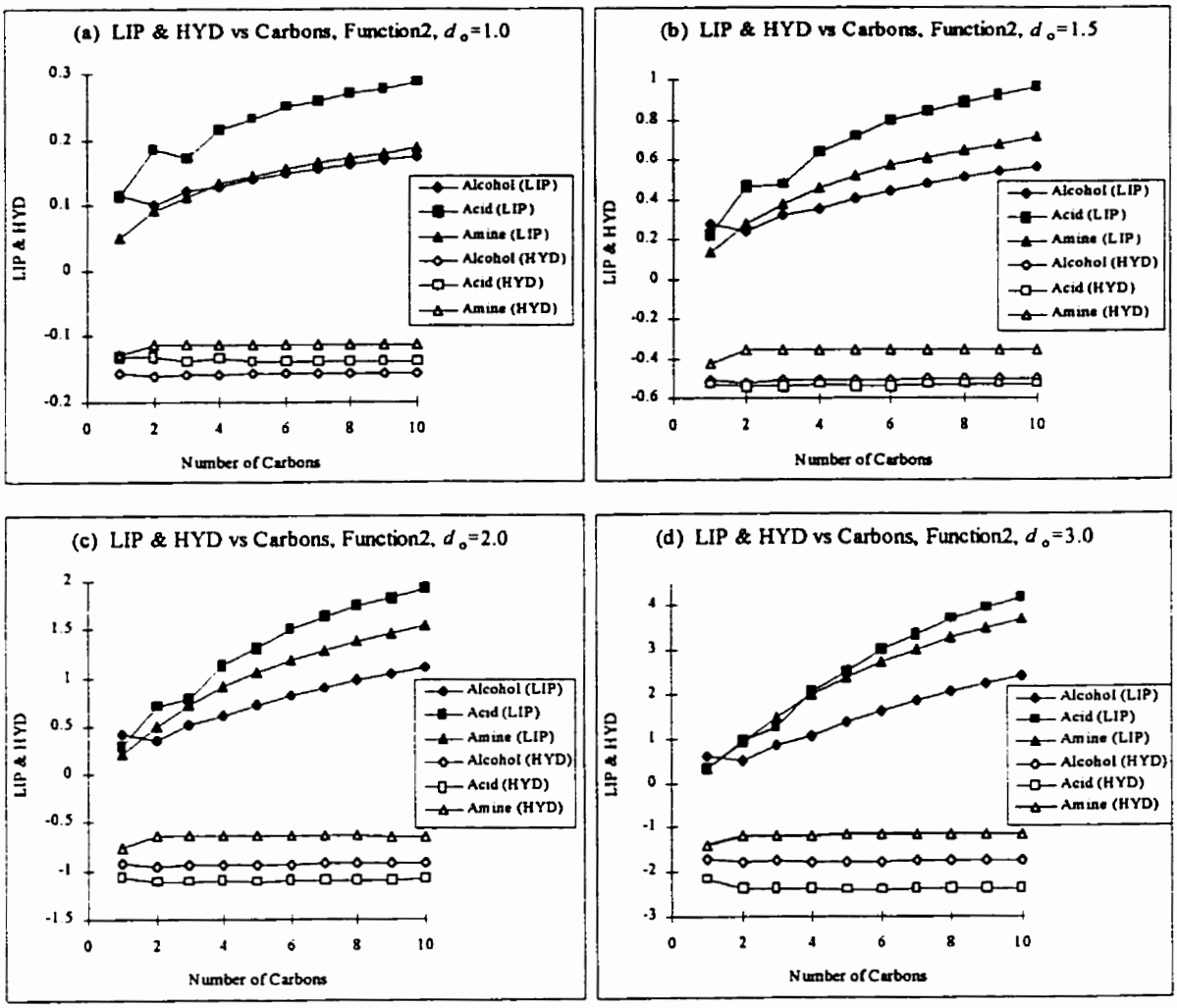


Figure 4-9. Molecular lipophilic indices, L_M, and molecular hydrophilic indices, H_M, are shown as functions of the number of carbons for aliphatic alcohols, aliphatic carboxylic acids, and aliphatic amines. Parameter d_0 in screening function 2 takes several values: (a) d_0=1.0, (b) d_0=1.5, (c) d_0=2.0, and (d) d_0=3.0.

4.3.3 Partition Coefficients as a Criterion of Optimization

Fig. 4-10 shows the experimental logP_{ow} data [Leo *et al.* 1971] as a function of the number of carbon atoms, which are nearly linear for the three types of compounds. However, unlike Fig. 4-8 and 4-9, I cannot find any difference in the slopes for the three series of compounds from Fig. 4-10. The three lines show similar dependence, and the alcohol and acid lines almost overlap. Generally speaking, the logP_{ow} values for the

amines are smaller than the acids and alcohols, and the logP_{ow} of the acids and alcohols are almost the same.

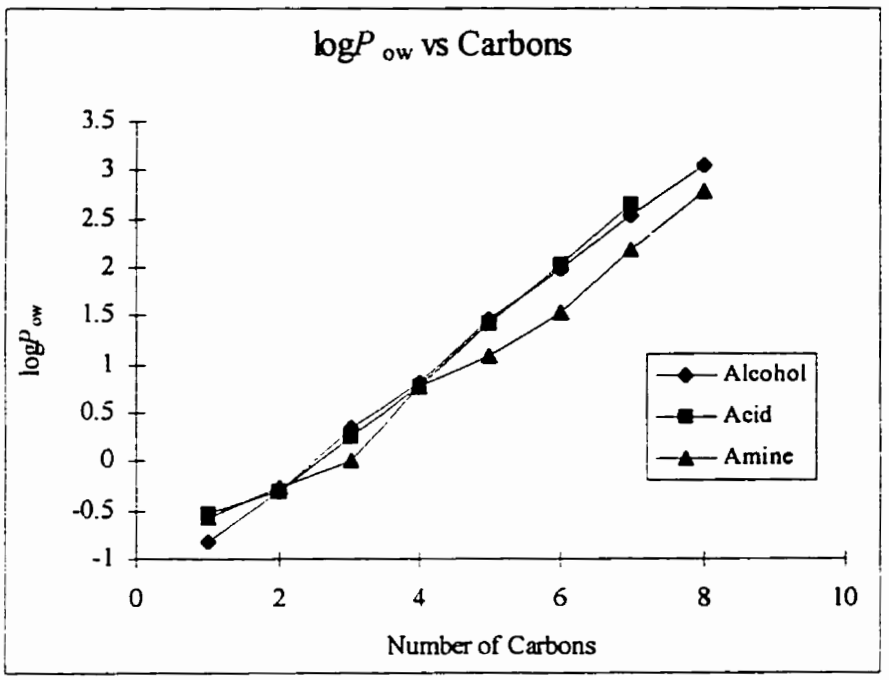


Figure 4-10. Experimental partition coefficients logP_{ow} of aliphatic alcohols, aliphatic carboxylic acids, and aliphatic amines as functions of the number of carbons [Leo *et al.* 1971].

If one assume that the relationship between logP_{ow} and indices L_M and H_M is linear,

$$\log P_{ow} = C_0 + C_1 L_M + C_2 H_M. \qquad (4\text{-}11)$$

Then a fair correlation coefficient and standard deviation are obtained for 23 molecules of the three types of compounds, r=0.833, σ=0.698, using screening function 1 and γ=2.0. Similar results are obtained using screening function 2 and d_0=2.0. I am not satisfied with these results, however, maybe the experimental logP_{ow} data are too poor. (See Fig. 4-10.) The results of correlation calculations of logP_{ow} for each family of the three types of

compounds (aliphatic alcohols, aliphatic carboxylic acids, and aliphatic amines) are much better than for miscellaneous compounds. In §4.4, I will discuss this observation.

4.4 Discussions and Conclusions

In this section, I will analyze the results of § 4.3 and try to find criteria for the optimization of the parameters, judge the effects of the three screening functions, and draw some conclusions.

4.4.1 Distance-dependent Functions in Screening Functions

Fig. 4-11 shows the curves for the two types of distance-dependent functions used in screening functions 1 and 2. One is a power function used in screening function 1, $1/\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}$, the other is an exponential function used in screening function 2, $exp(-\|\mathbf{R}_i-\mathbf{r}\|/d_0)$. Actually, the distance-dependent function in screening function 3, $exp[-\xi(\|\mathbf{R}_i-\mathbf{r}\|/|b_i|)]$, is the same as in screening function 2 if one takes $b_i=1$ and $\xi=1/d_0$. In the defining equation of HMLP, eq. (3-6), the summation does not include the atom α on which point $\mathbf{r}$ is located. Therefore, there is an exclusive region around point $\mathbf{r}$. Typically, in atomic units, the radius of the exclusive region is 2~4 a_0 (1 a_0= 0.509 Å). The largest difference between exponential and power distance-functions is around zero ($\mathbf{r} \sim 0$), where the exclusive region originates. Outside this exclusive region, the two types of distance-dependent functions are very similar. However, there still are two notable differences. First, the exponential distance-function has larger values at short distances than the power distance-function. This means that the exponential distance-function is more influenced by its nearest neighbors than the power distance-function. Second, the power distance-function has larger values at long distances than the exponential distance-function. This means that the power distance-function is more affected by atoms at longer distances.

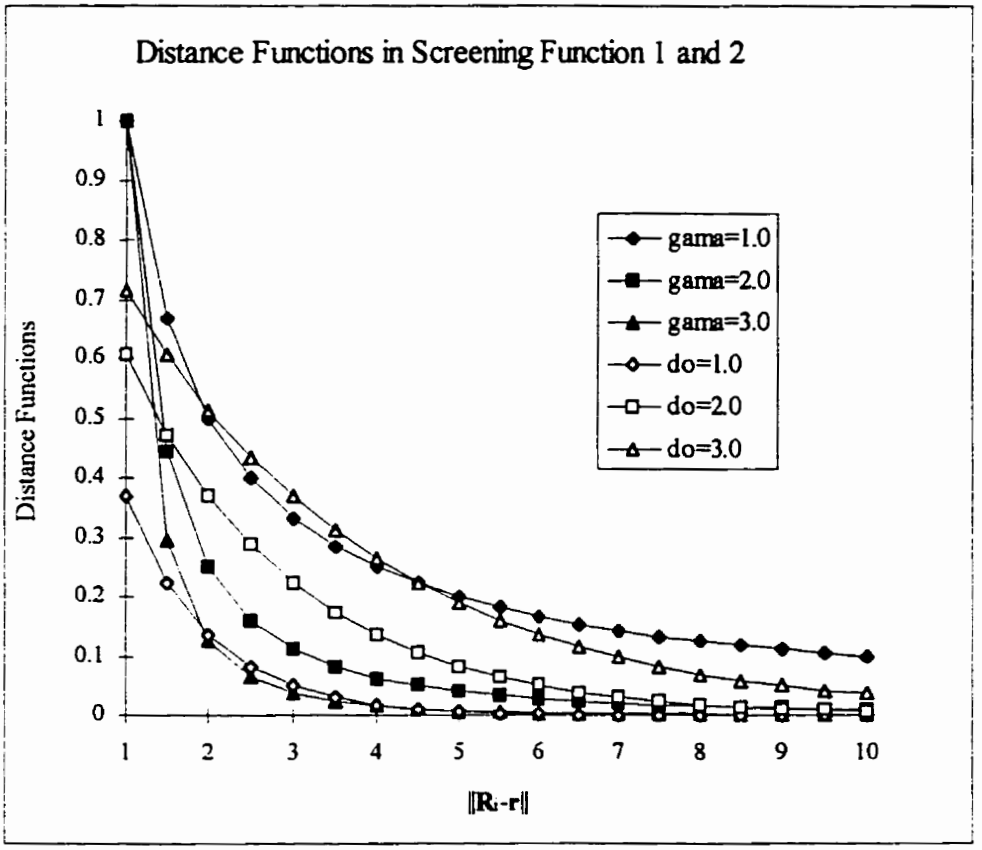


Figure 4-11. The curves of the power and exponential distance-functions in screening functions 1 and 2, using different values of parameters γ, d_0. The black symbols are for the power distance-function $1/\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}$, and the white symbols are for the exponential distance-function $exp(-\|\mathbf{R}_i-\mathbf{r}\|/d_0)$. The curves of the distance-function in screening function 3 for ξ=1, 1/2 and 1/3 are the same as those of the exponential distance-function in screening function 2 for d_0=1, 2 and 3.

From the results of calculations in §4.3, I can say that both screening functions 1 (eq. (4-8)) and 2 (eq. (4-9)) give basically chemically reasonable results. However, the results of screening function 3 (eq. (4-10)) are not good. The distance-dependent functions in screening functions 2 and 3 are the same, however, the mathematical roles of atomic surface-MEP descriptors b_i's in the two screening functions are different. In screening function 2, b_i is a factor of the distance-dependent function, while in screening function 3, b_i is part of the exponent in the exponentially distance-dependent function,

$exp[-\xi(\|\mathbf{R}_i-\mathbf{r}\|/|b_i|)]$. In some cases, the b_i's are too small to provide non zero values. For example, in propane, both carbon and hydrogen atoms have small atomic surface-MEP descriptors b_i's, therefore, their atomic lipophilicity indices are almost zero. (cf. Fig. 4-3 (d)). For the hydrophilic groups, e.g., the hydroxyl group OH, the carboxyl group COOH, and the amino group NH_2, the results of screening function 3 are not bad because of the big atomic surface-MEP descriptors in the functional groups. However, for the hydrocarbon chains of the three compounds (ethanol, propionic acid, and ethylamine), the screening function 3 does not give reasonable results because of very small atomic surface-MEP descriptors b_i's. Molecular lipophilic indices, L_M, of the three types of compounds (alcohols, acids, and amines) do not increase much with increasing number of carbon atoms in the hydrocarbon chains, using eq. (4-10) as the screening function; cf. Fig. 4-7. This is a fatal fault with screening function 3. However, when parameter ξ becomes smaller, the results are better. As possible future research, more calculations and optimizations could be done for a more detailed investigation of the screening function 3.

In both screening functions 1 (eq. (4-8)) and 2 (eq. (4-9)), atomic surface-MEP descriptors b_i's are factors of the distance-dependent functions. The difference is the distance-dependent functions. Screening function 1 uses a power decay function, $1/\|\mathbf{R}_i-\mathbf{r}\|^\gamma$, and screening function 2 uses an exponential decay function, $exp(-\|\mathbf{R}_i-\mathbf{r}\|/d_0)$. In Figs. 4-1 and 4-2, the atoms in the hydrophilic functional groups, OH, COOH, and NH_2, have negative atomic lipophilicity indices, and the atoms in the lipophilic hydrocarbon chains have positive atomic lipophilicity indices. Based on Figs. 4-1 and 4-2, the parameters γ and d_0 in eqs. (4-8) and (4-9) affect the values of various indices in two different ways: the values of the indices decrease with increasing γ in screening function 1, and increase with increasing d_0 in screening function 2. In the carboxyl functional group COOH, there are 3 strongly charged atoms, therefore their atomic lipophilicity indices are very sensitive to the values of γ and d_0. I find an order of magnitude change of the indices l_H, l_O, and $l_{=O}$, caused by changes in the values of parameters γ and d_0; cf. Fig. 4-4 (a) and (c). In the range $2.0<\gamma<4.0$ for eq. (4-8), the order of the indices is $l_H<l_{=O}<l_O$. In the range $1.5<d_0<2.6$ for eq. (4-9), the order of the indices is the same as in eq. (4-8). In a carboxyl group, H tends to become more like the hydronium ion H_3O^+. It should be the most

hydrophilic atom among the three atoms. Carbonyl oxygen, =O, has two lone electron pairs and a widely exposed surface area, therefore it is more hydrophilic than the oxygen in the hydroxyl group, OH. The order $l_H<l_{=O}<l_O$ is reasonable, and it gives us a range for the optimization of parameters γ and d_0. In the amino group NH_2, there are three heavily charged atoms, and two hydrogens have the same sign of MEP. Their atomic lipophilicity idices are sensitive to the parameters γ and d_o, too. An interesting thing is that for both γ and d_0, there is a minimum in the atomic lipophilicity indices l_H. In the optimization of γ in screening function 1, the minimum of the index l_H is -0.0113, at γ=2.3. In the optimization of d_0, a minimum, -0.0411, in index l_H is found at d_0=1.9; cf. Fig. 4-4 (b) and (d). The minimum in l_H means that the effects from neighboring atoms are a maximum. This phenomena provides a useful suggestion for the selection of parameters γ and d_o. However, it is not the only condition for the optimizations of parameters γ and d_o.

4.4.2 Effectiveness of Indices L_M and H_M

For a long time, an additivity scheme of fragmental or atomic contributions to molecular lipophilicity has been used in the calculation of molecular partition coefficients. It is the basic principle of empirical MLP, which is the summation of atomic lipophilicity indices; cf. eq. (3). Israelachvili [1992 p. 135] said "The hydrophilic and hydrophobic interactions, unlike electrostatic and dispersion interactions, are interdependent and therefore not additive". It is obvious that simple additivity is not accurate enough for the calculation of molecular lipophilicity. However, much experimental evidence shows that the whole molecular lipophilicity is roughly the sum of all constituent fragments or atoms [Ghose and Crippen 1986]. Here I do not wish to discuss and check the additivity of molecular lipophilicity, rather I focus on the validity of using partition coefficients as a criterion for HMLP, and discuss the meaning of the two indices: molecular lipophilic index L_M and molecular hydrophilic index H_M.

In the defining equations (3-10) and (3-11), L_M is the sum of atomic lipophilicity indices of all lipophilic atoms (l_a>0), and H_M is the sum of atomic lipophilicity indices of all hydrophilic atoms (l_a<0). MLP is a three dimensional representation of molecular lipophilicity. The indices L_M and H_M are the simplified and approximate representation of

MLP. A rough picture is that the atomic lipophilicity indices l_a's tell us the local lipophilicity on a molecule, while the molecular lipophilic index L_M is the overall measure of the lipophilic part of a molecule, and the hydrophilic index H_M is the overall measure of the hydrophilic part of a molecule, respectively. Reducing the three dimensional molecular lipophilicity potential into two simple scalar descriptors, L_M and H_M, is a big approximation, therefore the two indices are rather rough measures of molecular lipophilicity and hydrophilicity. However, these two indices are very convenient and useful for the approximate comparisons of molecular lipophilicity.

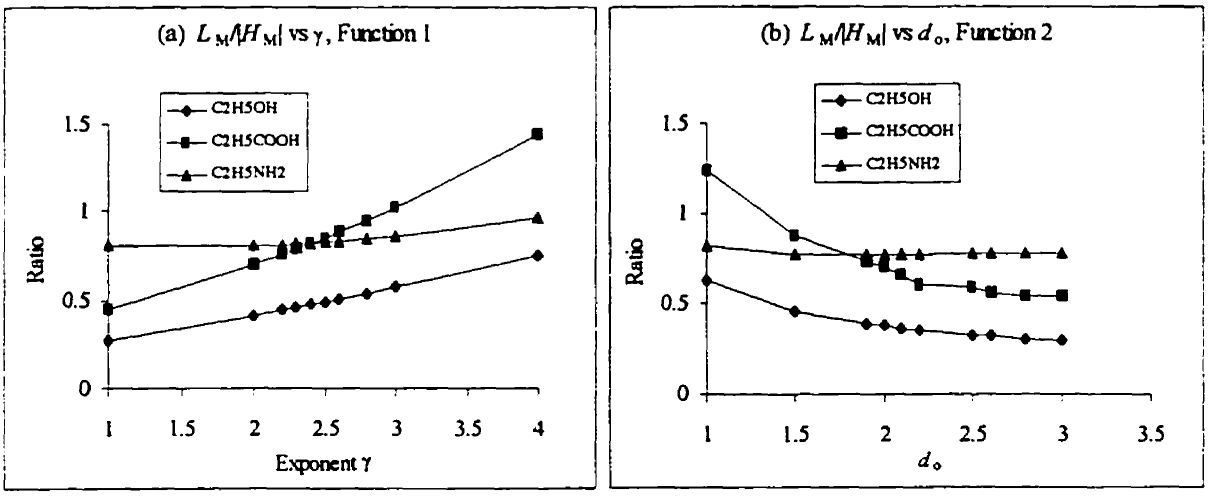


Figure 4-12. Ratio of indices $L_M/|H_M|$ of ethanol, propionic acid, and ethylamine as functions of parameters γ and d_o in screening functions 1 (a) and 2 (b).

Figs. 4-1, 4-2 and 4-3 show atomic lipophilic indices l_a, L_M and H_M as functions of parameters (γ, d_0, and ξ) for ethanol, propionic acid, and ethylamine. Fig. 4-12 shows the ratio of $L_M/|H_M|$ for these three compounds as functions of parameters γ and d_o. From these figures, I find that parameters (γ, d_0, and ξ) affect both the values and ratio of L_M and H_M. Fig. 4-12 (a) tells us that the absolute values of L_M and H_M decrease with increasing γ, but the ratio $L_M/|H_M|$ increases with increasing γ in screening function 1 for ethanol and propionic acid. The ratio $L_M/|H_M|$ of ethylamine has a minimum at γ=2.3. In Fig.4-12 (b), for screening function 2, the absolute values of L_M and H_M increase with increasing d_o, but the ratio $L_M/|H_M|$ of propionic acid decreases with d_o, and the ratio

$L_M/|H_M|$ of ethylamine has a minimum at d_o=1.9. The order of the ratio $L_M/|H_M|$ for the three molecules changes with the parameters γ and d_0.

Figs. 4-8 and 4-9 compare values of the L_M and H_M indices of the three series of compounds for various values of γ and d_o. The molecular hydrophilic index H_M is a rough measure of the hydrophilic strength of the functional groups in a molecule. Figs. 4-8 and 4-9 show that in the ranges $1.0<\gamma<3.0$ and $1.5<d_0<3.0$ the order of H_M using screening functions 1 and 2 is the same: acids<alcohols<amines. This may be a reasonable order for the three functional groups (COOH, OH, and NH_2). This result provides another criterion for the selection of parameters γ and d_0. As mentioned before, the molecular lipophilic index L_M is a rough measure of the lipophilic strength of the lipophilic part of a molecule. As shown by Figs. 4-8 and 4-9, the increase of L_M with the number of carbon atoms is approximately linear for the three types of compounds, albeit with different slopes. The order of the slopes of L_M for the three compounds is alcohols< amines< acids when $1.5<\gamma<3.0$ using screening function 1. The same order is found in Fig. 4-9 in the range $1.5<d_o<3.0$ using screening function 2. This means that the hydrocarbon chains of aliphatic acids are the most lipophilic and the hydrocarbon chains of aliphatic alcohols are the least lipophilic among the three types of compounds. In Fig. 4-12, I find the ratio $L_M/|H_M|$ of ethanol is the smallest for the three compounds when using these two screening functions.

4.4.3 Experimental Criteria for Screening Functions

Partition coefficients log P_{ow} have been used as the overall measure of molecular lipophilicity for a long time. As an equilibrium constant for a two-phase system, log P_{ow} is determined by the difference of the solvation free energies $\Delta G°$ of the solute in each phase. This difference is represented by the molar standard free energy of transfer $\Delta \overline{G}_{tr}^{o}$, from the aqueous phase to the organic phase,

$$\Delta \overline{G}_{tr}^{o} = \Delta \overline{H}_{tr}^{o} - T\Delta \overline{S}_{tr}^{o}. \qquad (4\text{-}12)$$

Common chemical knowledge is: a larger positive $\log P_{ow}$ means the molecule is more lipophilic; otherwise, a larger negative $\log P_{ow}$ means the molecule is more hydrophilic. Transfer-free energy has two components, enthalpy and entropy. As pointed out by Israelachvili [1992 p. 284], lipophilic interaction is an entropic controlled phenomenon. Similarly, one can say that hydrophilic interaction is an enthalpic controlled phenomenon. A good guess is that there is a close relationship between the molecular lipophilic index L_M and transfer entropy $\Delta\bar{S}^{o}_{tr}$, and a close relationship exists between the molecular hydrophilic index H_M and transfer enthalpy $\Delta\bar{H}^{o}_{tr}$. Many authors [Cabani and Gianni 1979, Cabani *et al.* 1981] point out that values of partition coefficients are not very sensitive to the changes of molecular structures. The reason is that sometimes the change of the two components of transfer-free energy, enthalpy and entropy, caused by changes in chemical structure, cancel each other. Partition coefficients are also affected by the intramolecular hydrogen bonding. Therefore, partition coefficients cannot be used as a good criterion for HMLP in the comparison of miscellaneous compounds, however, in a family of compounds, partition coefficients are good criterion for HMLP.

HMLP gives a local description of molecular lipophilicity distributions in a molecular surface or space, and atomic lipophilic indices l_a's are the representations of atomic lipophilicity. The experimental method that is best suited to provide the answers to the calculation results of HMLP is the nuclear magnetic resonance (NMR) [Lee and Rossky 1994], which gives information about water structure near certain atomic groups [Piculell 1986, Piculell and Halle 1986, Halle and Piculell 1986, Halle 1981]. However, in this study, I have not compared the calculation results with experimental data of NMR, because the experimental data are not available. Besides NMR, there are several other experimental methods that may be used to get information for the improvement of HMLP. The methods most often used for the description of molecular motions in a solution are infrared absorption, Raman and Rayleigh scattering, coherent and incoherent neutron scattering, dielectric and Kerr relaxation, and fluorescence depolarization. Measurements of heat capacities also can provide useful information for the development of a quantitative HMLP. More work should be done to check the calculation results of HMLP with various experimental data.

Generally speaking, both screening functions 1 and 2 give good results, and there is no big difference between the two screening functions. For screening function 1, the suitable range of parameter γ is $1.0<\gamma<3.0$, and $\gamma=2.3$ is recommended, while $\gamma=2.0$ is the nearest integral value. For screening function 2, the suitable range of parameter d_0 is $1.5<d_0<3.0$, and $d_0=2.0$ is recommended. Further comparison of screening functions 1 and 2 needs more accurate and well designed experiments. Table 4-1 gives a summary of the optimization of parameters γ and d_0 in the screening functions 1 and 2.

Table 4-1. Summary of the optimization of parameters γ and d_0 in screening functions 1 and 2.

Criteria	γ in Function 1	d_0 in Function 2
$l_H<l_{=O}<l_O$ in COOH	$2.0<\gamma<3.5$	$1.5< d_0 <2.6$
Minimum of l_H in NH_2	$\gamma=2.3$	$d_0 = 1.9$
Minimum of $L_M/\|H_M\|$	$\gamma=2.3$	$d_0 = 1.9$
Order of H_M's Acids<alcohols<amines	$1.0<\gamma<3.0$	$1.5< d_0 <2.6$
Best value	$\gamma=2.3$	$d_0 = 2.0$

Chapter 5: An Application of HMLP to a Small Molecular System of Pyrazole and Its Derivatives

Summary

In this chapter, heuristic molecular lipophilicity potential (HMLP) is used in the study of a real drug molecular system—pyrazole and its derivatives. HMLP's are calculated using a program developed by the author based on the *ab initio* electron density ρ(r) calculated by Gaussian 92 at RHF/6-31G* level. The molecular lipophilic index, L_M, the molecular hydrophilic index, H_M, atomic lipophilicity indices l_a's, and lipophilic indices l_s's and hydrophilic indices h_s's of substituents (atomic groups) are used in the study of QSAR (quantitative structure activity relationship) in this small molecular system. HMLP indices are classified into three types: *endogenous, exogenous, and general indices*. In the correlation between molecular biological activities and HMLP indices, a new strategy is developed for the analysis and recognition of small molecular systems using various HMLP indices. Multiple linear regression (MLR), variance analysis, and principal component analysis (PCA) are used in the study of relationships between various indices of HMLP and the biological activities of molecules. In the multiple linear regression, the best result is achieved using a combination of all three types of variables. Various statistical criteria are constructed for the judgment of the data matrix and the prediction model. The mechanism of inhibition of LADH caused by pyrazole and its derivatives is explained based on the calculated results of HMLP indices. The strategy developed in this research is a potential tool in the combinatorial chemistry of small molecules, QSAR studies, and computer-aided rational drug design.

5.1 Introduction

In computer-aided rational drug design, there are two very challenging questions. (1) What molecular properties should be used to screen the lead drug molecules: molecular shape? electrostatic potential? molecular lipophilicity? or others? (2) How does one correlate the molecular biological activities with their molecular properties? At the points of a cubic grid in molecular space? On the molecular surface? Or on several atoms or fragments of the molecule?

5.1.1 Three Types of Molecular Interactions in Ligand-Receptor Complex

For the first question, it is obvious that all three types of properties (steric, electrostatic and lipophilic) are important [Bone and Villar 1995, Kollman 1994, Balbes *et al.* 1994]. In rational ligand design, it is becoming accepted that consideration should be given to a combination of steric, electrostatic, and lipophilic factors [Bone and Villar 1995]. Since the introduction of CoMFA (Comparative Molecular Field Analysis) nine years ago, it has become the dominant technique in the study of QSAR and drug design [Cramer *et al.* 1988]. In the standard CoMFA program, only steric (van der Waals) and electrostatic potential energy fields are used [Waller and Kellogg 1996]. Great efforts have been made to include molecular lipophilicity potential in the drug design strategy [Klebe *et al.* 1994, Jain *et al.* 1994, Leow *et al.* 1993, Kellogg and Abraham 1992, Cramer *et al.* 1988]. Except for the molecular lipophilicity field, more and more molecular potential fields are included in CoMFA, such as the H-bonding field [Kim 1993], the desolvation field [Gilson and Honig 1987], the molecular orbital field [Waller and Marshall 1993], and the electrotopological field [Waller and Kellogg 1996].

In the CoMFA technique, steric and electrostatic potential energies are stored in one molecular field, which is obtained from the potential energy-calculations between an sp^3 probe carbon atom charged +1 and a molecule of ligand based on the 6-12 Lennard-Jones potential function and Coulombic law. All other fields are individual potential energy fields. Usually, chemists correlate biological activities with various molecular potential fields one by one. First, biological activities are correlated with steric and electrostatic potential fields by partial least square (PLS) regression, then the remaining

biological activities are correlated with the second potential field, and so on [Waller and Kellogg 1996]. It is obvious that if one changes the correlation order, the results may be different.

5.1.2 Correlation Activities with Molecular Structure

A basic principle in molecular modeling is that all molecular properties are decided by the molecular structure. Therefore, there are certain relationships between molecular biological activities and molecular structure. In drug design, if the 3-dimensional structure of a receptor of drugs is not available experimentally, it is often useful to collect structural information of the active site on the receptor from the structures of ligands. The techniques that are used for this purpose can be divided into three categories [Walters 1996]:

1) Models based on points of a cubic grid surrounding one or more active compounds;
2) Models based on a surface constructed over one or more active compounds;
3) Models based on a set of atoms or fragments (amino acid side chains) surrounding one or more active compounds.

Cubic grid-based models are the starting point of CoMFA developed by Cramer *et al.* [1988]. A regular three-dimensional lattice is set up, large enough to surround all of the compounds of interest, and with 2.0Å separation between lattice points. Field properties are calculated with respect to the ligands at each grid point. Quantitative structure-activity relationship (QSAR) methods such as principal component analysis and partial least squares are used to carry out a statistical analysis of the relationship between field interaction energies and biological activities. There are two inherent problems in using the CoMFA model in a practical study. One of them is that the number of variables (field descriptors at all of the grid points) is much larger than the number of compounds in the training set. Too many variables may cause over correlation and are hardly expected to give an explanation for the correlation results. The second problem is that the values of potential energies at grid points may be too different in magnitude to give a

good correlation result. Some grid points are deep inside atoms, some points are on the molecular surface, and some points are far away from the molecular surface. For the field of MEP, the values of points inside an atom may be as high as 10^4, however, the values of points outside atoms may be as small as 10^{-4}. It is difficult to give a reliable correlation result from such a data set, and it is also difficult to give a reasonable explanation for the correlation results based on the huge numbers of variables.

A surface model is regarded as the simplest approach for receptor model construction. Molecular surface is the interface of molecular interactions, therefore, more attention should be placed on this part in a molecule. In this model, a series of active compounds are superimposed, followed by the construction of van der Waals surface surrounding all active regions in the set. Such surfaces convey the steric requirements of the receptor binding site. With computer graphics, it is possible to map properties such as electrostatic potential onto this surface, to provide some information about the electronic properties of active analogs. This approach was used successfully in building a receptor model for high potency sweeteners by Culberson and Walters [1991].

In the atom-based and fragment-based models, a receptor model can be made by placing atoms or groups (such as amino acid side chains) around a set of active ligands. In the case that amino acid side chains are used as the fragments, certain amino acid side chains are docked into the region based on the biological activity analysis. Whenever one tries to construct a receptor model using atoms or fragments, one is immediately struck by the arbitrary nature of the choices that must be made. In many cases, the number of possible models that could be constructed is beyond comprehension. This is a highly combinatorial problem for which a systematic solution seems impossible.

5.1.3 Indices of HMLP

In this research, I explore the two questions given in §5.1, using HMLP introduced in Chapters 3 and 4 [Du *et al.* 1997]. For convenience, I rewrite the definition equations of HMLP below,

$$\underset{\mathbf{r}\in S_a}{L(\mathbf{r})} = V(\mathbf{r})\sum_{i\neq a} M_i(\mathbf{r};\mathbf{R}_i,b_i), \tag{3-6}$$

where $V(\mathbf{r})$ is the molecular electrostatic potential (MEP) at point **r**, and **r** is on the surface S_a of atom a. In the sum, $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$ is the screening function on position **r** from atom *i*. In eq. (3-6) summation is over all constituent atoms except atom a. In the screening function $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$, $\mathbf{R}_i$ is the nuclear position of atom *i*, and b_i is the atomic surface-MEP descriptor of atom *i* [Du and Arteca 1996],

$$b_i = \sum_{k\in S_i} V(\mathbf{r}_k)\Delta s_k\ , \tag{3-7}$$

where Δs_k is the area element on the surface of atom *i*. Summation is over all exposed surfaces of atom *i*. As discussed in Chapter 4, the atom-based screening functions can involve a number of functions. Here I use the power distance-dependent function,

$$M_i(\mathbf{r};\mathbf{R}_i,b_i) = \frac{r_0^{\gamma}}{b_0}\frac{b_i}{\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}} = \varsigma\frac{b_i}{\|\mathbf{R}_i-\mathbf{r}\|^{\gamma}}. \tag{4-8}$$

In eq. (4-8), r_0, b_0, and g are parameters. The unit of b_0 is the same as b_i (energy•area); r_0 has a unit of length. Therefore, $M_i(\mathbf{r};\ \mathbf{R}_i,\ b_i)$ is a dimensionless function, and the unit of HMLP $L(\mathbf{r})$ is the same as MEP $V(\mathbf{r})$. In eq. (4-8), $\zeta=(r_0)^{\gamma}/b_0$ is a simple scaling factor. In the later calculations, I take ζ=1, and exponent γ=2.5 based on the optimization results in Chapter 4.

HMLP is a three dimensional, structure-based, unified molecular lipophilicity and hydrophilicity potential, requiring no empirical indices of atomic lipophilicity. The HMLP has an advantage in answering the first question: three types of interactions (steric, electrostatic, and lipophilic) in the ligand-receptor complex are included in one field. For the second question, I combine the grid-based model, the surface-based model, and the atom-based and fragment-based models into one model, and correlate molecular biological activities with various HMLP indices: the molecular lipophilic index, L_M, the

hydrophilic index, H_M, atomic lipophilicity indices l_a's, lipophilic indices l_s's, and hydrophilic indices h_s's of substituents (atomic groups).

In Chapter 3, I introduced three HMLP indices: atomic lipophilicity indices l_a's, the molecular lipophilic index L_M, and the molecular hydrophilic index H_M. The atomic lipophilicity index l_a is defined as,

$$l_a = \sum_{i \in S_a} L(\mathbf{r}_i)\Delta s_i \ , \tag{3-9}$$

where the summation is over all the exposed area S_a of atom a. If $l_a>0$, then atom a is lipophilic, whereas, if $l_a<0$, then atom a is hydrophilic. The molecular lipophilic index (L_M) and the hydrophilic index (H_M) are the sum of the corresponding values for all lipophilic atoms and hydrophilic atoms, respectively,

$$L_M = \sum_{\substack{a \\ (l_a>0)}} l_a \ , \tag{3-10}$$

$$H_M = \sum_{\substack{a \\ (l_a<0)}} l_a \ . \tag{3-11}$$

In order to study the effects of the lipophilicities of substituents, here I define the lipophilic indices l_s's and hydrophilic indices h_s's of substituents (atomic groups). The atomic group lipophilic index l_s is the sum of all positive atomic lipophilic indices in a substituent, and the atomic group hydrophilic index h_s is the sum of all negative atomic lipophilic indices in a substituent,

$$l_S = \sum_{\substack{a \\ (l_a>0)}} l_a \ , \tag{5-1}$$

$$h_S = \sum_{\substack{a \\ (l_a<0)}} l_a \ . \tag{5-2}$$

5.2 Calculations of Pyrazole and its Derivatives

Today drug developers generally prefer to focus on small organic molecules with molecular weights of about 500 daltons or less—the class of compounds from which most successful drugs have traditionally emerged [Borman 1996]. Several laboratories have made great efforts to create combinatorial libraries for small molecules [Bunin and Ellman 1992, Murphy *et al.* 1995]. The direction in which combinatorial chemistry is headed is to combine combinatorial chemistry with computational drug-design strategies for molecular recognition. This research aims to contribute a new strategy for drug design and combinatorial chemistry by the study of small molecules based on the use of heuristic lipophilicity potential.

5.2.1 Calculation Algorithm

Eighteen molecules of pyrazole derivatives are used in this research. Molecular geometries are optimized using Gaussian 92 at the RHF/STO-3G level. Molecular van der Waals surfaces are built using the program MS [Connolly 1983 a, b, 1985], and atomic radii are optimized based on MEP criteria [Du and Arteca 1996 a]. Molecular surfaces are described by sets of grid points, and point density is 25 points/Å^2. Electrostatic potentials, V(r), at grid points are calculated by Gaussian 92 using RHF with the basis set 6-31G*. Then atomic surface-MEP descriptors b_i's and molecular lipophilicity potential L(r)'s are calculated based on eqs. (3-6) and (3-7). The power screening function eq. (4-8) is used in the HMLP calculations. Various indices of HMLP are calculated according to eqs. (3-9), (3-10), (3-11), (5-1), and (5-2) using the program developed by the author. Finally, multiple linear regression (MLR), variation analysis, and principal component analysis (PCA) are carried out between HMLP indices and the biological activities of molecules.

5.2.2 Pyrazole and its Derivatives

Pyrazole and its derivatives form a small molecular system that has been studied extensively through experiments and drug design theories. Pyrazole is known to be a potent inhibitor of the enzyme liver alcohol dehydrogenase (LADH), and is applied in the

study of alcohol metabolism [Rozas *et al.* 1995, 1992, 1991, Cornell *et al.* 1983, Tolf *et al.* 1979, Dahlbom *et al.* 1974, Theorell *et al.* 1969]. After the discovery of the inhibition of LADH by pyrazole, it was found that the inhibitory power of pyrazole and its derivatives is affected by the properties of substituents and their positions of those on the pyrazole ring (see Fig. 5-1). It was found that the lipophilic substituents on position 4 enhance the inhibitory power. Studies also show that hydrophilic substituents, such as -COOH, on position 4 decrease the inhibitory power. On the other hand, the inhibitory activity is lowered by both lipophilic and hydrophilic substituents on position 3 [Theorell *et al.* 1969, Fries *et al.* 1979, Rozas *et al.* 1991, 1995]. Therefore, the small molecular system, pyrazole and its derivatives, provides a good example for checking my model of HMLP.

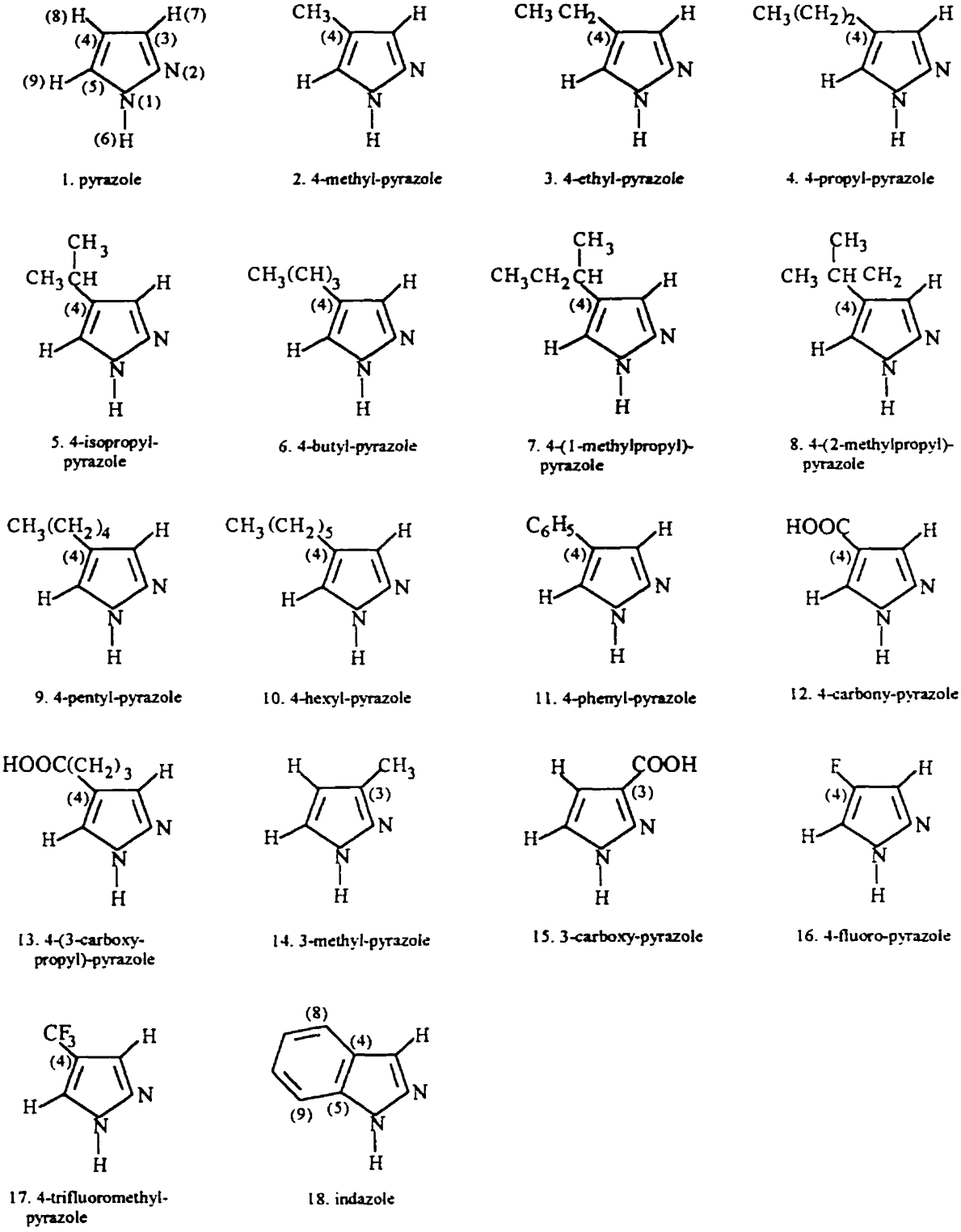


Figure 5-1. Pyrazole and its derivatives used in this research. Indazole is regarded as a derivative of pyrazole. The positions of atoms are shown in the structure of pyrazole. Substituents are either in positions 3 or 4.

5.2.3 Calculation Results

Table 5-1 lists the atomic lipophilicity indices, l_a, substituent lipophilic indices, l_s, hydrophilic indices, h_s, the molecular lipophilic index, L_M, and the hydrophilic index, H_M, of eighteen molecules. The substituents in this research are either on positions 3 or 4. In Table 1, $h_s^{(7)}$ and $l_s^{(7)}$ are the indices of substituents on position 3, which replace the hydrogen H(7) of pyrazole. In the same way, $h_s^{(8)}$ and $l_s^{(8)}$ are the indices of substituents on position 4, which replace the hydrogen H(8) of pyrazole. As shown in Table 5-1, for pyrazole, N(1), C(3), C(4), C(5) and H(9) are lipophilic atoms (l_a>0); N(2), H(6) and H(7) are hydrophilic atoms (l_a<0); and atom H(8) is almost neutral ($l_a^{(8)}\approx 0.0$). Carbon C(5) ($l_a^{(5)}$=0.1268) is the most lipophilic atom, and nitrogen N(2) ($l_a^{(2)}$=-0.09813) is the most hydrophilic atom in pyrazole. For the lipophilic substituents of hydrocarbons, the hydrophilic indices h_s's are zero, and the lipophilic indices increase with increasing number of carbon atoms. However, the carboxy and hcarboxypropyl substituents have both nonzero lipophilic indices, l_s, and hydrophilic indices, h_s. Long ago chemists found that the branched and cyclized substituents are less lipophilic than straight chains with the same number of carbon atoms. These phenomena have been demonstrated by Hansch *et al.* using π substitutent analysis experimentally [Leo *et al.* 1971]. All these well-known phenomena are reflected in Table 5-1, where the lipophilic indices, $l_s^{(8)}$, of the 1-methylpropyl and 2-methylpropyl groups are smaller than the same index of the butyl groups. The index $l_s^{(8)}$ of the phenyl group is smaller than the same index for the hexyl group. All these results are reasonable from a chemical point of view. Table 5-1 shows that different substituents on positions 3 and 4 affect the atomic lipophilicities of pyrazole in different ways and to different degrees.

Table 5-1. Lipophilic and hydrophilic indices of pyrazole and its derivatives*

Molecules	$I_a^{(1)}$	$I_a^{(2)}$	$I_a^{(3)}$	$I_a^{(4)}$	$I_a^{(5)}$	$I_a^{(6)}$	$I_a^{(7)}$
1. pyrazole	0.0204	-0.0981	0.0391	0.0029	0.1268	-0.0302	0.0000
2. 4-methyl-pyrazole	0.0141	-0.0939	0.0425	-0.0084	0.1156	-0.0385	0.0000
3. 4-ethyl-pyrazole	0.0150	-0.0907	0.0472	-0.0077	0.0944	-0.0373	0.0000
4. 4-propyl-pyrazole	0.0165	-0.0861	0.0469	-0.0070	0.0889	-0.0412	0.0000
5. 4-isopropyl-pyrazole	0.0157	-0.0903	0.0463	-0.0068	0.0882	-0.0420	0.0000
6. 4-butyl pyrazole	0.0205	-0.0891	0.0482	-0.0076	0.0926	-0.0370	0.0000
7. 4-(1-methylpropyl)-pyrazole	0.0210	-0.0925	0.0454	-0.0073	0.0917	-0.0412	0.0000
8. 4-(2-methylpropyl)-pyrazole	0.01243	-0.0937	0.0413	-0.00912	0.0985	-0.0349	0.0000
9. 4-pentyl-pyrazole	0.0187	-0.0866	0.0484	-0.0069	0.0923	-0.0414	0.0000
10. 4-hexyl-pyrazole	0.0158	-0.0866	0.0488	-0.0074	0.0907	-0.0426	0.0000
11. 4-phenyl-pyrazole	0.0250	-0.1147	0.0306	-0.0077	0.1535	-0.0096	0.0000
12. 4-carboxy-pyrazole	0.0850	-0.1243	0.0121	-0.0019	0.2207	0.0574	0.0000
13. 4-(3-carboxy-propyl)-pyrazole	0.0236	-0.1266	0.0315	-0.0146	0.1271	-0.0085	0.0000
14. 3-methyl-pyrazole	0.0164	-0.0848	0.0357	0.0026	0.1094	-0.0406	0.0303
15. 3-carboxy-pyrazole	0.0799	-0.0290	0.0128	0.0248	0.2645	0.0575	0.2388
16. 4-fluoro-pyrazole	0.0477	-0.1231	0.0173	0.0031	0.1979	0.0294	0.0000
17. 4-trifluoromethyl-pyrazole	0.1135	-0.1365	0.0268	0.0000	0.2810	0.1097	0.0000
18. Indazole	0.0119	-0.0854	0.0407	-0.0007	0.0355	-0.0414	0.0000

(Continued)

Molecule	$h_s^{(7)}$	$l_s^{(8)}$	$h_s^{(8)}$	$l_a^{(9)}$	L_M	H_M	$log\ K_I$ (expt.)	$log\ K_I$ ‡ (calc.)
1.	-0.0210	0.0000	-0.0001	0.0378	0.2271	-0.1493	-0.660[a]	-0.675
2.	-0.0186	0.0704	0.0000	0.0348	0.2774	-0.1593	-1.886[b]	-1.593
3.	-0.0182	0.1554	0.0000	0.0450	0.3570	-0.1540	-2.155[b]	-1.775
4.	-0.0184	0.1886	0.0000	0.0413	0.3940	-0.1527	-2.398[b]	-1.961
5.	-0.0161	0.1641	0.0000	0.0413	0.3623	-0.1551	-2.097[b]	-1.883
6.	-0.0181	0.2675	0.0000	0.0438	0.4727	-0.1517	-2.745[b]	-2.357
7.	-0.0158	0.2350	0.0000	0.0421	0.4352	-0.1567	-1.854[d]	-2.415
8.	-0.0170	0.2552	0.0000	0.0532	0.4605	-0.1547	-1.886[d]	-2.425
9.	-0.0188	0.3022	0.0000	0.0431	0.5047	-0.1537	-3.097[b]	-2.707
10.	-0.0186	0.3477	0.0000	0.0432	0.5462	-0.1552	-3.523[d]	-3.054
11.	0.0086	0.1025	-0.0091	0.0722	0.3839	-0.1497	-1.000[b]	-0.550
12.	-0.0109	0.0596	-0.2357	0.0631	0.4979	-0.3727	3.000[a]	3.679
13.	-0.0073	0.5396	-0.2482	0.0723	0.7941	-0.4050	1.176[a]	0.533
14.	0.0000	0.0000	-0.0032	0.0272	0.2280	-0.1349	2.097[a]	2.100
15.	-0.1865	0.0126	0.0000	0.1040	0.7949	-0.2155	3.813[a]	3.813
16.	-0.0158	0.0000	-0.0254	0.0679	0.3633	-0.1643	——	-0.042
17.	-0.0022	0.0429	-0.0117	0.1128	0.6867	-0.1503	-1.097[b]	-1.500
18.	-0.0171	0.0266	0.0000	0.1020	0.2168	-0.1447	1.146[a]	-0.404

* The units of the inhibition constant K_I are in μM.

† For indazole, the atomic lipophilicities of the side ring on positions 4 and 5 are divided equally, and assigned to the imaginary atoms 8 and 9, with positions taken from pyrazole.

‡ Calculated by the author using eq. (5-4).

a [Rozas *et al.* 1991].

b [Tolf *et al.* 1979].

c [Dahlbom *et al.* 1974].

d [Tolf *et al.* 1985].

It was thought that the nonprotonated nitrogen N(2) binds to the zinc cation Zn(II) and the protonated N(1) forms a weak bond to the C(4) atom of the nicotinamide ring [Rozas and Arteca 1992, Rozas *et al.* 1991] (see Fig. 5-6). The interaction with the metallic cation seems to be essential for the release of the N-H proton in order to create the bond between pyrazole and NAD [Rozas and Arteca 1992, Rozas *et al.* 1991]. Calculated results of this work show that nonprotonated nitrogen N(2) is the most hydrophilic atom, therefore it is easy to bind with the zinc cation through electrostatic interaction. It was found that the inhibitory power increases by a factor of two for each added methylene group, CH_2, of the hydrocarbon substituents in position 4 [Rozas *et al.* 1991, 1995]. In Table 5-1, the lipophilic indices $I_s^{(8)}$'s of hydrocarbon substituents increase regularly with the number of carbon atoms by an increment of 0.033 per CH_2, consistent with experiments. In the next section, we study the relationship between molecular bioactivities and lipophilicity indices of pyrazole and its derivatives.

5.3 The Relationship between Molecular Bioactivities and Various Indices

In this section, the relationship between molecular bioactivities and lipophilicity indices will be studied using three different methods: multiple linear regression (MLR), variation analysis, and principal component analysis (PCA).

5.3.1 Multiple Linear Regression

The logarithms of experimental inhibition constants, $\log K_I$, are correlated with various indices using linear functions. The experimental inhibition constant of 4-fluorol-pyrazole is not available, therefore seventeen molecules are used in the regression calculation and 4-fluorol-pyrazole is used as an example of prediction. In the first trial,

two parameters, the total molecular lipophilic index L_M and hydrophilic index H_M, are used to correlate with $\log K_I$. A very poor correlation coefficient r=0.551 and standard deviation s=1.969 are obtained. This means that the inhibition constants are not determined by L_M and H_M. However, a much better correlation equation (r=0.947) is obtained using the lipophilic and hydrophilic indices of substituents on positions 3 and 4: $l_s^{(7)}$, $h_s^{(7)}$, $l_s^{(8)}$ and $h_s^{(8)}$,

$$\log K_I = -0.0821(\pm 0.363) + 46.91(\pm 14.27) l_s^{(7)} + 38.25(\pm 18.93) h_s^{(7)} - 6.772(\pm 1.613) l_s^{(8)} - 18.73(\pm 2.809) h_s^{(8)} \quad (5\text{-}3)$$

(n=17, r=0.947, s=0.818, F=26.11, Sig.=0.000).

Experiments show that the inhibitory power of pyrazole and its derivatives is enhanced by lipophilic substituents on position 4. Both lipophilic and hydrophilic substituents on position 3 lower the inhibitory power. The calculated results of multilinear regression illustrate that the lipophilicities of substituents on positions 3 and 4 are important for the inhibition of LADH by pyrazole and its derivatives. These results agree with the experiments by Rozas *et al.* [1991]. There are many different selections of parameters by the combination of the 13 HMLP indices. A much better correlation equation is achieved using six parameters: $l_s^{(7)}$, $h_s^{(7)}$, $l_s^{(8)}$ $h_s^{(8)}$, L_M and H_M (r=0.966),

$$\log K_I = 9.017(\pm 6.905) + 50.29(\pm 14.14) l_s^{(7)} + 9.077(\pm 31.62) h_s^{(7)} - 2.389(\pm 2.538) l_s^{(8)} - 77.93(\pm 48.58) h_s^{(8)} - 3.789(\pm 1.797) L_M + 57.92(\pm 49.41) H_M \quad (5\text{-}4)$$

(n=17, r=0.966, s=0.725, F=23.02, Sig.=0.000).

Table 5-1 lists the $\log K_I$ values calculated using eq. (5-4). A predicted value of $\log K_I$=-0.042 for 4-fluorol-pyrazole is obtained using eq. (5-4). As shown in Table 5-1, $\log K_I$ of indazole has the largest error (1.550). This means that indazole is an exception in this small molecular system. If indazole is omitted, an even better correlation coefficient,

r=0.987, is obtained. Later I will show that there is a strong correlation between $l_s^{(8)}$ and L_M. If one omits $l_s^{(8)}$, the results are still good (r=0.963) using 5 parameters,

$$\log K_I = 13.55(\pm 4.924) + 51.85(\pm 13.97) l_s^{(7)} - 4.703(\pm 27.89) h_s^{(7)} - 108.9(\pm 35.60) h_s^{(8)} - 4.732(\pm 1.484) L_M + 89.40(\pm 36.19) H_M \quad (5\text{-}5)$$

(n=17, r=0.963, s=0.722, F=27.73, Sig.=0.000).

In the above two tests, I use HMLP indices of substituents ($l_s^{(7)}$, $h_s^{(7)}$ and $h_s^{(8)}$) and molecules (L_M and H_M). If two atomic lipophilicity indices on the pyrazole ring, $l_a^{(5)}$ of C(5) and $l_a^{(6)}$ of H(6), which have larger variances as shown in the next part (variation analysis), are added, using 7 indices, the result of regression is further improved (r=0.987, s=0.479, F=47.19),

$$\log K_I = 6.654(\pm 4.326) - 29.68(\pm 7.99) l_a^{(5)} + 51.01(\pm 12.78) l_a^{(6)} + 74.49(\pm 10.95)\, l_s^{(7)} + 51.52(\pm 23.77) h_s^{(7)} - 19.21(\pm 36.00) h_s^{(8)} - 7.81(\pm 1.462) L_M - 0.575(\pm 36.32) H_M \quad (5\text{-}6)$$

(n=17, r=0.987, s=0.479, F=47.19, Sig.=0.000).

Generally speaking, the more parameters used, the better the correlation coefficients obtained. However, correlation equations using too many parameters are meaningless. The selection of parameters for multiple linear regression, in a certain sense, is arbitrary [Bowerman and O'Connell, 1990, Erricker 1971]. In the screening of leading small molecules by multiple linear regression, one of the most important aspects is to make sure the variables used in multiple linear regression are independent, and contain all necessary information. Also we want to know in what qualities that the data base of a group of samples (Table 5-1) describes the nature of a small molecular system. This is helpful for the selection of parameters in multiple linear regression. In the next step, the variation analysis is used to explore the above questions.

5.3.2 Variance Analysis

In Table 5-1 there are n=18 samples (molecules) and m=13 variables (HMLP indices), forming an 18×13 data matrix, $\mathbf{X}_{18\times13}$. The sample mean of the ith variable is

$$\bar{x}_i = \frac{1}{n}\sum_{r=1}^{n} x_{ri}, \tag{5-7}$$

and the sample variance of the ith variable [Mardia *et al.* 1979] is

$$s_{ii} = \frac{1}{n}\sum_{r=1}^{n}(x_{ri} - \bar{x}_i)^2 = s_i^2. \tag{5-8}$$

The sample covariance between the ith and jth variables is

$$s_{ij} = \frac{1}{n}\sum_{r=1}^{n}(x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j). \tag{5-9}$$

The sample correlation coefficient between variables ith and jth is

$$r_{ij} = s_{ij}/(s_i s_j). \tag{5-10}$$

Unlike covariance s_{ij}, the correlation coefficient r_{ij} is invariant under both changes of scaling and origin shift of the ith and jth variables. The matrix

$$\mathbf{S}=(s_{ij}) \tag{5-11}$$

is the covariance matrix , and the matrix

$$\mathbf{R}=(r_{ij}) \tag{5-12}$$

is the sample correlation matrix. Clearly $|r_{ij}|^2<1$ and $r_{ii}=1$. If **R=I**, in this situation, the variables are said to be uncorrelated. Both covariance and correlation matrices are symmetrical.

Tables 5-2 and 5-3 show the sample covariance matrix **S** and the correlation matrix **R** of pyrazole derivatives. Only elements of the lower triangle of the matrices **S** and **R** are shown in Tables 5-2 and 5-3 because of the symmetry of matrices. Comparing the diagonal terms of **S**, the largest variance (s_{12}=2.9042) is on the molecular lipophilic index L_M. This means that the 18 molecules are quite different in lipophilicities. The substituents on position 4 have quite different lipophilic and hydrophilic indices (see Table 5-1). Therefore, variances corresponding to the indices $l_s^{(8)}$ and $h_s^{(8)}$ are the second and third largest (s_9=2.0968 and s_{10}=0.5679). The variances of the atomic lipophilicity indices l_a's of atoms on the pyrazole frame can be thought of as the measurement of influence on the lipophilicity caused by various substituents at positions 3 and 4. The strongest effect is on the atom C(5) (s_5=0.4186) and the second largest effect is on the atom H(6) (s_6=0.1949). Covariance matrix S is one possible multivariate generalization of the invariant notion of variance, measuring scatter about the mean. Covariance matrices are useful in the comparison of several small molecular sets to find out to what extent the data matrices describe the small molecular systems. For this purpose, it is convenient to use a single number to measure multivariate scatter. Two common measurements are (1) generalized variance, $|S|=\Pi_i l_i$, and (2) total variation, $trS=\Sigma_i\ l_i$, where l_i's are eigenvalues of S [Mardia *et al.* 1979, Kleinbaum *et al.* 1988]. For both measurements, large values indicate a high degree of scatter around $\bar{x}$ and low values represent concentration around $\bar{x}$. Generalized variance plays an important role in maximum likelihood estimation and the total variation is a useful concept in principal component analysis.

Table 5-2. Covariance matrix S of pyrazole derivatives*

Indices	$I_a^{(1)}$	$I_a^{(2)}$	$I_a^{(3)}$	$I_a^{(4)}$	$I_a^{(5)}$	$I_a^{(6)}$	$I_s^{(7)}$	$h_s^{(7)}$	$I_s^{(8)}$	$h_s^{(8)}$	$I_a^{(9)}$	L_M	H_M
1. $I_a^{(1)}$	0.0838												
2. $I_a^{(2)}$	-0.0156	0.0533											
3. $I_a^{(3)}$	-0.0268	0.0049	0.0144										
4. $I_a^{(4)}$	0.0128	0.0101	-0.0062	0.0069									
5. $I_a^{(5)}$	0.1752	-0.0285	-0.0659	0.0317	0.4186								
6. $I_a^{(6)}$	0.1254	-0.0314	-0.0439	0.0186	0.2731	0.1949							
7. $I_s^{(7)}$	0.0612	0.0911	-0.0320	0.0383	0.1725	0.0888	0.2996						
8. $h_s^{(7)}$	-0.0387	-0.0712	0.0191	-0.0268	-0.1072	-0.0541	-0.2132	0.1635					
9. $I_s^{(8)}$	-0.1550	-0.0387	0.0773	-0.0792	-0.3659	-0.2280	-0.2133	0.1193	2.0968				
10. $h_s^{(8)}$	-0.0652	0.0858	0.0433	0.0127	-0.1302	-0.1118	0.0438	-0.0415	-0.3741	0.5679			
11. $I_s^{(9)}$	0.0486	-0.0051	-0.0180	0.0094	0.0947	0.0801	0.0556	-0.0376	-0.0825	-0.0310	0.0618		
12. L_M	0.2759	0.0019	-0.0784	0.0233	0.6157	0.4278	0.4283	-0.3137	1.1596	-0.5490	0.2141	2.9042	
13. H_M	-0.0712	0.0554	0.0462	0.0051	-0.1541	-0.1192	-0.0362	0.0230	-0.3952	0.5496	-0.0412	-0.6992	0.5609

* The values of data in Table 5-2 are 100 times the original data (original data are multiplied by a factor of 100).

Table 5-3. Correlation matrix **R** of pyrazole derivatives

indices	$I_a^{(1)}$	$I_a^{(2)}$	$I_a^{(3)}$	$I_a^{(4)}$	$I_a^{(5)}$	$I_a^{(6)}$	$I_s^{(7)}$	$h_s^{(7)}$	$I_s^{(8)}$	$h_s^{(8)}$	$I_a^{(9)}$	L_M	H_M
1. $I_a^{(1)}$	1.0000												
2. $I_a^{(2)}$	-0.2333	1.0000											
3. $I_a^{(3)}$	-0.7731	0.1783	1.0000										
4. $I_a^{(4)}$	0.5321	0.5290	-0.6181	1.0000									
5. $I_a^{(5)}$	0.9353	-0.1909	-0.8487	0.5902	1.0000								
6. $I_a^{(6)}$	0.9808	-0.3078	-0.8295	0.5089	0.9560	1.0000							
7. $I_s^{(7)}$	0.3860	0.7209	-0.4868	0.8432	0.4873	0.3673	1.0000						
8. $h_s^{(7)}$	-0.3303	-0.7625	0.3932	-0.7976	-0.4098	-0.3030	-0.9634	1.0000					
9. $I_s^{(8)}$	-0.3697	-0.1159	0.4448	-0.6589	-0.3906	-0.3566	-0.2692	0.2037	1.0000				
10. $h_s^{(8)}$	-0.2987	0.4931	0.4785	0.2030	-0.2670	-0.3359	0.1061	-0.1362	-0.3428	1.0000			
11. $I_s^{(9)}$	0.6750	-0.0881	-0.6038	0.4551	0.5886	0.7301	0.4085	-0.3736	-0.2292	-0.1655	1.0000		
12. L_M	0.5592	0.0049	-0.3836	0.1650	0.5584	0.5685	0.4592	-0.4552	0.4699	-0.4274	0.5053	1.0000	
13. H_M	-0.3284	0.3206	0.5139	0.0824	-0.3180	-0.3606	-0.0884	0.0758	-0.3644	0.9737	-0.2212	-0.5478	1.0000

Correlation matrix **R** describes the relationship among these indices. From the first column of Table 5-3, one finds that the atomic lipophilic index $I_a^{(1)}$ of atom N(1) correlates highly with the atomic lipophilic indices $I_a^{(5)}$ (0.9353) and $I_a^{(6)}$ (0.9808). Between $I_a^{(1)}$ and $I_a^{(3)}$ there is a negative correlation (-0.7731). Furthermore, the atomic lipophilic indices $I_a^{(5)}$ and $I_a^{(6)}$ also correlate highly with each other. Table 5-3 tells us that the molecular hydrophilic index H_M is mainly determined by $h_s^{(8)}$, the hydrophilic index of substituents on position 4, because the correlation coefficient between H_M and $h_s^{(8)}$ is 0.9737. Also, substituents on position 3 affect the lipophilicities of atoms N(2) and C(4) more than other atoms, because the correlation coefficients between $I_s^{(7)}$ and $h_s^{(7)}$ to $I_a^{(2)}$ and $I_a^{(4)}$ are much higher than others. All this information is useful for the selection of parameters in the multiple linear regression. Figure 5-2 shows the correlations of $I_a^{(1)}$ to $I_a^{(3)}$, $I_a^{(5)}$ and $I_a^{(6)}$.

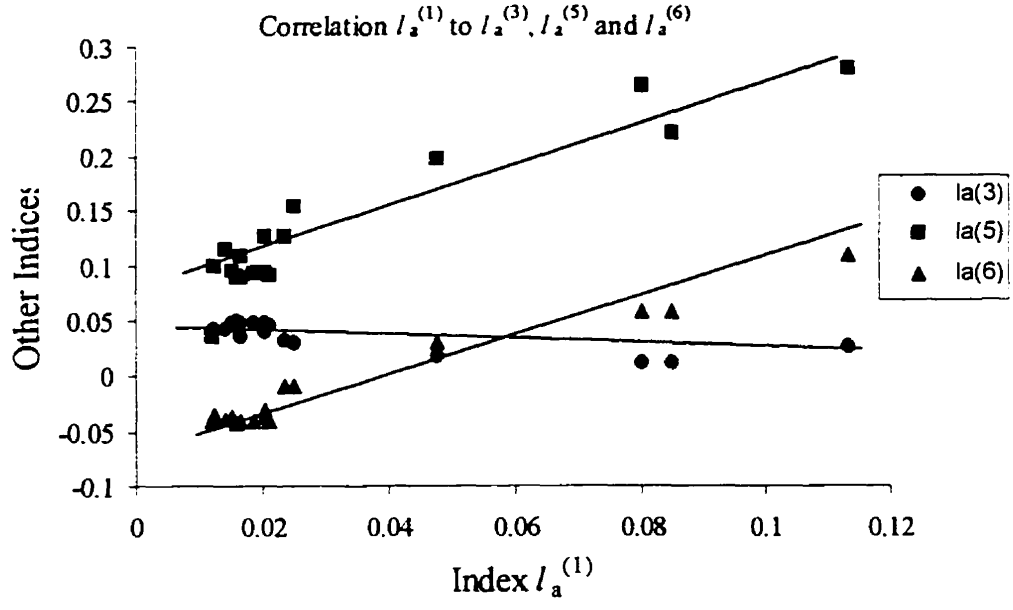


Figure 5-2. Correlations of $I_a^{(1)}$ to $I_a^{(3)}$, $I_a^{(5)}$ and $I_a^{(6)}$. The correlations of $I_a^{(1)}$ to $I_a^{(5)}$ and $I_a^{(6)}$ are positive, however, the correlation of $I_a^{(1)}$ to $I_a^{(3)}$ is negative

5.3.3 Principal Component Analysis

Principal component analysis (PCA) uses a linear transformation of original variables to reduce the dimensions of the data matrix, and simplify the structure of the covariance matrix, making the interpretation of data more straightforward. In this research, principal component analysis is performed using nonlinear iterative partial least squares (NIPALS) [Geladi and Kowalski 1986 a, b]. Data matrix $\mathbf{X}_{n\times m}$ is the mean-centered HMLP indices of pyrazole and its derivatives as shown in Table 5-1, where n=18 is the number of molecules and m=13 is the number of HMLP indices. If the rank of matrix **X** is r and $r \le n$ and m, then **X** can be written as the sum of r matrices of rank 1,

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 + \ldots\ldots + \mathbf{M}_r. \tag{5-13}$$

Rank is a number expressing the true underlying dimensionality of a matrix. Each rank 1 matrix $\mathbf{M}_i$ is the outer product of two vectors, a score vector t_i and a loading vector p_i,

$$\mathbf{M}_i = t_i p_i'. \tag{5-14}$$

If the rank 1 matrices with small norm $\|\mathbf{M}_i\|$ are thought of as the random error, then the original data matrix can be expressed as the product of the score matrix $\mathbf{T}_{n\times h}$ and the loading matrix $\mathbf{P}_{m\times h}$ ($h<r$),

$$\mathbf{X}_{n\times m} = \sum_{i=1,h} \mathbf{t}_i \mathbf{p}_i' + \mathbf{E} = \mathbf{T}_{n\times h} \mathbf{P}'_{m\times h} + \mathbf{E}. \tag{5-15}$$

The analysis has a simple geometrical representation as a linear projection of the m-dimensional **X** space on the h-dimensional subspace **T** using the projector matrix **P**. The loading vectors $\{p_j\}$ are the direction vectors of angle cosines. The h loading vectors, p_j, with larger eigenvalues are principal component directions, and span an h-dimensional orthogonal space. Score vectors $\{t_i\}$ are the projections of sample points on the principal component directions in the h-dimensional space. Actually, the loading vectors $\{p_j\}$ are the normalized eigenvectors of matrix **X'X**

$$(\mathbf{X'X})_{m\times m}=\mathbf{P}_{m\times h}\mathbf{L}_h\mathbf{P'}_{m\times h}\,, \tag{5-16}$$

where $\mathbf{L}_h$ is a diagonal matrix consisting of eigenvalues l_1, l_2, ...,l_h. Loading vectors $\{p_j\}$ represent the non-correlated linear combinations of the variables. Score vectors $\{t_i\}$ are the eigenvectors of matrix **XX'**, however, they are not normalized,

$$(\mathbf{XX'})_{n\times n}=\mathbf{T}_{n\times h}\mathbf{I}_h\mathbf{T'}_{n\times h}\,. \tag{5-17}$$

The eigenvalues of **XX'** are the inner products of score vectors $l_i=\mathbf{t}_i'\mathbf{t}_i$. The score vectors, t_l, with larger eigenvalues are principal components of the data matrix in the *h*-dimensional orthogonal space. The eigenvalues are the variances of principal components. In practice one hopes to use the principal components with higher variances in order to reduce the dimension of the sample space. The principal components with smaller eigenvalues are thought of as the noise.

Loading vectors and the score vectors of pyrazole and its derivatives are listed in Table 5-4 and Table 5-5 respectively, obtained by nonlinear iterative partial least squares (NIPALS). In Table 5-4 the loading vectors are the linear combination of the original indices. The first five loading vectors have larger eigenvalues, and the first three eigenvalues are much greater than the other two. Therefore the number of dimensions of the orthogonal data space is 3 to 5 based on their eigenvalues. In the first principal component direction (loading vector), the coefficients of L_M (0.7941) and $l_s^{(8)}$ (0.4781) are the largest two, followed by H_M (0.-0.2477) and $h_s^{(8)}$ (-0.2140). In the second loading vectors, $l_s^{(8)}$ (-0.7408), $l_a^{(5)}$ (0.3922), and L_M (0.3059) have the largest portions. Indices $h_s^{(8)}$ (-0.6661), H_M (-0.5668), L_M (-0.2725), and $l_s^{(7)}$ (-0.2336) have the largest portions in the third loading vectors. Table 5-5 shows the score vectors in the orthogonal sample space. The first five score vectors have non zero eigenvalues, and the eigenvalues of the first three score vectors are much greater. Figure 5-3 shows the quotients of indices in the first 8 loading vectors, which are the squares of variables. The sum of the squares of all indices is 1 in each loading vector because the loading vectors are normalized to1.

Table 5-4. Loading vectors (principal component directions) of pyrazole derivatives calculated by NLPALS

indices	vector 1	vector 2	vector 3	vector 4	Vector 5	vector 6	vector 7	vector 8	vector 9
$l_a^{(1)}$	0.0527	0.1683	0.0537	0.1686	0.0429	0.5325	0.0412	0.2149	0.2627
$l_a^{(2)}$	-0.0110	0.0208	-0.1625	-0.3104	-0.0143	0.3509	0.1335	0.0963	0.6035
$l_a^{(3)}$	-0.0150	-0.0673	-0.0557	0.0146	0.0126	0.2710	0.0173	-0.5636	-0.2911
$l_a^{(4)}$	-0.0031	0.0495	-0.0208	-0.0565	0.0024	0.0028	0.0253	0.3921	-0.5232
$l_a^{(5)}$	0.1154	0.3922	0.0931	0.2869	-0.5798	-0.3950	-0.2054	-0.0450	0.2730
$l_a^{(6)}$	0.0840	0.2556	0.1005	0.2758	0.1183	0.0983	-0.1476	0.3397	-0.0872
$l_s^{(7)}$	0.0724	0.2513	-0.2336	-0.5854	-0.0965	-0.3339	0.4243	0.0732	-0.0263
$h_s^{(7)}$	-0.0556	-0.1646	0.1920	0.4528	0.0233	-0.1405	0.7387	-0.1861	0.1257
$l_s^{(8)}$	0.4781	-0.7408	-0.1329	0.0266	-0.1370	-0.1539	-0.1381	0.2023	0.1324
$h_s^{(8)}$	-0.2140	0.0131	-0.6661	0.1538	-0.0361	-0.0203	-0.2557	-0.3139	0.0663
$l_s^{(9)}$	0.0421	0.1083	-0.0032	0.0653	0.7782	-0.4217	-0.1577	-0.0286	0.2573
L_M	0.7941	0.3059	-0.2725	0.1290	0.1101	0.1396	0.1549	-0.1546	-0.1315
H_M	-0.2477	-0.0241	-0.5668	0.3470	-0.0025	-0.0391	0.2378	0.3893	-0.0647
Eigenvalues	0.7551	0.3679	0.1492	0.0526	0.0080	0.0020	0.0014	0.0003	0.0001

Table 5-5. Score vectors (principal components) of pyrazole derivatives calculated by NIPALS

molecule	vector 1	vector 2	vector 3	vector 4	vector 5	vector 6	vector 7	vector 8	vector 9
1.	-0.2654	0.0327	0.0426	-0.0157	-0.0180	-0.0011	-0.0109	-0.0020	-0.0044
2.	-0.1920	-0.0129	0.0227	-0.0171	-0.0192	-0.0010	-0.0094	-0.0094	0.0019
3.	-0.0915	-0.0586	-0.0158	-0.0085	-0.0017	0.0038	-0.0041	-0.0034	-0.0004
4.	-0.0476	-0.0751	-0.0327	-0.0066	-0.0023	0.0095	0.0002	-0.0016	-0.0002
5.	-0.0842	-0.0676	-0.0184	-0.0104	-0.0021	0.0068	-0.0005	-0.0034	-0.0017
6.	0.0534	-0.1062	-0.0637	0.0102	-0.0039	0.0076	-0.0001	0.0033	0.0003
7.	0.0088	-0.0953	-0.0457	0.0034	-0.0048	0.0062	-0.0003	0.0003	0.0000
8.	0.0395	-0.0982	-0.0554	0.0107	0.0002	-0.0061	-0.0043	0.0026	0.0023
9.	0.0954	-0.1235	-0.0770	0.0118	-0.0062	0.0069	0.0002	0.0033	0.0015
10.	0.1500	-0.1460	-0.0939	0.0167	-0.0071	0.0047	0.0005	0.0041	0.0012
11.	-0.0847	0.0197	0.0062	0.0411	0.0002	-0.0277	0.0069	-0.0035	0.0052
12.	0.1067	0.1459	0.2732	-0.0167	-0.0087	0.0183	-0.0002	0.0028	0.0040
13.	0.5624	-0.1840	0.1369	-0.0303	0.0100	-0.0148	0.0022	-0.0030	-0.0037
14.	-0.2702	0.0259	0.0284	-0.0327	-0.0201	-0.0027	0.0302	0.0025	-0.0023
15.	0.2631	0.3838	-0.1495	-0.1249	-0.0007	-0.0025	-0.0013	-0.0002	0.0003
16.	-0.1319	0.1254	0.0511	0.0459	-0.0115	-0.0168	-0.0138	0.0101	-0.0021
17.	0.1599	0.2582	-0.0363	0.1664	0.0174	0.0080	0.0053	-0.0041	-0.0020
18.	-0.2717	-0.0241	0.0275	-0.0432	0.0783	0.0008	-0.0006	0.0016	0.0001
Eigenvalues	0.7551	0.3679	0.1492	0.0526	0.0080	0.0020	0.0014	0.0003	0.0001

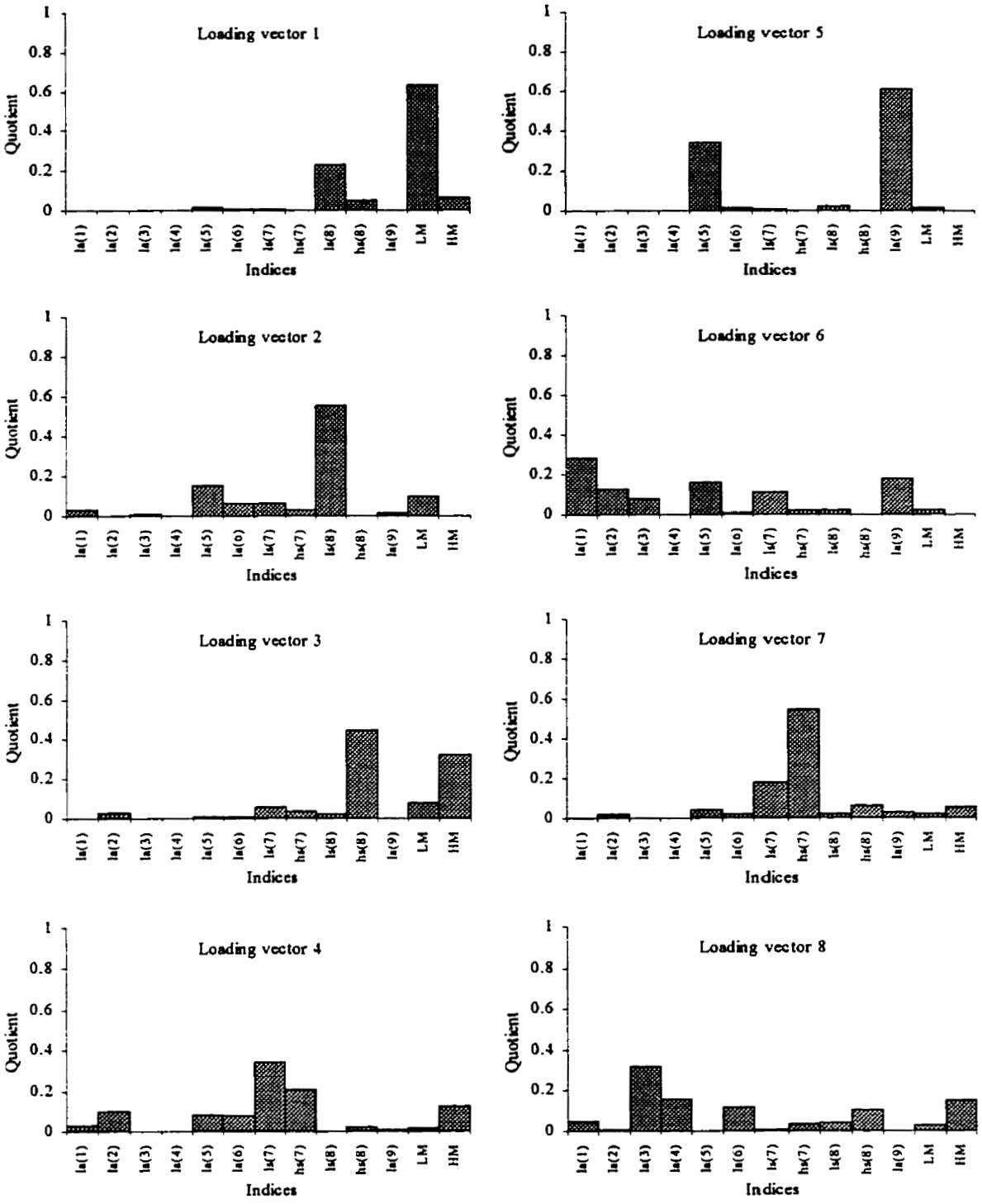


Figure 5-3. Quotients of indices in the first 8 loading vectors. The loading vectors are normalized to 1, therefore the sum of squares of indices is 1 in each loading vector.

Based on principal component analysis, one can estimate how much information is involved in every principal component and how much information remains in the residual matrix. For this purpose, the residual matrices $\mathbf{E}_h$ are calculated after each iteration,

$$\mathbf{E}_h = \mathbf{E}_{h-1} - \mathbf{t}_h\mathbf{p}_h' \quad \text{and} \quad \mathbf{X} = \mathbf{E}_0, \tag{5-24}$$

where the subscript "*h*" is the number of iterations. After every iteration, the sum of squares is calculated for every variable in matrix $\mathbf{E}_h$, and the results are shown in Table 5-6. Also, the sum of squares is calculated for every sample, and the results are shown in Table 5-7. In Tables 5-6 and 5-7, the values are the ratios of the sums of squares in matrices, $\mathbf{E}_h$, to the corresponding sums of squares in matrix $\mathbf{E}_0$. The values in Tables 5-6 and 5-7 tell us how much information remains in the residual matrix $\mathbf{E}_h$ after *h* iterations, or after *h* principal components are removed. Table 5-6 shows that for L_M, more that 90% of the information is contained in the first principal component. However, for $l_a^{(1)}$, $l_a^{(3)}$, $l_a^{(4)}$, $l_a^{(5)}$, and $l_a^{(6)}$, most information is in the second principal component. For $l_a^{(9)}$, more than 40% of the information is in the fifth principal component. After 4 principal components are removed, for most variables the remaining information is less that 5%, however, for $l_a^{(3)}$, $l_a^{(4)}$ and $l_a^{(9)}$, there is still 10.7%, 7.89% and 47.1% of the information, respectively. Table 5-7 shows that for molecules (1) pyrazole, (2) 4-methyl pyrazole, and (14) 3-methyl pyrazole, more than 90% of the information is in the first principal component. On the other hand, for molecule (12) 4-carboxy-pyrazole, most of the information (69%) is in the third principal component. After 4 principal components are removed, the information content is almost exhausted for most molecules, however, there is still 8.5% and 7.4% remaining for the molecules 4-phenyl pyrazole and indazole, respectively. This might mean that indazole and 4-phenyl pyrazole have some special characteristics in this small molecular system. Figure 5-4 shows the sta*.Utistics of variables and samples graphically. It gives an intuitive understanding of the principal component analysis.

Table 5-6. Statistics for the indices: sums of squares for indices in residual matrices, by PCA

The sums are scaled to 1 for each index.

indices	$l_a^{(1)}$	$l_a^{(2)}$	$l_a^{(3)}$	$l_a^{(4)}$	$l_a^{(5)}$	$l_a^{(6)}$	$l_s^{(7)}$	$h_s^{(7)}$	$l_s^{(8)}$	$h_s^{(8)}$	$l_a^{(9)}$	L_M	H_M
Iteration	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Iteration	0.8612	0.9905	0.9340	0.9943	0.8665	0.8481	0.9267	0.9206	0.5427	0.6617	0.8797	0.0891	0.5410
Iteration	0.1702	0.9738	0.2898	0.2667	0.1155	0.1634	0.4960	0.5819	0.0077	0.6611	0.4916	0.0232	0.5389
Iteration	0.1417	0.5632	0.1112	0.2146	0.0983	0.1205	0.3450	0.3950	0.0007	0.0136	0.4914	0.0020	0.0641
Iteration	0.0425	0.0347	0.1069	0.0789	0.0408	0.0064	0.0105	0.0284	0.0006	0.0014	0.4713	0.0003	0.0014
Iteration	0.0415	0.0345	0.1064	0.0789	0.0052	0.0032	0.0091	0.0282	0.0002	0.0013	0.0370	0.0002	0.0014
Iteration	0.0031	0.0083	0.0484	0.0789	0.0010	0.0026	0.0049	0.0269	0.0001	0.0013	0.0043	0.0001	0.0013
Iteration	0.0029	0.0057	0.0483	0.0782	0.0002	0.0017	0.0002	0.0005	0.0000	0.0004	0.0011	0.0000	0.0005
Iteration	0.0019	0.0054	0.0089	0.0384	0.0002	0.0007	0.0001	0.0001	0.0000	0.0001	0.0011	0.0000	0.0000

Table 5-7. Statistics for samples: sums of squares for samples in residual matrices, by PCA. The sums are scaled to 1 for each molecule. The number of molecules can be found in Fig. 5-1.

molecule	Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6	Iteration 7	Iteration 8
1.	1.0000	0.0488	0.0344	0.0100	0.0066	0.0022	0.0022	0.0006	0.0006
2.	1.0000	0.0398	0.0355	0.0220	0.0144	0.0048	0.0048	0.0025	0.0001
3.	1.0000	0.3133	0.0312	0.0107	0.0048	0.0046	0.0034	0.0020	0.0011
4.	1.0000	0.7512	0.1335	0.0166	0.0118	0.0113	0.0013	0.0013	0.0010
5.	1.0000	0.4177	0.0426	0.0147	0.0058	0.0055	0.0017	0.0017	0.0007
6.	1.0000	0.8448	0.2312	0.0106	0.0049	0.0041	0.0009	0.0009	0.0003
7.	1.0000	0.9932	0.1931	0.0093	0.0083	0.0063	0.0029	0.0029	0.0029
8.	1.0000	0.8919	0.2253	0.0132	0.0053	0.0053	0.0027	0.0015	0.0010
9.	1.0000	0.7020	0.2023	0.0080	0.0034	0.0022	0.0006	0.0006	0.0003
10.	1.0000	0.5755	0.1735	0.0071	0.0019	0.0009	0.0005	0.0005	0.0002
11.	1.0000	0.2930	0.2549	0.2511	0.0848	0.0848	0.0090	0.0043	0.0031
12.	1.0000	0.8946	0.6976	0.0066	0.0040	0.0033	0.0002	0.0002	0.0001
13.	1.0000	0.1455	0.0540	0.0034	0.0009	0.0007	0.0001	0.0001	0.0000
14.	1.0000	0.0504	0.0417	0.0312	0.0173	0.0121	0.0120	0.0002	0.0001
15.	1.0000	0.7280	0.1492	0.0613	0.0000	0.0000	0.0000	0.0000	0.0000
16.	1.0000	0.5487	0.1408	0.0732	0.0185	0.0151	0.0078	0.0029	0.0002
17.	1.0000	0.7897	0.2419	0.2310	0.0034	0.0009	0.0004	0.0002	0.0000
18.	1.0000	0.1123	0.1054	0.0962	0.0738	0.0000	0.0000	0.0000	0.0000

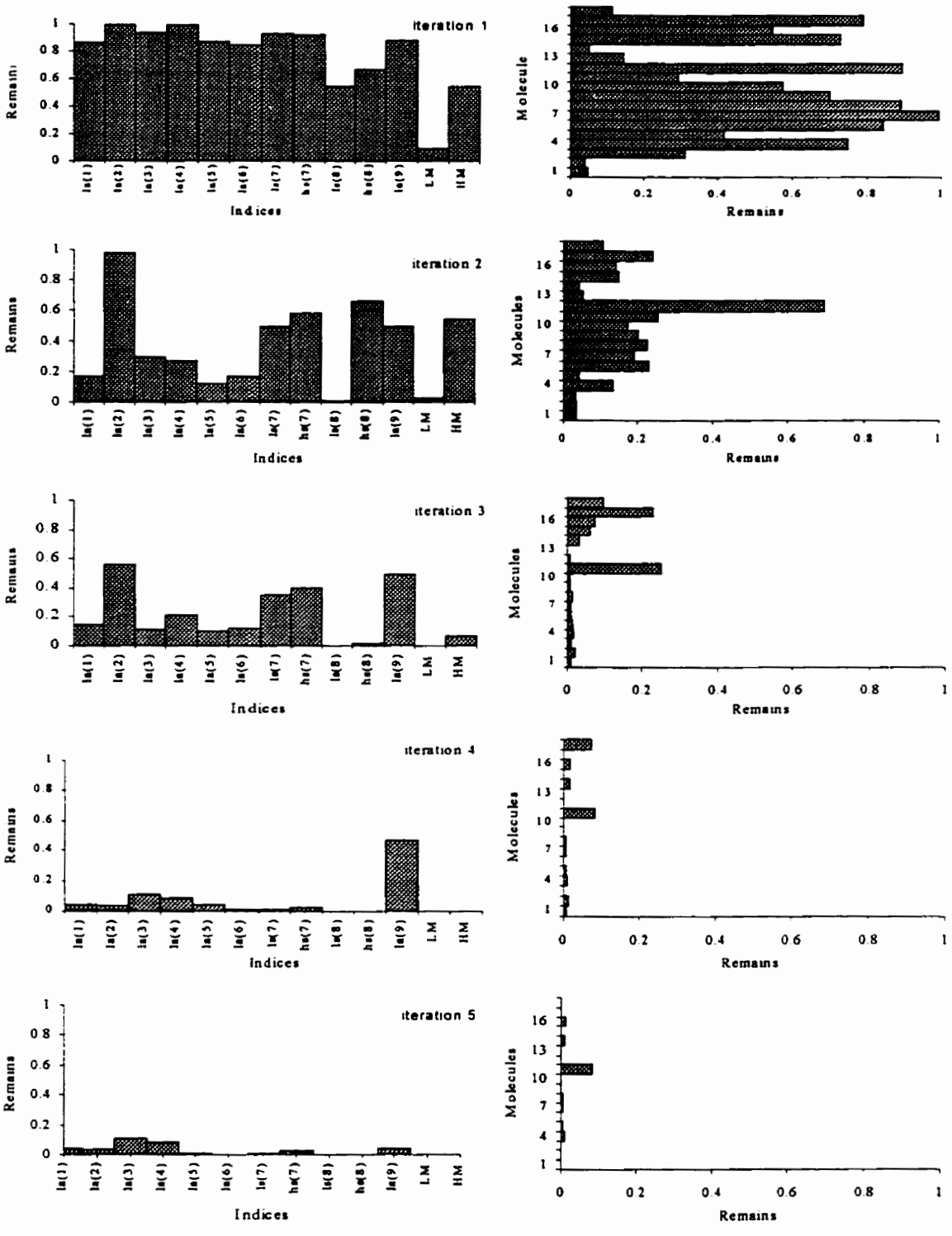


Figure 5-4. Sums of squares in residual matrices $\mathbf{E}_h$ for variables and samples. The sums of squares are normalized to 1 for both variable vectors and sample vectors in the original matrix $\mathbf{E}_0=\mathbf{X}$. The values of other matrices $\mathbf{E}_h$ are the ratios of square sums to the corresponding square sums of $\mathbf{E}_0$.

Figure 5-5 (a) shows the plot of eigenvalues of loading vectors to the number of iterations. Figure 5-5 (b) gives the plot of norms of $\mathbf{E}_h$ *vs* the number of iterations. The values in Fig. 5-5 (b) are the ratios of norms of $\mathbf{E}_h$ to the norm of $\mathbf{E}_0$. Figure 5-5 helps one find the number of principal components. The evaluation of the number of principal components is analogous to the concept of detection limits between signal and noise. Based on Figure 5-5, there are 3 to 5 principal components for this small molecular system. If more principal components are used, noise and interferences may be involved, however, if fewer principal components are used, useful information may be lost.

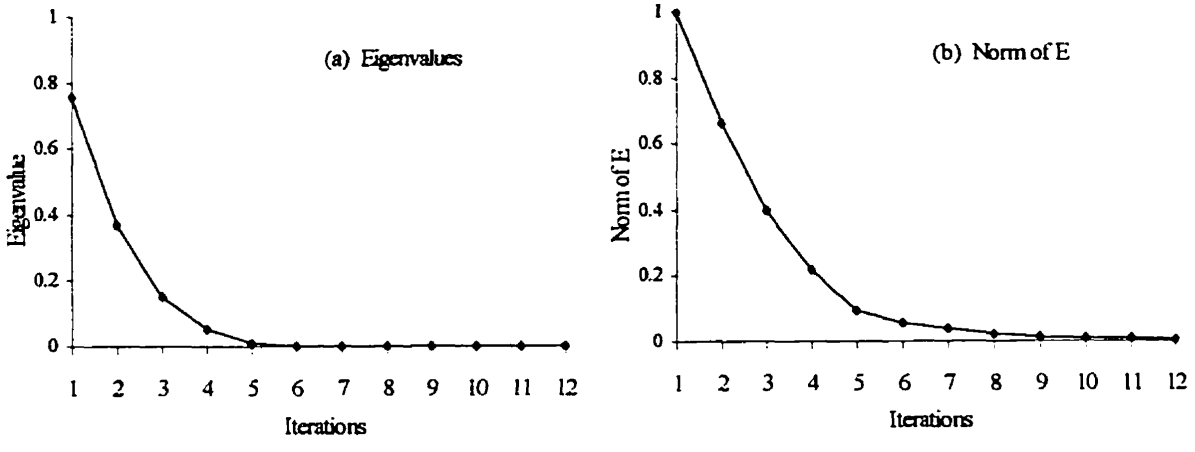


Figure 5-5. (a) Eigenvalues of loading vectors (principal components) *vs* number of iterations. (b) Norms of residual matrices $\|\mathbf{E}_h\|$ *vs* number of iterations.

5.4 Conclusions and Discussions

The small molecular system of pyrazole and its derivatives is a good example for the examination of HMLP and the strategy developed in this study. In §5.1, I presented two questions in computer-aided drug design and QSAR studies. In this section, I examine the results of using HMLP and present a more detailed discussion of answers to the above two questions.

5.4.1 Three Types of Variables in HMLP

For the analysis of small molecular systems using heuristic lipophilicity indices, it is convenient to classify variables into three types: *Endogenous variables, exogenous variables, and general variables*. Endogenous variables are the lipophilicity indices of atoms in the frame of a small molecular system, which are fixed for every molecule and are not substituted in a small molecular system. On the other hand, exogenous variables are the lipophilic and hydrophilic indices of substituents, which are different for every molecule in the small molecular system. Usually for exogenous variables, the lipophilic and hydrophilic indices of substituents, change with the structure of the substituents. General variables are the molecular lipophilic index, L_M, and the hydrophilic index, H_M, which are the general descriptors of molecular lipophilicity and hydrophilicity. For the small molecular system of pyrazole and its derivatives, endogenous variables are indices $l_a^{(1)}$, $l_a^{(2)}$, $l_a^{(3)}$, $l_a^{(4)}$, $l_a^{(5)}$, $l_a^{(6)}$, and $l_a^{(9)}$, which are the indices of atoms on the pyrazole ring and the atoms on the non-substituted side positions. Exogenous variables are indices $l_s^{(7)}$, $h_s^{(7)}$, $l_s^{(8)}$ and $h_s^{(8)}$, which are the indices of substituents on positions 3 and 4. General variables are L_M and H_M, which measure lipophilicity and hydrophilicity of the whole molecule. Classification of endogenous and exogenous variables in a small molecular system depends on the analysis, and the strategy by which experiments are carried out. For example, if substituents are added on position 9, there is no endogenous variable $l_a^{(9)}$, and two more exogenous variables $l_s^{(9)}$ and $h_s^{(9)}$ appear.

These three types of variables have different properties and different behavior. The variances of endogenous variables are smaller than those of exogenous variables and general variables, and fall around the values of the parent molecule in the small molecular system, which is pyrazole. Endogenous variables are affected by the properties of substituents. Variances of endogenous variables are thought to be a measure of the influences of substituents. The values of exogenous variables are decided by the structures of the substituents. The variances of exogenous variables may be much larger than those of endogenous variables. Although the variances of endogenous variables are smaller than those of exogenous and general variables, this does not mean that endogenous variables are less important than exogenous variables and general variables.

Actually, the lipophilicity indices of the key atoms (N(1), N(2) and H(6)) in the inhibition of LADH are endogenous variables ($l_a^{(1)}$, $l_a^{(2)}$ and $l_a^{(6)}$), and have small variances. The exogenous variables play active roles because they affect the properties of endogenous variables and general variables. In the multiple linear regression, the best result is achieved using a combination of the three types of variables: endogenous indices $l_a^{(5)}$ and $l_a^{(6)}$, exogenous indices $l_s^{(7)}$, $h_s^{(7)}$, and $h_s^{(8)}$, and general indices L_M and H_M, see eq. (5-12).

5.4.2 Roles of the Three Types of HMLP Indices in Activity Analysis

Not all variables are independent. Certain relationships exist among all three types of variables, and the relationships depend on the chemical nature of the small molecular system. There is a higher correlation among $l_a^{(1)}$, $l_a^{(3)}$, $l_a^{(5)}$ and $l_a^{(6)}$ based on the variance analysis. In the multiple linear regression between the logarithms of inhibition constants logK_I and HMLP indices, the exogenous variables ($l_s^{(7)}$, $h_s^{(7)}$, $l_s^{(8)}$ $h_s^{(8)}$) and general variables (L_M and H_M) are important. The reason for this is that exogenous and general variables possess higher variances and contain key information in the data set. However, the best result is obtained using a combination of all three types of variables.

Principal component analysis simplifies the structure of the data matrix of a small molecular system considerably. The dimensions of the data matrix of pyrazole derivatives is reduced to 3 - 5 from the original 13 by principal component analysis. The loading vectors of principal components are the linear combinations of original variables with the large eigenvalues, which represent the main variances of principal components in orthogonal space. The components with small eigenvalues may be the noise or other interferences. For pyrazole derivatives, the first four principal components contain 90% of the information (see Fig. 5-5 (b)), however, there is still considerable information in the fifth and other components. After 4 principal components are removed, the remaining information for $l_a^{(1)}$, $l_a^{(3)}$, $l_a^{(4)}$, $l_a^{(5)}$ and $l_a^{(9)}$ is 4.25%, 10.7%, 7.89%, 4.08% and 47.1%, respectively. It seems that $l_a^{(9)}$ has little effect in this small molecular system. The sums of squares for variables and samples in residual matrices, $\mathbf{E}_h$, are used to analyze the information involved in every principal component. Statistical criteria, such as

eigenvalues of loading vectors and norms of residual matrices can be used to judge the number of principal components.

The strategy developed in this research is a combination of a grid-based model, a surface model, and atom-based and fragment-based models. A grid is set on the molecular surface, followed by the construction of the HMLP indices for atoms, substituents, and the molecule based on the lipophilicity potential $L(\mathbf{r})$ on the molecular surface. However, I use the points of grid located on the molecular surface instead of on a regular cubic grid. It is common knowledge that the molecular surface is the interface of molecular interactions. The points deep inside atoms and far from the surface are less important in the study of molecular interactions, and may invalidate the correlation calculations. It is also common knowledge that atoms are the natural unit in chemical interactions. In my strategy, atomic lipophilicity indices are the basic unit. In studies of small molecular systems, the atoms in a substituent are a group, as one unit in the structure-activity analysis. This classification is consistent with chemical conventions and very convenient to chemists. All three types of variables, endogenous, exogenous, and general variables, are measured in the same physical unit, and are calculated using the same method. This characteristic of HMLP indices makes it much easier to provide a rational explanation for the calculation results.

5.4.3 Analysis of the Inhibition of LADH by HMLP Indices

The essential processes involved in the inhibition of LADH (liver alcohol dehydrogenase) are (i) the interaction of a ligand with the metal cation Zn(II), (ii) the transfer of a proton from this ligand to an acceptor, mediated by the presence of the metal cation, and (iii) the binding of the deprotonated ligand to the nicotinamide adenine dinucleotide (NAD^+) coenzyme [Fries *et al.* 1979]. It has long been known that LADH catalyzes the first step in alcohol metabolism, and pyrazole and its derivatives are the strongest inhibitors known for the coenzyme NAD^+ [Eklund *et al.* 1982, Horjales *et al.* 1987]. The interactions between LADH and pyrazole derivatives can be illustrated by Fig. 5-6.

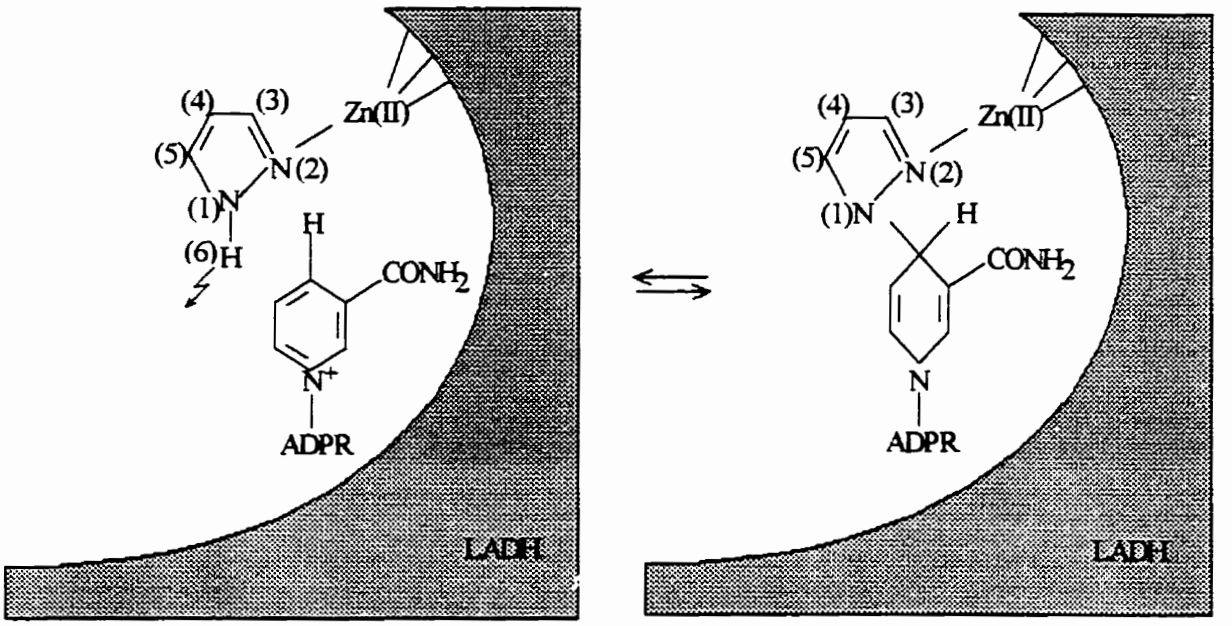


Fig. 5-6. The interaction between LADH, pyrazole and coenzyme NAD^+.

In the model of HMLP, the interaction ability of N(2) with Zn(II) is measured by the negative value of the atomic lipophilicity index $I_a^{(2)}$, which describes the electrostatic interactions. In Table 5-1, $I_a^{(2)}$ has the most negative value (-0.0981) in the pyrazole ring. Both lipophilic and hydrophilic substituents on position 4 have little affect on $I_a^{(2)}$, however, the hydrophilic substituent (COOH, molecule 15) on position 3 decreases the negative value of $I_a^{(2)}$ (-0.0290). On the other hand, lipophilic and hydrophilic substituents on positions 3 and 4 have quite different effects on $I_a^{(6)}$ of proton H(6) bonding with N(1). Index $I_a^{(6)}$ in pyrazole is -0.0302. All lipophilic substituents of the alkyl group on position 4 (molecules 2-10) increase the negative values of $I_a^{(6)}$. The phenyl group (C_6H_5, molecule 11) on position 4 decreases the negative value of $I_a^{(6)}$ (-0.0096), and the hydrophilic substituent (COOH, molecule 12) on position 4 turns the $I_a^{(6)}$ to a positive value (0.0574). When the hydrophilic group COOH is separated by 3 CH_2 (molecule 13), $I_a^{(6)}$ becomes a very small negative value (-0.0085). Hydrophilic substituent (COOH, molecule 15) on position 3 turns the $I_a^{(6)}$ to a positive value (0.0575), too. A higher negative value of $I_a^{(6)}$ helps to release H(6) from N-H, and to create the

bond between pyrazole and NAD [Rozas and Arteca 1991]. Calculation results in this study provide a good explanation for the mechanism of the inhibition of LADH. Another effect caused by the lipophilicity of substituents is that lipophilic substituents on position 4 cause an entropic pressure in an aqueous solution, and help the hydrophilic side of pyrazole (N(2) and H(6)) enter the hydrophilic cavity of LADH, see Fig. 5-6, otherwise, lipophilic substituents on position 3 repel pyrazole derivatives away from the hydrophilic cavity of LADH.

Chapter 6: Further Discussions and Conclusions

Summary

In this chapter, I present some further discussions about three topics: 1) division of molecular surface into atomic pieces as used in HMLP, 2) further tests and improvements of screening function and HMLP, and 3) visualization of HMLP using computer graphics technique. The discussions about these three topics present some ideas for further solutions of the problems found in HMLP. The discussions are outlined below.

1). HMLP is an atom-based technique focusing on the MEP distribution on the molecular surface, and the division of atomic pieces in a molecular surface is necessary. In this part, I review the approaches used in the division of molecular space or surface into atomic pieces and suggest a new division method using fuzzy sets and fuzzy logic.

2). In this part, two compounds, polymethylene oxide (a known hydrophobic compound) $[—CH_2—O—]_n$, and polyethylene oxide (a known hydrophilic compound), $[—CH_2—CH_2—O—]_n$, are used to check my HMLP. Some possible improvements and modifications of the screening function are discussed in this part.

3). Using color pictures to map MLP distribution on molecular surfaces is a useful tool in molecular modeling and drug design. In this part, I will suggest two possible visualization approaches for HMLP: a two-color system and a three-color system. Some possible applications of the three-color system will be discussed.

6.1 Division of Molecular Surface into Atomic pieces

HMLP is an atom-based technique and is a modified MEP. The defining equations (3-6), (2-13) and (3-7) of HMLP, MEP, and atomic surface-MEP descriptor b_i's are rewritten below,

$$\underset{\mathbf{r}\in S_\alpha}{L(\mathbf{r})} = V(\mathbf{r})\sum_{i=\alpha} M_i(\mathbf{r};\mathbf{R}_i, b_i), \tag{3-6}$$

$$V(\mathbf{r}) = \sum_{A=1}^{n} \frac{Z_A}{\|\mathbf{R}_A - \mathbf{r}\|} - \sum_{\mu,\nu} P_{\mu\nu} \int \frac{\phi_\mu^*(\mathbf{r}')\phi_\nu(\mathbf{r}')d\mathbf{r}'}{\|\mathbf{r}' - \mathbf{r}\|}, \tag{2-13}$$

$$b_i = \sum_{k \in S_i} V(\mathbf{r}_k)\Delta s_k, \tag{3-7}$$

where Δs_k is the area element on the surface of atom i and summation is over all exposed surfaces of atom i. It is obvious that the method by which the molecular surface is divided into atomic pieces is important. The division method affects various HMLP indices, too, based on the eq. (3-9),

$$l_\alpha = \sum_{i \in S_\alpha} L(\mathbf{r}_i)\Delta s_i. \tag{3-9}$$

6.1.1 A Brief Review about Division of Molecular Surface

From the viewpoint of quantum mechanics, after atoms combine into a molecule, it is hard to say what part belongs to a certain atom. However, because the atom is an important concept in molecular modeling, deeply rooted in every part of chemistry, one has to discuss the approaches used in the division of molecules into atomic pieces. In this section, I present a mini review of this topic. There are three different approaches used in the division of molecular space and surface: the van der Waals fused-spheres approach, the topological approach, and the LCAO-MO approach.

The van der Waals fused-spheres approach is built on the geometrical point of

view: surrounding each atomic nucleus, there is a space or surface belonging to this atom which is determined by fused-spheres or by atomic van der Waals radii. The fused-sphere van der Waals approach is easily accepted by chemists and widely used in molecular modeling. However, there are several shortcomings to this approach. i) Atomic van der Waals radii have no clear theoretical definitions and are different by different experimental methods. ii) Sometimes, atomic spaces in a molecule are not spheres, particularly for the atoms connected by polar covalent bonds. In this case, the electron density of some atoms may be spread farther along the chemical bonds than other atoms. iii) The border between two atoms may not be formed by the fused spheres, and may be an irregular border. A good computer program for the calculation of van der Waals fused-sphere surfaces is presented by Connolly [1983 a, b, 1985].

The topological approach is established by Bader and coworkers [Bader *et al.* 1979 a, b, Bader 1980]. On the basis of extensive studies of molecular charge distributions, Bader and coworkers have found that molecular charge density, $\rho(\mathbf{r})$, is the universal property that can be used in the definitions of atoms and chemical bonds in a molecule. This universal property may be characterized in terms of the gradient vector field of the charge density, $\nabla\rho(\mathbf{r})$. The properties of this field, and hence the principal characteristics of a charge distribution, are totally determined by the number and character of its critical points, points at which the field vanished. The trajectories of $\nabla\rho(\mathbf{r})$, all of which terminate at particular critical points, define the atoms in a molecule.

The LCAO-MO approach is based on the LCAO-MO approximation, and has been successfully used in the Mulliken population analysis of atomic charges. Actually, LCAO-MO is an approach for the division of atomic contributions to a physical property, not for the division of molecular space and surface. The atomic contributions to a physical property $\langle \mathbf{F} \rangle$ are calculated according to the integrals of atomic orbitals over all space,

$\mathbf{F}_{ij}^{k} = \mathbf{F}_{ij}$ if both $\varphi_i(\mathbf{r})$ and $\varphi_j(\mathbf{r})$ are AO's centered on the *k*th nuclei,

$= 0.5\mathbf{F}_{ij}$ if only one of $\varphi_i(\mathbf{r})$ and $\varphi_j(\mathbf{r})$ is centered on the *k*th nuclei,

$= 0$ otherwise.

This approach has been applied in various quantum chemical programs because coding of the corresponding computer program is easy. When this approach is used in the division of molecular space or surface, there are some questions that should be answered. In the next section, I suggest a division method of molecular surface within the LCAO-MO approach using fuzzy sets and fuzzy logic.

6.1.2 Division of Molecular Surface Using Fuzzy Sets and Logic

In the HMLP approach established in Chapter 3, a molecular surface is described by a set of grid points on the molecular surface S_M,

$$\mathsf{P} = \{p_1, p_2, \ldots\ldots p_n\}, (p_i \in S_M). \tag{6-1}$$

The atomic surface of atom *a* in a molecule is described by a subset of points belonging to atom *a*,

$$\mathsf{P}^{(a)} = \{p^{(a)}_1, p^{(a)}_2, \ldots\ldots p^{(a)}_m\}, (p^{(a)}_i \in S_A). \tag{6-2}$$

The set of grid points of a molecule is the union of all subsets of constituent atoms,

$$\mathsf{P} = \bigcup_a \mathsf{P}^{(a)}. \tag{6-3}$$

The membership of a point in an atomic subset is a fuzzy concept. There are many different ways to determine to which atomic subset a point p_i belongs. So far, in my

thesis research, I use the van der Waals fused-sphere surface, and the atomic pieces are separated according to the borders of fused spheres. This may not be the best method for separating a molecular surface, as discussed in §3.5.2 and §6.1.1. Actually, HMLP is a modified MEP. According to eq. (2-13), for a point p_i on the molecular surface S_M, all atoms have contributions to $V(\mathbf{r}_i)$. Therefore the division of a molecular surface into atomic pieces based on MEP is a question of fuzzy sets and fuzzy logic. For every atomic subset $\mathbf{P}^{(a)}$, a membership set $\mathbf{M}^{(a)}$ can be defined on the set of real numbers $\Re$,

$$\mathbf{M}^{(a)} = \{m_i^{(a)} | m_i^{(a)} = \frac{|V^{(a)}(\mathbf{r}_i)|}{|V(\mathbf{r}_i)|}\}, \quad (0 \le m_i^{(a)} \le 1, \text{ and } i=1,2,\ldots n), \tag{6-4}$$

where $V^{(a)}(\mathbf{r}_i)$ is the contribution from atom a to $V(\mathbf{r}_i)$. Membership sets are normal fuzzy sets in the closed interval [0,1],

$$\mathbf{M}^{(a)}: \Re \rightarrow [0,1]. \tag{6-5}$$

Atomic subset $\mathbf{P}^{(a)}$ can be built based on the α-cut [Klir and Yuan 1995] of atomic membership set $\mathbf{M}^{(a)}$,

$${}^{\alpha}\mathbf{P}^{(a)} = \{p_i | \mathbf{M}^{(a)}(p_i) \ge \alpha\}, \ \alpha \in [0,1], \text{ and } i=1,2,\ldots m \tag{6-6}$$

where α is a cut-off value for the membership by which a point p_i belongs to the atomic subset ${}^{\alpha}\mathbf{P}^{(a)}$, and a reasonable value of α is α=0.5. In this way, the set of grid points P of a molecule is separated into atomic subsets ${}^{\alpha}\mathbf{P}^{(a)}$ using fuzzy sets and fuzzy logic. The method described in this part can be illustrated using Fig. 6-1.

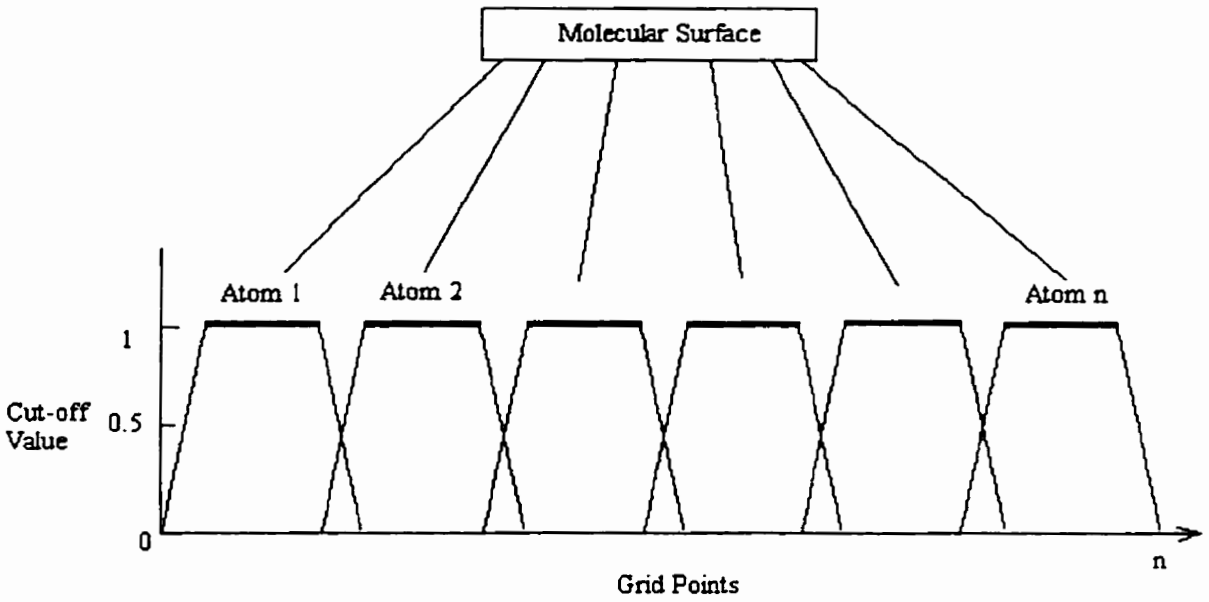


Figure 6-1. Separation of a molecular surface into atomic pieces using fuzzy sets and fuzzy logic.

Under the LCAO-MO approach of quantum mechanical calculations of molecular electrostatic potential, $V(\mathbf{r}_i)$ is calculated in terms of the atomic orbitals, as shown by eq. (2-13). The following rules can be used to assign the atomic contributions to $V(\mathbf{r}_i)$:

$V^{(a)}(\mathbf{r}_i) = V(\mathbf{r}_i)$ if both $\varphi_i(\mathbf{r})$ and $\varphi_j(\mathbf{r})$ are AO's centered on nuclei a,

$= 0.5V(\mathbf{r}_i)$ if only one of $\varphi_i(\mathbf{r})$ and $\varphi_j(\mathbf{r})$ is an AO centered on nuclei a,

$= 0$ otherwise.

The total atomic contribution of atom a is the sum of nuclear contribution and electron contribution,

$$V^{(a)}(\mathbf{r}) = \frac{Z_A}{\|\mathbf{R}_A - \mathbf{r}\|} - \sum_{\mu,\nu \in A} P_{\mu\nu} \int_{\infty} \frac{\phi_\mu^*(\mathbf{r}')\phi_\nu(\mathbf{r}')d\mathbf{r}'}{\|\mathbf{r}' - \mathbf{r}\|} - \frac{1}{2} \sum_{\mu \in A, \nu \in A} P_{\mu\nu} \int_{\infty} \frac{\phi_\mu^*(\mathbf{r}')\phi_\nu(\mathbf{r}')d\mathbf{r}'}{\|\mathbf{r}' - \mathbf{r}\|} \quad (6\text{-}7)$$

In this way one can get the atomic contributions $V^{(a)}(\mathbf{r}_i)$, then find the membership $m^{(a)}{}_i = |V^{(a)}(\mathbf{r}_i)|/|V(\mathbf{r}_i)|$ in set $M^{(a)}$. Finally, atomic subset ${}^{a}P^{(a)}$ can be obtained for atom a

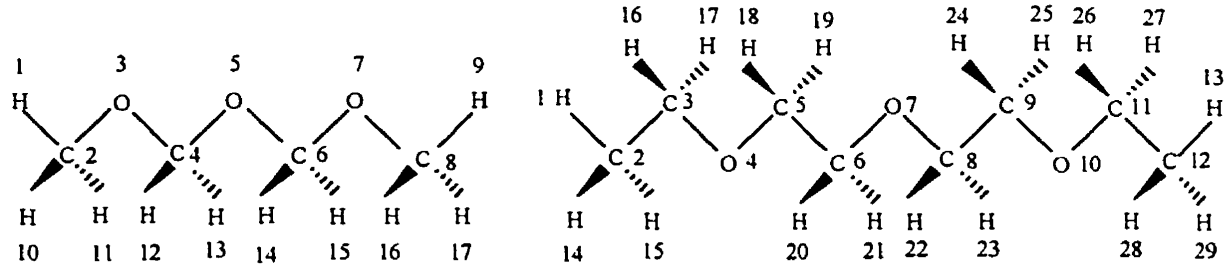
which consists of all points with $m^{(a)}_i \geq \alpha$. This method can be applied to any type of molecular surface. There are no atomic radii required, therefore it can be used in the molecular electron isodensity surface, too.

There is a problem in the above approach: atomic contributions $V^{(a)}(\mathbf{r}_i)$ to MEP $V(\mathbf{r}_i)$ could be positive or negative, and the sum of memberships, $m^{(a)}_i$, of all atoms may not be equal to 1, therefore the cut-off value α is sometimes difficult to select. The following approach can be used to solve this problem. If $V^{(a)}(\mathbf{r}_i)$ has the same sign as $V(\mathbf{r}_i)$, atom a has a positive contribution to $V(\mathbf{r}_i)$. If $V^{(a)}(\mathbf{r}_i)$ has the opposite sign of $V(\mathbf{r}_i)$, atom a has a negative contribution to $V(\mathbf{r}_i)$. Point p_i belongs to the atom having the largest positive contribution to $V(\mathbf{r}_i)$.

6.2 Further Tests and Improvements of HMLP

In this part, I check my HMLP model using two types of compounds, polymethylene oxide, $[—CH_2—O—]_n$, known as a hydrophobic compound, and polyethylene oxide, $[—CH_2—CH_2—O—]_n$, known as a hydrophilic compound. I also want to present an improved screening function and modified HMLP.

6.2.1 Further Tests of HMLP



(a) Polymethylene Oxide (b) Polyethylene Oxide

Figure 6-2. Structures and atomic numbers of polymethylene oxide and polyethylene oxide.

In polymethylene oxide there are more oxygen atoms than in polyethylene oxide

and in polyethylene oxide there are more lipophilic segments CH_2 than in polymethylene oxide. However, the former is lipophilic and the latter is hydrophilic. This is quite strange in the common chemical point of view, and it is difficult to give a reasonable explanation for this phenomenon [Israelachvili 1991]. Two molecules are used in this test: CH_3-O-CH_2-O-CH_2-O-CH_3 and CH_3-CH_2-O-CH_2-CH_2-O-CH_2-CH_2-O-CH_2-CH_3. The first molecule is a small polymethylene oxide and the second molecule is a small polyethylene oxide. The geometries of the two molecules are optimized using Gaussian 92 at the RHF/STO-3G level, and their molecular structures are shown in Fig.6-2. HMLP indices are calculated using RHF with basis set 6-31G*. The power screening function is used, and exponent γ=2.5. Calculation results are listed in Tables 6-1 and 6-2.

Table 6-1. Atomic charges, surface areas, surface-MEP-descriptors, and atomic lipophilicity indices of polymethylene oxide. Calculated using Gaussian 92 at the level RHF/6-31G*, in atomic units.

Number	Atom	q_i	S^+	S^-	S_{total}	b_i^+	b_i^-	b_i	l_a
1	H	0.0786	13.43	0.000	13.43	0.3377	0	0.3377	0.0063
2	C	-0.0658	72.06	28.28	100.3	1.398	-0.6272	0.7706	0.1001
3	O	-0.2328	0	17.77	17.77	0	-1.432	-1.432	-0.0253
4	C	0.1269	36.48	31.38	67.86	1.149	-1.335	-0.1856	0.2119
5	O	-0.2340	0	17.16	17.16	0	-1.513	-1.513	0.0254
6	C	0.1268	36.12	30.28	66.40	1.106	-1.320	-0.2138	0.2063
7	O	-0.2336	0	17.77	17.77	0	-1.434	-1.434	-0.0244
8	C	-0.0658	72.06	29.37	101.4	1.417	-0.6623	0.7546	0.1052
9	H	0.0780	13.13	0	13.13	0.3272	0	0.3272	0.0060
10	H	0.0581	13.58	0	13.58	0.4337	0	0.4337	0.0116
11	H	0.0581	13.73	0	13.73	0.4379	0	0.4379	0.0116
12	H	0.0469	14.49	0	14.49	0.5373	0	0.5373	0.0002
13	H	0.0469	14.49	0	14.49	0.5383	0	0.5383	0.0003
14	H	0.0467	14.04	0	14.04	0.5214	0	0.5214	-0.0001
15	H	0.0467	14.19	0	14.19	0.5259	0	0.5259	-0.0002
16	H	0.0591	13.89	0	13.89	0.4469	0	0.4469	0.0113
17	H	0.0591	14.04	0	14.04	0.4511	0	0.4511	0.0114

Table 6-2. Atomic charges, surface areas, surface-MEP-descriptors, and atomic lipophilicity indices of polyethylene oxide. Calculated using Gaussian 92 at the level RHF/6-31G*, in atomic units.

Number	Atom	q_i	S^+	S^-	S_{total}	b_i^+	b_i^-	b_i	l_a
1	H	0.0634	13.73	0	13.73	0.2975	0	0.2975	0.0034
2	C	-0.1819	56.18	35.21	91.39	0.3945	-0.4016	-0.0071	0.0214
3	C	0.0161	41.41	16.05	57.46	0.6241	-0.3430	0.2811	0.0381
4	O	-0.2518	0	16.85	16.85	0	-1.1434	-1.1435	-0.0282
5	C	0.0085	29.55	26.27	55.82	0.2717	-0.4724	-0.2008	0.0255
6	C	0.0092	29.92	27.36	57.28	0.2641	-0.5058	-0.2417	0.0250
7	O	-0.2526	0	16.85	16.85	0	-1.0650	-1.0650	-0.0083
8	C	0.0093	29.19	26.63	55.82	0.2467	-0.4794	-0.2327	0.0228
9	C	0.0085	30.28	27.00	57.28	0.2896	-0.5011	-0.2114	0.0279
10	O	-0.2520	0	16.85	16.85	0	-1.1450	-1.1450	-0.0286
11	C	0.0162	41.41	14.59	56.00	0.6103	-0.3074	0.3029	0.0345
12	C	-0.1819	56.19	34.84	91.03	0.3899	-0.3954	-0.0055	0.0201
13	H	0.0635	13.13	0	13.13	0.2853	0	0.2853	0.0033
14	H	0.0678	12.83	0.1509	12.98	0.1562	-0.0001	0.1560	0.0009
15	H	0.0678	12.98	0.1509	13.13	0.1583	-0.0001	0.1581	0.0009
16	H	0.0493	13.89	0	13.89	0.3180	0	0.3180	0.0025
17	H	0.0493	14.04	0	14.04	0.3212	0	0.3212	0.0025
18	H	0.0563	13.58	0	13.58	0.2444	0	0.2444	-0.0027
19	H	0.0563	13.73	0	13.73	0.2470	0	0.2470	-0.0028
20	H	0.0579	13.89	0	13.89	0.2358	0	0.2358	-0.0047
21	H	0.0579	14.04	0	14.04	0.2382	0	0.2382	-0.0047
22	H	0.0579	13.58	0	13.58	0.2313	0	0.2313	-0.0046
23	H	0.0579	13.73	0	13.73	0.2338	0	0.2338	-0.0046
24	H	0.0563	13.88	0	13.89	0.2496	0	0.2496	-0.0028
25	H	0.0563	14.04	0	14.04	0.2521	0	0.2521	-0.0029
26	H	0.0493	13.58	0	13.58	0.3108	0	0.3108	0.0026
27	H	0.0493	13.73	0	13.73	0.3137	0	0.3137	0.0026
28	H	0.0679	13.28	0	13.28	0.1602	0	0.1602	0.0009
29	H	0.0679	13.43	0	13.43	0.1622	0	0.1622	0.0009

There are three oxygen atoms in the polymethylene oxide molecule. In Table 6-1, although the first and third oxygen atoms, O(3) and O(7), which are on the two ends of the molecule and are connected with the CH_3 group, are hydrophilic ($f^{(3)}_a$=-0.0253 and $f^{(7)}_a$=-0.0244), the second oxygen (O(5)) atom in the middle is lipophilic ($f^{(5)}_a$=0.0254), and all carbon atoms are lipophilic, too. It is easy to understand that in a polymethylene oxide macromolecule, except for the two oxygens on the two ends of the chain, all oxygens in the middle and all carbons are lipophilic, therefore polymethylene oxide is a lipophilic compound. In Table 6-2, all three oxygen atoms (O(4), O(7), and O(10)) in polyethylene oxide are hydrophilic ($f^{(4)}_a$=-0.0282, $f^{(7)}_a$=-0.0083 and $f^{(10)}_a$=-0.0286, respectively), and all carbon atoms are almost 5 to 10 times less lipophilic than carbon atoms in polymethylene. Therefore it is a hydrophilic compound. These calculated results are consistent with the chemical and physical properties of these two molecules, demonstrating the success of HMLP. Atomic charges q_i and atomic MEP-surface descriptors b_i^+ and b_i^- in Tables 6-1 and 6-2, help one to understand this phenomenon. In the polymethylene oxide molecule, the central O(5) has two neighbors, C(4) and C(6), which are positively charged (q_i=0.127) and have large and almost equal b_i^+ (1.149 and 1.106) and b_i^- (-1.335 and -1.320). The first neighbors of O(5) give almost equal positive and negative influences on the lipophilicity of O(5). The secondary neighbors of O(5) are O(3) and O(7) having large negative b_i (-1.432 and -1.434). The influences from the secondary neighbors make O(5) a lipophilic atom. In the polyethylene oxide molecule, the conditions of the first neighbors of the central O(7) are the same as those in polymethylene oxide, however, the conditions of the second neighbors, two carbon atoms, are totally different from those atoms, two oxygens, in the polymethylene oxide. Therefore the influences from the first and second neighbors of O(7) make it a hydrophilic atom.

6.2.2 Improvements of Screening Function

The molecular lipophilicity potential is a real physical potential describing the distribution of interaction potential energies between a macromolecule and a huge number of water molecules. The most important difference between heuristic and empirical lipophilicity potential is that HMLP has a certain physical meaning and

theoretical background. In this section, I try to present a more theoretical screening function and an improved HMLP.

Molecular electrostatic potential $V(\mathbf{r})$ is the measurement of the interaction ability of a solute molecule with a unit test charge at point $\mathbf{r}$. Suppose a water molecule is a dipole consisting of two opposite point charges (q^{+}_{w} and q^{-}_{w}), then $V(\mathbf{r}_A)$ and $V(\mathbf{r}_B)$ are the measurements of interaction abilities of atom A and atom B in a solute molecule with water molecules at points $\mathbf{r}_A$ and $\mathbf{r}_B$, respectively. As discussed in §4.2.1, the lipophilicity of atom A is affected by all neighboring atoms. Considering the size of a water molecule, the influences of neighboring atoms only be effective at some positions, and a new screening function is suggested as follows,

$$M_i(\mathbf{r};\mathbf{R}_i,b_i) = \frac{r_0}{b_0}\frac{b_i}{\|\mathbf{R}_i-\mathbf{r}\|}\{\cos([\frac{\|\mathbf{R}_i-\mathbf{r}\|}{\lambda_0}-\frac{1}{2}]\pi)+1\}$$

$$= \zeta\frac{b_i}{\|\mathbf{R}_i-\mathbf{r}\|}\{\cos([\frac{\|\mathbf{R}_i-\mathbf{r}\|}{\lambda_0}-\frac{1}{2}]\pi)+1\}, \qquad (6\text{-}8)$$

where λ_0 is the dimension of a water molecule. Comparing this with the power screening function eq. (4-8) in Chapter 4, there are two modifications: (1) exponent γ is assigned the value 1, and (2) a distance-dependent periodic factor is added to the screening function. The graph of the new screening function, eq. (6-8), is shown in Fig. (6-3).

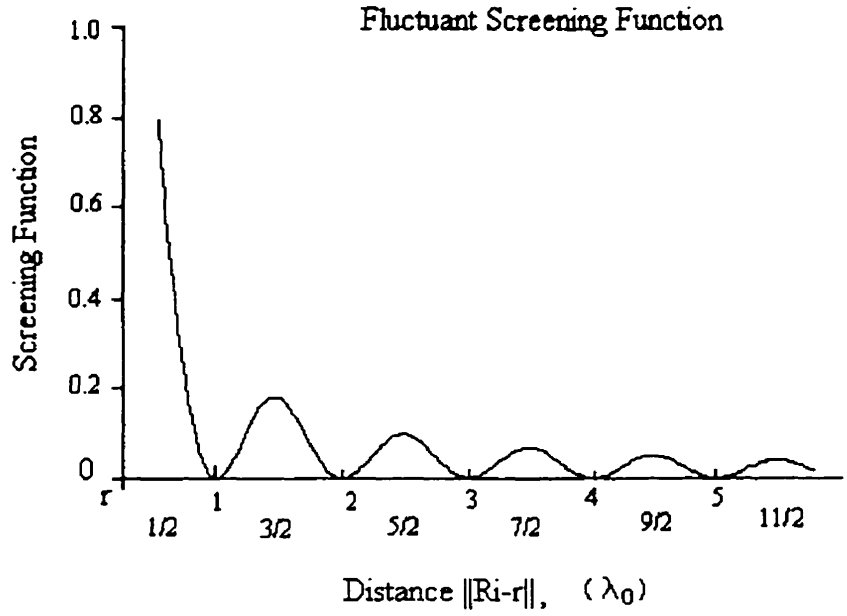


Figure 6-3. Fluctuating decaying screening function. Here, λ_0 is the dimension of a water molecule. The maxima are at integer factors of the half dimension ($\lambda_0/2$) of a water molecule.

In Fig. (6-3) the maxima are at integer factor of the half dimension ($\lambda_0/2$) of a water molecule, the minima are at integer times of λ_0, and the maxima decay with the distance as a function $1/x$. The reason why the exponent γ in the power distance-decaying screening function takes the value 1 is that the physical nature of interactions among water molecules and the interactions between atoms in a solute molecule and water molecules are electrostatic interactions. If the above assumptions are true, there should be a direct connection between HMLP and hydration free energies. The scale factor ζ in eq. (6-8) may be derived from the experimental data of hydration free energies. A fluctuating decaying screening function is expected to be effective for the fluorinated organic compounds.

6.3 Visualization of HMLP on a Molecular Surface

Just like empirical MLP [Kellogg and Abraham 1992], heuristic MLP, possibly in combination with a computer graphic technique to show the distribution of MLP on the

molecular surface in different colors, provides more detailed information and gives the expressive visualization of lipophilicity. This technique is expected to be valuable for the study of complementary lipophilic and electrostatic maps on molecular surfaces in both direct and indirect drug design [Kellogg and Abraham 1992]. In this section, I suggest several methods for the visualization of HMLP.

6.3.1 Two-color System

As shown in §3.2.1, in the definition of HMLP, positive values are used for lipophilicity and negative values are used for hydrophilicity. A two-color system for the molecular lipophilicity map is suggested based on this definition. In the two-color system, the color red is assigned for lipophilicity, while the color green is used to represent hydrophilicity. A molecular lipophilicity map can be drawn on the molecular surface according to the values of $L(\mathbf{r})$ on the surface grid. The color distributions are from deep red to light pink and from deep green to light green. Then a complementary lipophilicity map can be designed based on the original lipophilicity map.

Molecular lipophilicity maps can also be drawn according to atomic lipophilicity indices l_a's. Each atom has one color based on the value of l_a. In this way molecular lipophilicity maps are not limited to molecular surfaces. It can be used in any form of representations of molecular structures, such as three dimensional stick-structures, stick-ball structures, stick structures, and fused-sphere structures.

6.3.2 Three-color System

HMLP is a unified lipophilicity and hydrophilicity potential, and most electrostatic interactions are included in hydrophilicity potential. In the study of complementary electrostatic interaction, one needs to know the positive and negative electrostatic interactions. In the two-color system of HMLP, negative values are used for hydrophilicity (electrostatic interactions). Both negative and positive electrostatic interactions (hydrophilicity) use one color. This is a shortcoming of the two-color system. In the three-color system, neutral numbers are used for lipophilicity, and positive and negative values are used for positive and negative hydrophilicity (electrostatic

interactions), respectively. For hydrophilicity, if the original MEP is negative, HMLP gets a negative value, if the original MEP is positive, HMLP gets a positive value. Three prime colors are used in the three-color system: yellow for neutral values (lipophilicity), red for positive values (positive hydrophilicity), and green for negative values (negative hydrophilicity).

The three-color system contains more information than the two-color system. In the three-color system, hydrogen-bond donors and acceptors can be shown in different color on the HMLP map according to the positive and negative values of HMLP, respectively. In the studies of molecular similarity, dissimilarity, recognition, and complementary interactions of ligand-receptor complex, the three-color system is expected to be valuable. However, new mathematical tools, tertiary algebra, are needed in the three-color system for the study of molecular similarity, dissimilarity, recognition, and complementarity.

Some color pictures of HMLP maps of small molecules in two-color system are shown in the following pages. Generally speaking, these HMLP maps give very reasonable representations of lipophilicity distribution on mo'm(.Ulecular surfaces. Some deep red spots are errors caused by incorrect partitions of atomic surfaces, or incorrect atomic van der Waals radii used in calculations. However, based on my observations, these errors are tolerable, and can be minimized through optimizations of atomic van der Waals radii. These pictures are made by AVS. HMLP data are converted from surface-dots to cubic grid. This is the second source of errors. The third source of errors is from the computer program MS [Connolly 1983 a, b, 1985] for the calculations of molecular surfaces, which may leave some hollows near the borders between atoms.

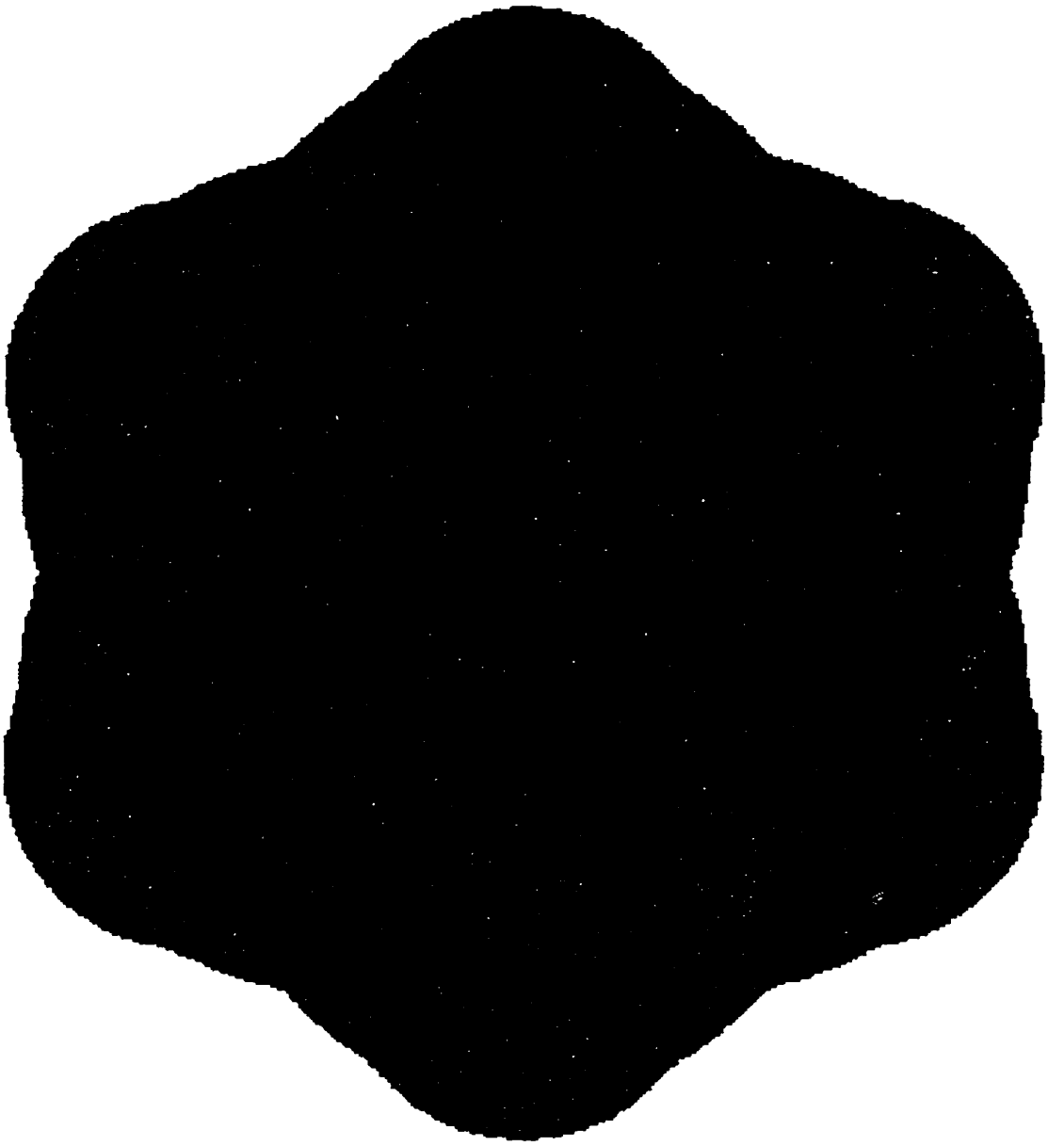


Figure 6-4. Benzene is a molecule almost neutral in lipophilicity. Both molecular lipophilic index, L_M=0.01807, and molecular hydrophilic index, H_M=-0.01627 are very small. The whole molecule is little lipophilic. When a hydrophilic functional group is put on the ring, the ring becomes much more lipophilic. Please see C_6H_5COOH and $C_6H_5NO_2$.

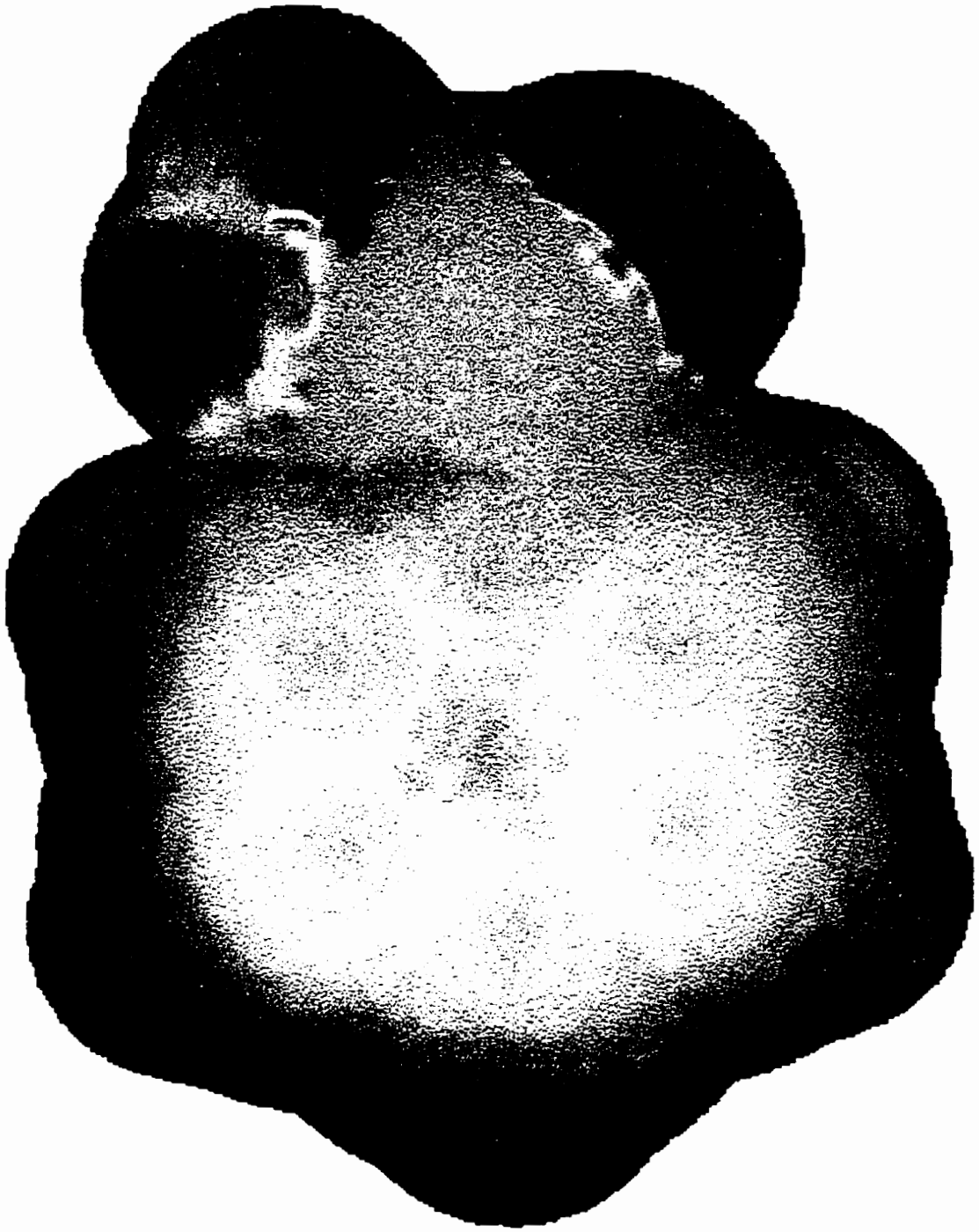


Figure 6-5. HMLP map of C_6H_5COOH in two-color system. The blue and green parts are the hydrophilic areas. The red and yellow parts are the lipophilic areas. H's are more lipophilic than C's in the benzene ring. The two red spots between H and O in hydroxyl group, and between H and carboxyl C in the COOH group are errors that may be caused by the incorrect partitioning of atomic surfaces.

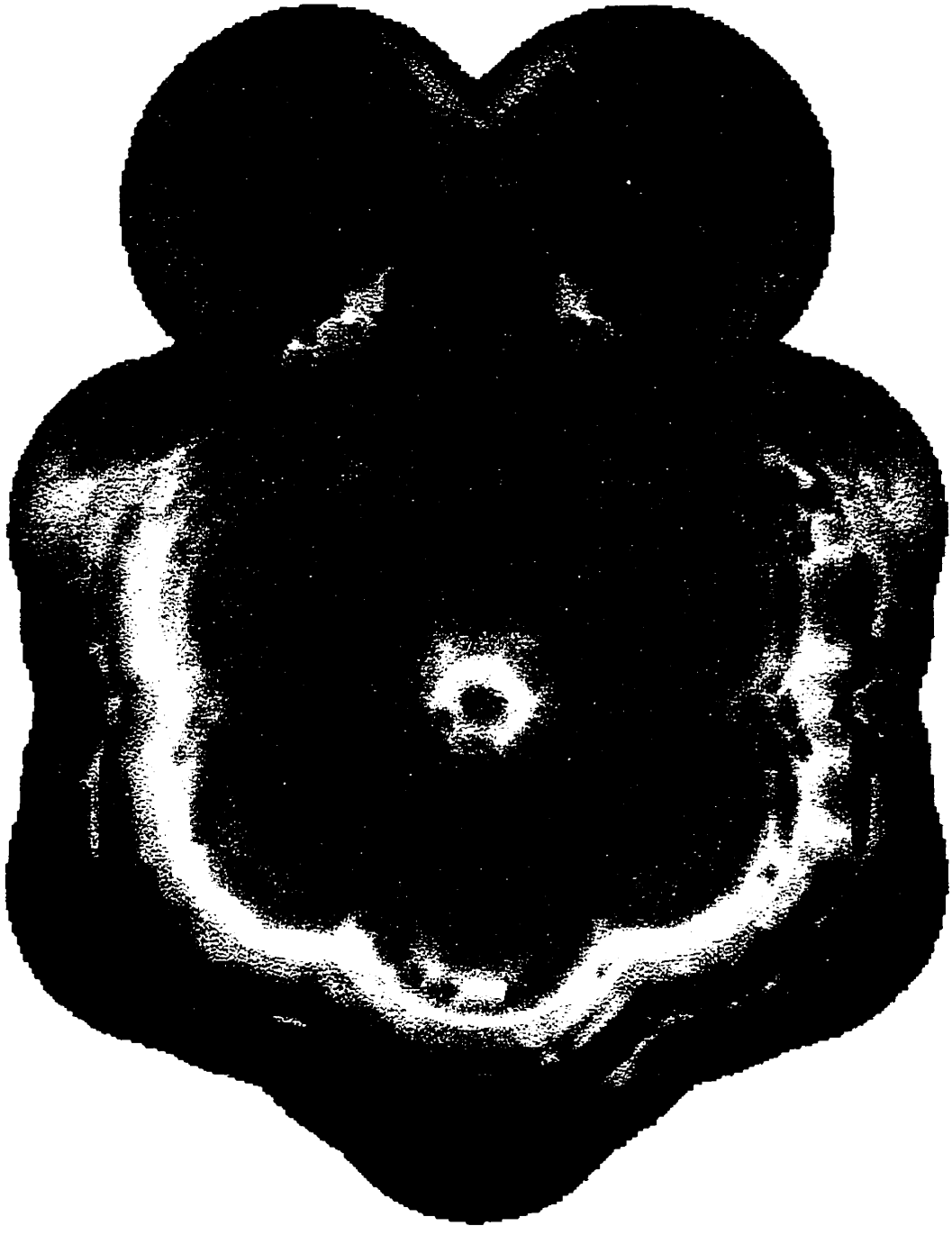


Figure 6-6. HMLP map of $C_6H_5NO_2$ in two-color system. The blue and green parts are the hydrophilic areas. The red and yellow parts are the lipophilic areas. H's are much more lipophilic than C's in the benzene ring. This may be caused by the electron conjugation between C_6H_5 and NO_2.

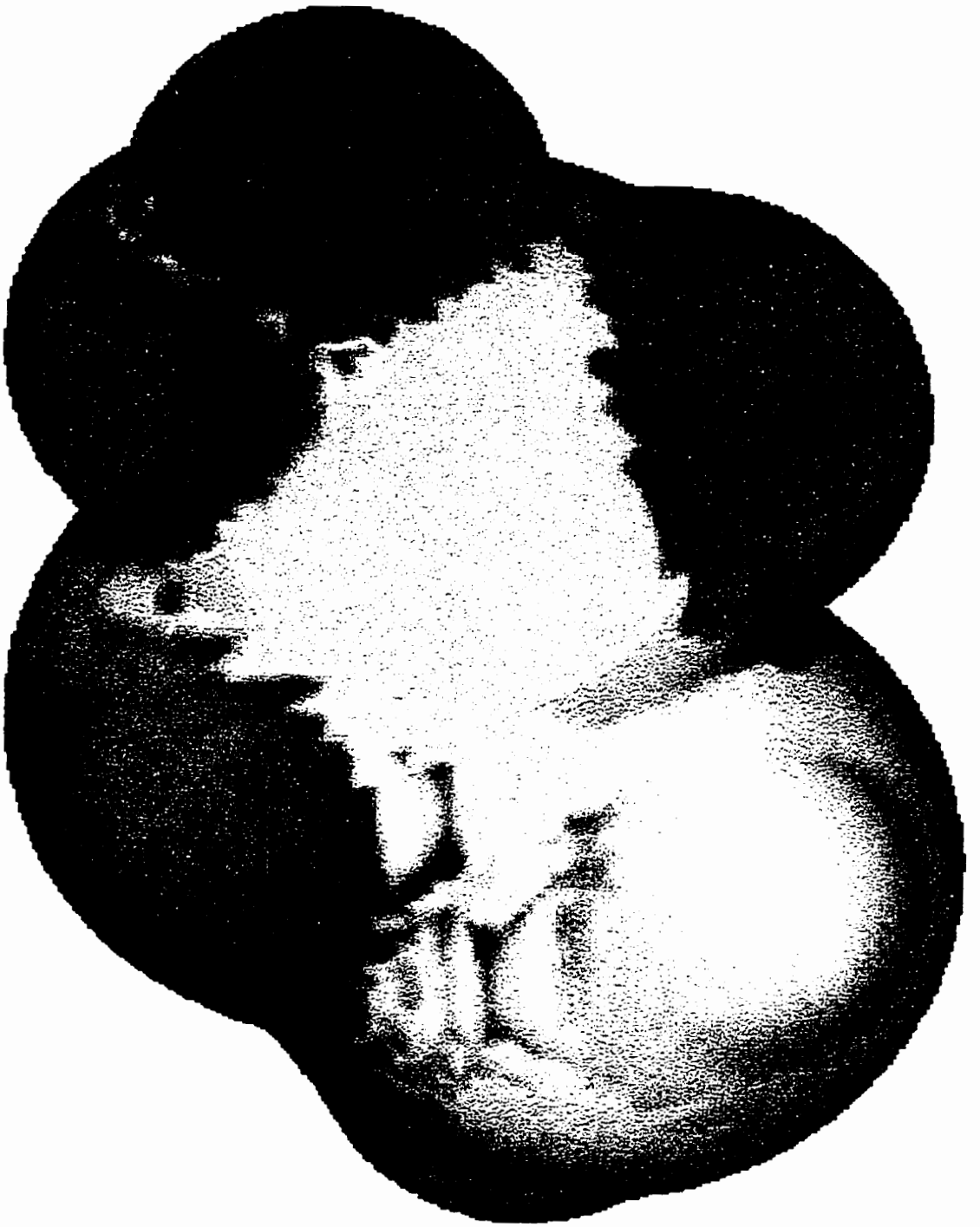


Figure 6-7. HMLP map of C_2H_5COOH in two-color system. The blue and green parts are the hydrophilic areas. The red and yellow parts are the lipophilic areas. CH_2 is more lipophilic than CH_3. The two red spots between H and O in the hydroxyl group, and between H and carboxyl C in the COOH group are errors that may be caused by the incorrect partitioning of atomic surfaces.

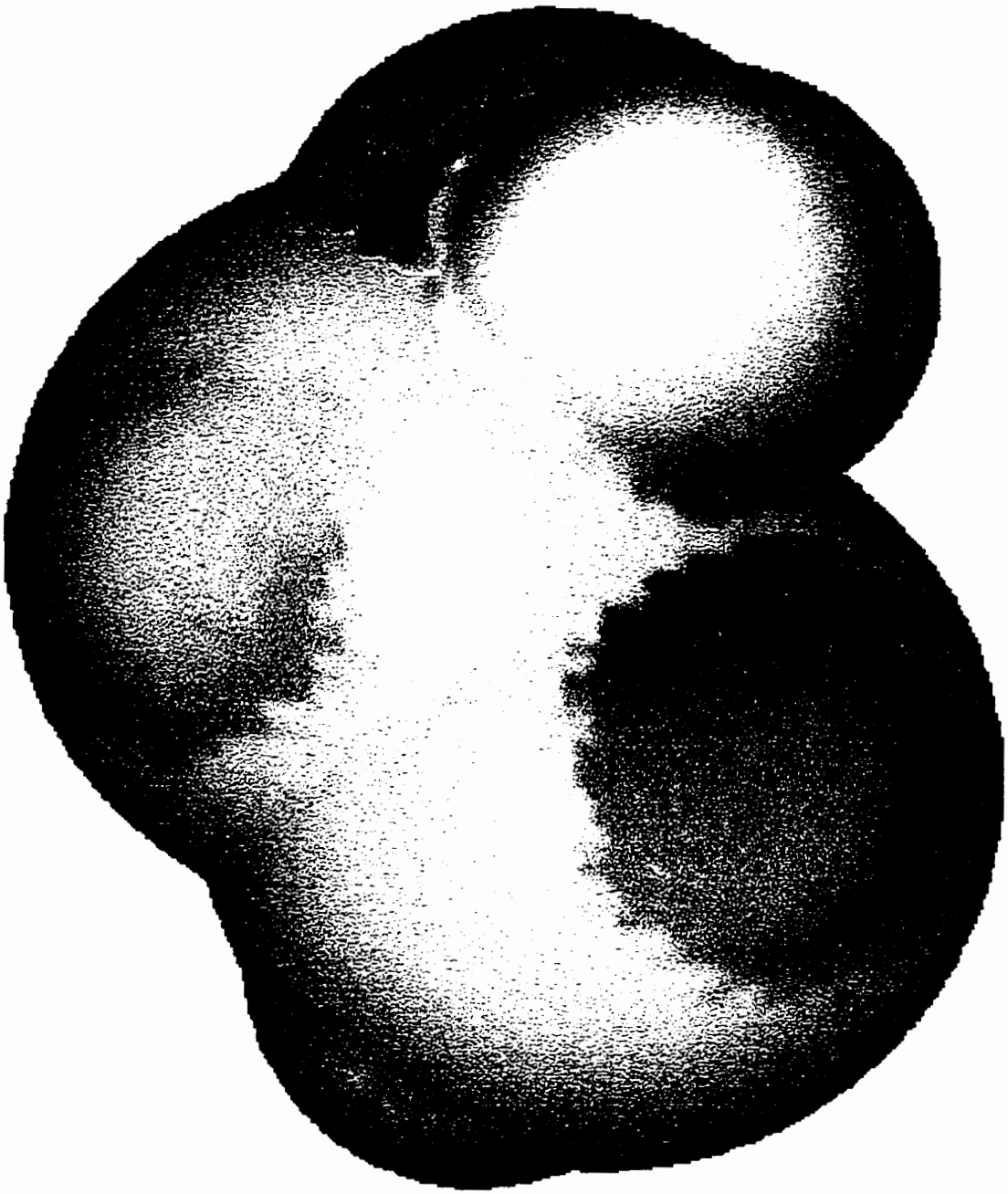
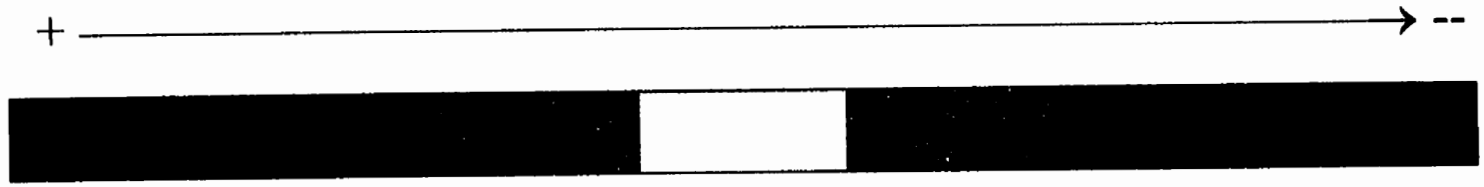


Figure 6-8. HMLP map of $C_2H_5NH_2$ in two-color system. The blue and green parts are hydrophilic areas. The red and yellow parts are lipophilic areas. The nitrogen in NH_2 is much more hydrophilic than two hydrogens. This is consistent with the chemical property that NH_2 is a Lewis base. Some errors can be found between N and the two H's. The reason may be that the van der Waals radius (1.46Å) of N is smaller than it should be (1.55Å). A fused-sphere van der Waals surface was used.

Two-Color System:

(Red) Lipophilicity

Hydrophilicity (Blue)

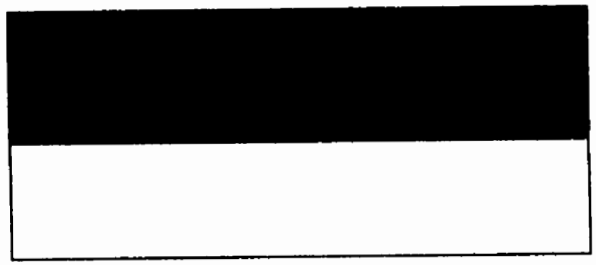


Three-Color System:

Lipophilicity:
neutral ( )

Hydrophilicity:
Positive (Red), Negative (Blue)

+Hydrophilicity



-Hydrophilicity

Lipophilicity

Figure 6-9. Visualization of HMLP on a Molecular Surface

203

References

Abraham, M.H., Doherty, R.M., Kamlet, M.J., and Taft, R.W., *Chem. Br.*, 22 (**1986**), 551.

Abraham, M.H., Grellier, P.L., Abboud, J.-M., Doherty, R.M., and Taft, R.W., *Can. J. Chem.*, 66 (**1988**), 2673.

Abraham, M.H., *Chem. Soc. Rev.*, 22 (**1993**), 73.

Akahane, K., Nagano, Y., and Umeyama, H., *Chem. Pharm. Bull.*, 37 (**1989**), 86.

Alder, B.J. and Wainwright, T.E., *J. Chem. Phys.*, 27 (**1957**), 1208.

Alder, B.J. and Wainwright, T.E., *J. Chem. Phys.*, 31 (**1959**), 459.

Alkorta, I. and Villar, H.O., *Int. J. Quantum. Chem.*, 44 (**1992**), 203.

Alkorta, I., Villar, H.O., and Arteca, G.A., *J. Comput. Chem.*, 14 (**1993**), 530.

Armstrong, D.W., *Sep. Purif. Methods*, 14 (**1985**), 213.

Arteca, G.A., Jammal, V.B., Mezey, P.G., Yadav, J.S., Hermsmeier, M.A., and Gund, T.M., *J. Mol. Graphics*, 6 (**1988**), 45.

Arteca, G.A., Grant, N.D., and Mezey, P.G., *J. Comput. Chem.*, 12 (**1991 a**), 1198.

Arteca, G.A., Hernández-Laguna, A., Rández, J.J., Smeyers, Y.G., and Mezey, P.G., *J. Comput. Chem.*, 12 (**1991 b**), 705.

Atkins, P.W., (**1994**) "*Physical Chemistry*", Freeman, W.H. and Co., New York, Fifth Edition.

Audry, E., Dubost, J.P., Colleter, J.C., and Dallet, P., *Eur. J. Med. Chem.*, 24 (**1989 a**), 71.

Audry, E., Dubost, J.P., Dallet, P., Langlois, M.H., and Colleter, J.C., *Eur. J. Med. Chem.*, 24 (**1989 b**), 155.

Audry, E., Dubost, J.-P., Colleter, J.-C and Dallet, P., *Eur. J. Med. Chem. Chim. Thér.*, 21 (**1986**), 71.

Bachrach, S.M., in: Lipkowitz, K.B. and Boyd, D.B. (Eds.), (**1994**) "*Reviews of Computational Chemistry*", vol.5, VCH, New York.

Bader, R.F.W., Carroll, M.T., Cheeseman, J.R., and Chang, C. J., *J. Am. Chem. Soc.*, 109 (**1987**), 7968.

Bader, R.F.W., *J. Chem. Phys.*, 73 (**1980**), 2871.

Bader, R.F.W., Anderson, S.G., and Duke, A.J., *J. Am. Chem. Soc.*, 101 (**1979**), 1389.

Bader, R.F.W., Nguyen-Dang, T.T., and Tal, Y., *J. Chem. Phys.*, 70 (**1979**), 4316.

Balbes, L.M., Mascarella, S.W. and Boyd, D.B., (**1994**) *in: "Reviews in Computational Chemistry, A perspective of modern methods in computer-aided drug design"*, Lipkowitz, K.B. and Boyd, D.B. (Eds), Vol. 5, VCH, Weinheim, Germany, P. 337.

Baricic, P. and Mackov, M., *J. Mol. Graph.* **12**, 49. *J. Comput.-Aided Mol. Design*, 7 (**1994**), 503.

Bartell, L. S., *J. Chem. Educ.*, 45 (**1968**), 754.

Ben-Naim, A., (**1987**) "*Solvation Thermodynamics*", Plenum Press, New York and London.

Ben-Naim, A., (**1980**) "*Hydrophobic Interactions*", Plenum Press, New York, Chap. 1 and App. 1.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., Di Nola, A., and Haak, J.R., *J. Chem. Phys.*, 81 (**1984**), 3684.

Binkley, J.S., Whiteside, R.A., Krishnan, R., Seeger, R., DeFrees, D.J., Schlegel, H.B., Topiol, S., Kahn, L.R., and Pople, J.A., (**1980**) "*Gaussian 80*," Carnegie-Mellon University, Pittburgh PA.

Binkley, J. S., Whiteside, R.A., Krishnan, R., Seeger, R., DeFrees, D. J., Schlegel H. B., Topiol, S., Kahn, L. R., and Pople, J. A., (**1981**) *QCPE Bull*, 13.

Bodor, N., Gabanyi, Z., and Wong, C.-K., *J. Am. Chem. Soc.*, 111 (**1989**), 3783.

Bodor, N. and Huang, M.-J.,*J. Pharmacol. Sci.*, 81 (**1992**), 272.

Bonaccorsi, R., Palla, P., Tomasi, J., *J. Am. Chem. Soc.*, 106 (**1984**), 1945.

Bonaccorsi, R., Floris, F., Palla, P., and Tomasi, J., *Thermochim. Acta*, 162 (**1990**), 213.

Bondi, A., (**1968**) "*Physical Properties of Molecular Crystals, Liquids, and Gasses*", Wiley, New York.

Bone, R.G.A. and Villar, H.O., *J. Mol. Graph.*, 13 (**1995**), 201.

Borman, S., *Chem. Eng. News*, August 14, **1995**, 29.

Borman, S., *Chem. Eng. News*, February 12, **1996**, 28.

Bowen, J.P., and Allinger, N.L., (**1991**) in: "*Reviews of Computational Chemistry*", Lipkowitz, K.B. and Boyd, D.B. (Eds.), vol.2, VCH, New York.

Bowerman, B.L. and O'Connell, R.T., (**1990**) "*Linear Statistical Models, an Applied Approach*", PWS-KENT Publishing Company, Boston.

Brandstrom, A., *Acta. Chim. Scand.*, 17 (**1963**), 1218.

Braumann, T., *J. Chromatogr.* , 373 (**1986**), 191.

Brinck, T., Murray, J.S., and Politzer, P., *J. Org. Chem.*, 56 (**1991**), 5012.

Brinck, T., Murray, J.S., and Politzer, P., *Mol. Phys.*, 76 (**1992 a**), 609.

Brinck, T., Murray, J.S., and Politzer, P., *Int. J. Quantum Chem., Quantum Biol. Symp.*, 19 (**1992 b**), 57.

Brinck, T., Murray, J.S., and Politzer, P., *J. Org. Chem.*, 58 (**1993**), 7070.

Brooks, B.K., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M., *J. Comput. Chem.*, 4 (**1983**), 187.

Bunin, B.A. and Ellman, J.A., *J. Am. Chem. Soc.*, 114 (**1992**), 10997.

Burdett, J.K., (**1980**) "*Molecular Shapes (Theoretical Models of Inorganic Stereochemistry)*", Wiley, New York.

Burkert, U. and Allinger, N.J., (**1982**) "*Molecular Mechanics*", American Chemical Society, Washington, DC.

Cabani, S. and Gianni, P., *J. Chem. Soc., Faraday Trans. I.*, 75 (**1979**), 1184.

Cabani, S., Conti, G. and Lepori, L., *Trans. Faraday Soc.*, 67 (**1971**), 1933.

Cabani, S., Gianni, P. Mollica, V. and Lepori, L., *J. Solu. Chem.*, 10 (**1981**), 563.

Cabani, S. Conti, G. and Matteoli, E., *J. Chem. Soc., Faraday Trans. I*, 74 (**1978**), 2408.

Canet, D. and Robert, J.B., (**1994**) in: "*Dynamics of Solutions and Fluid Mixtures by NMR*", Delpuech, J.-J. (Ed.), John Wiley and Sons, New York.

Carbó, R., Leyda, L., and Arnau, M., *Int. J. Quantum Chem.*, 17 (**1980**), 1185.

Carpenter, F., McGregor, W., and Close, J., *J. Am. Chem. Soc.*, 81 (**1959**), 849.

Carrupt, P-.A., Testa, B., and Gaillard, P., (**1997**) "*Computational Approaches to Lipophilicity: Methods and Applications*", in: "Reviews in Computational Chemistry", Volume 11, Edt. By Wiley-VCH, John Wiley ans Sons, Inc., New

York.

Cassidy, R. and Janoski, M., *LC-GC*, 10 (**1992**), 692.

Chan, D.Y.C., Mitchell, D.J., Ninham, B.W. and Pailthorpe, B.A., (**1979**) in: "*Water: A Comprehensive Treatise*", Vol. 6, Franks, F. (Ed.), Plenum Press, New York.

Chan, H.S. and Dill, K.A., *J. Chem. Phys.*, 101 (**1994**), 7007.

Cohn, E. and Edsal, J., (**1943**) "*Proteins, Amino Acids and Peptides*", Reinhold, New York, p. 200.

Connolly, M.L., *Science*, 221 (**1983 a**), 709.

Connolly, M.L., *J. Appl. Crystallog.*, 16 (**1983 b**), 548.

Connolly, M.L., *J. Am. Chem. Soc.*, 107 (**1985**), 1118.

Cornell, N.W., Hansch, C., Kim, K.H. and Henegar, K., *Arch. Biochem. and Biophys.*, 227 (**1983**), 81.

Cramer, C.J. and Truhlar, D.G., *J. Am. Chem. Soc.*, 113 (**1991**), 8305.

Cramer, C.J. and Truhlar, D.G., *Science*, 256 (**1992 a**), 213.

Cramer, C.J. and Truhlar, D.G., *J. Comput. Chem.*, 13 (**1992 b**), 1089.

Cramer, C.J. and Truhlar, D.G., *J. Comp.-Aided Mol. Des.*, 6 (**1992 c**), 629.

Cramer, R.D., Patterson, D.E. and Bunce, J.D., *J. Am. Chem. Soc.*, 110 (**1988**), 5959.

Cratin, P.D., *Ind. Eng. Chem.*, 60 (**1968**), 14.

Croizet, F., Langlois, M.H, Dubost, J.P., Braquet, P., Audry, E., Dallet, Ph., and Colleter, J.C., *J. Mol. Graphics*, 8 (**1990**), 153.

Crook, E., Fordyce, D., and Trebbi, G., *J. Colloid Sci.*, 20 (**1965**), 191.

Culberson, J.C. and Walters, D.E., (**1991**) in: "*Sweeteners, Discovery, Molecular Design, and Chemoreception*", Walters, D.E., Orthoefer, F.T., and DuBois, G.E. (Eds.); American Chemical Society, Washington, DC, p. 214-223.

Dahlbom, R. and Tolf, B.R., *Biochem. and Biophys. Resear. Commun.*, 57 (**1974**), 549.

Dean, P.M., (**1987**) "*Molecular Foundations of Drug-Receptor Interactions*", Cambridge University Press, Cambridge.

De Gennes, P.G. and Taupin, C., *J. Phys. Chem.*, 86 (**1982**), 2294.

Delpuech, J.-J, (**1994**) in: "*Dynamics of Solutions and Fluid Mixtures by NMR*", Delpuech, J.-J. (Ed.), John Wiley and Sons, New York.

Derjaguin, B.V. and Landau, L., *Acta Physicochim, URSS*, 14 , (**1941**), 633.

Derjaguin, B.V., *Kolloid Zeits*, 69 (**1934**), 155.

Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., and Stewart, J.J.P. *J. Am. Chem. Soc.*, 107 (**1985**), 3902.

De Young, L.R. and Dill, K.A., *J. Phys. Chem.*, 94 (**1990**), 801.

Diamond, J.M. and Katz, Y., *J. Membr. Biol.*, 17 (**1974**), 121.

Dogonadze, R.R., Kornyshev, A.A. and Ulstrup, J., (**1985**) in: "*The Chemical Physics of Solvation, Part A: Theory of Solvation*", Dogonadze, R.R., Kálmán, E., Kornyshev, A.A. and Ulstrup, J. (Eds.), Plenum, New York.

Dorsey, G.J., *Chromatogr.*, 27 (**1987**), 167.

Du, Q. and Arteca, G.A., *J. Comput. Chem.*, 17 (**1996 a**), 1258.

Du, Q. and Arteca, G.A., *J. Compt.-Aided Mol. Design*, 10 (**1996 b**), 133.

Du, Q., Arteca, G.A., Mezey, P.G., *J. Compt.-Aided Mol. Design*, 11 (**1997**), 503.

Dughan, L., Burt, C. and Richards, W.G., *J. Mol. Struct (Theochem)*, 235 (**1991**), 481.

Edward, J.T. and Farrell, P.G., *Can. J. Chem.* 53 (**1975**), 2965.

Eisenberg, D., Weiss, R.M. and Terwilliger, T.C., *Nature* 299 (**1982**), 371.

Eisenberg, D. and McLachlan, A.D., *Nature*, 319 (**1986**), 199.

Eklund, H., Samana, J.-P. and Wallén, L., *Biochemistry*, 21 (**1982**), 4858.

Erricker, B.C., (**1971**) "*Advanced General Statistics*", The English University Press LTD, London.

Famini, G.R., Penski, C.A., and Wilson, L.Y., *J. Phys. Org. Chem.* , 5 (**1992**), 395.

Fauchère, J.L., Quarendon, P., and Kaetterer, L., *J. Mol. Graphics*, 6 (**1988**), 203.

Fendler, J.H., *Chem. Eng. News*, 62 (**1984**), 25.

Field, M. J., Bash, P. A., and Karplus, M., *J. Comput. Chem.*, 11 (**1990**), 700.

Fink, M. and Bonham, R.A., (**1981**) in: "*Chemical Applications of Atomic and Molecular Electrostatic Potential*", Politzer, P. and Truhlar, D.G. (Eds.), Plenum, New York.

Flory, P.J., (**1969**) "*Statistical Mechanics of Chain Molecules*," Interscience, New York.

Flory, P.J., (**1953**) "*Principles of Polymer Chemistry*", Cornell University, Ithaca.

Fogh, J., Rasmussen, P.O. and Skadhauge, K., *Anal. Chem.*, 26 (**1954**), 392.

Forsman, J. and Jönsson, B., *J. Chem. Phys.*, 101 (**1994**), 5116.

Francl, M.M., Houk, R.F., and Hehre, W.J., *J. Am. Chem. Soc.*, 106 (**1984**), 563.

Frank, H. and Evans, M., *J. Chem. Phys.*, 13 (**1945**), 507.

Frank, R., (**1984**) "*Theoretical Drug Design Methods*", Akademie Verlag.

Frecer, V., Miertuš, S. and Majekova, M., *J. Mol. Struct. (Theochem)*, 216 (**1991**), 312.

Fries, R.W., Bohlen, D.P., and Plapp, B.V., *J. Med. Chem.*, 22 (**1979**), 356.

Friesner, R. and Levy, R.M., *J. Chem. Phys.*, 80 (**1984**), 4488.

Frisch, M.J., Trucks, G.W., Head-Gordon, M., Gill, P.M.W., Wong, M.W., Foresman, J.B., Johnson, B.G., Schlegel, H.B., Robb, M.A., Replogle, E.S., Gomperts, R., Andrés, J.L., Raghavachari, K., Binkley, J.S., González, C., Martin, R.L., Fox, D.J., DeFrees, D.J., Baker, J., Stewart, J.J.P., and Pople, J.A., Gaussian, Inc., (**1992**) "*Gaussian 92 (Revision E.2)*", Pittsburgh PA.

Furet, P., Sele, A. and Cohen, N.C., *J. Mol. Graphics*, 6 (**1988**), 182.

Gadre, S.R., Shrivastava, I.H., and Kulkarni, S.H., *Chem. Phys. Lett.*, 170 (**1990**), 271.

Gago, F., Builla, J.A., Elguero, J., and Diez-Masa, J., *Anal Chem.*, 59 (**1987**), 921.

Gaillard, P., Carrupt, P.-A. and Testa, B., *J. Mol. Graph.*, 12 (**1994**), 73.

Gaillard, P., Carrupt, P.-A., Testa, B. and Boudon, A., *J. Comput.-Aided Mol. Design*, 8 (**1994**), 83.

Gao, J., (**1994**) in: "*Modeling the Hydrogen Bond*", Smith, D.A. (Ed.), ACS, Washington, DC.

Gavezzotti, A., *J. Am. Chem. Soc.*, 105 (**1983**), 5220.

Geladi, P. and Kowalski, B., *Analytica Chimica Acta*, 185 (**1986 a**), 1.

Geladi, P. and Kowalski, B., *Analytica Chimica Acta*, 185 (**1986 b**), 19.

Ghose, A.K. and Crippen, G.M., *J. Med. Chem.*, 28 (**1985 a**), 333.

Ghose, A.K. and Crippen, G.M., *J. Comput. Chem.*, 6 (**1985 b**), 350.

Ghose, A.K. and Crippen, G.M., *J. Comput. Chem.*, 7 (**1986**), 565.

Ghose, A.K., Pritchett, A., and Crippen, G.M., *J. Comput. Chem.*, 9 (**1988**), 80.

Giesen, D.J., Cramer, C.J., and Truhlar, D.G., *J. Phys. Chem.*, 98 (**1994**), 4141.

Glidewell, C., *Inorg. Chim. Acta*, 12 (**1975**), 219.

Gordon, R.G., (**1968**) in: "*Advances in Magnetic Resonance*", vol. 3, Waugh, J.S. (Ed.), Academic Press, New York.

Grant, J.A. and Pickup, B.T., *J. Phys. Chem.*, 99 (**1995**), 3503.

Greenwald, H.L., Kice, E.K., Kenly, M., and Kelly, J., *Anal. Chem.*, 33 (**1961**), 465.

Guarnieri, F. and Still, W.C., *J. Comput. Chem.*, 15 (**1994**), 1302.

Guillot, B. Guissani, Y., and Bratos, S., *J. Chem. Phys.*, 95 (**1991**), 3643.

Guillot, B. and Guissani, Y., *Mol. Phys.*, 95 (**1993 a**), 53.

Guillot, B. and Guissani, Y., *J. Chem. Phys.*, 99 (**1993 b**), 3643.

Haile, J.M., (**1992**) "*Molecular Dynamics Simulation*", Wiley, New York.

Halle, B., Andersson, T. Forsén, S. and *Lindman B., J. Am. Chem. Soc*, 103 (**1981**), 500.

Halle, B. and Piculell, L., *J. Chem. Soc., Faraday Trans.* 82 (**1986**), 415.

Hallenga, K. and Koenig, S.H. *Biochemistry*, 15 (**1976**), 4255.

Hansch, C. and Fujita, T., *J. Am. Chem. Soc.* 86 (**1964**), 1616.

Hansch, C., *Acc. Chem. Res.*, 2 (**1969**), 232.

Hansch, C., (**1971**) in: "*Drug Design*", Ariens, E.J., (Ed.), Academic Press, New York, Vol. 1, p. 271.

Hansch, C., (**1978**) in: "*Correlation Analysis in Chemistry*", Chapman, N.B. and Shorter, J. (Eds.), Plenum, New York.

Hansch, C. and Leo, A.J., (**1979**) "*Substitute Constants for Correlation Analysis in Chemistry and Biology*", Wiley, New York.

Horjales, E., Eklund, H. and Branden, C.-I., *J. Mol. Biol.* 197 (**1987**), 685.

Head-Gordon, T., *J. Am. Chem. Soc.*, 117 (**1995**), 501.

Heiden, W., Moeckel, G., and Brickmann, J., *J. Comp.-Aided Mol. Des.*, 7 (**1993**), 503.

Helfrich, W. *Z. Naturforsch*, 33A (**1978**), 305.

Henczi, M., Nagy, J., and Weaver, D.F., *J. Liquid Chr.* 17 (**1994**), 2605.

Heermann, D.W., (**1990**) "*Computer Simulation Methods in Theoretical Physics*", Springer-Verlag, Berlin.

Hehre, W.J., Radom, L., Schleyer, P.v.R., and Pople, J.A., (**1986**) "*Ab Initio*

Molecular Orbital Theory", Wiley, New York.

Hermans, J., Berendsen, H.J.C., van Gunsteren, W.F., and Postma, J.P.M., *Biopolymers*, 23 (**1984**), 1513.

Herz, H.G., J. Soc., (**1973**) in: "*Water: A Comprehensive Treatise*", vol. 3, Franks, F. (Ed.), Plenum Press, New York.

Hirschfelder, J.O., Curtiss, C.F., and Bird, R.B., (**1964**) "*Molecular Theory of Gases and Liquids*", Wiley, New York.

Holtzer, A., *Biopolymers*, 32 (**1992**), 711.

Holtzer, A., *Biopolymers*, 34 (**1994**), 315.

Hopfinger, A.J., *J. Med. Chem.*, 26 (**1983**), 990.

Howarth, O.W., *J. Chem. Soc. Faraday Trans.*, 12 , (**1975**), 2303.

Hubbard, P.S., *J. Chem. Phys.*, 53 (**1970**), 985.

Israelachvili, J.N., (**1992**) "*Intermolecular and Surface Forces*", Academic Press, London.

Israelachvili, J.N. and Wennerstrom,H., *Nature*, 379 (**1996**), 219.

Israelachvili, J.N. and Pashley, R.M., *Nature*, 300 (**1982**), 341.

Iwase, K., Katsuichiro, K., Hirono, S., Nakagawa, S., and Moriguchi, I., *Chem. Pharm. Bull*, 33 (**1985**), 2114.

Jaffé, H. H., *Chem. Rev.*, 53 (**1953**), 191.

Jain, A.N., Koile, K. and Chapman, D., *J. Med. Chem.*, 37 (**1994**), 2315.

Johnson, M.A. and Maggiora, G.M. (Eds.), (**1990**) "*Concepts and Applications of Molecular Similarity*", Wiley, New York.

Jönsson, B., *Chem. Phys. Lett.*, 82 (**1981**), 520.

Jorgensen, W.L. and Tirado-Rives, J., *J. Am. Chem. Soc.*, 110 (**1988**), 1657.

Kahn, S.D., Pau, C.F., Overman, L.E., and Hehre, W.J., *J. Am. Chem. Soc.*, 108 (**1986**), 7381.

Kallszan, R., *J. Chromatogr. Sci.*, 22 (**1984**), 362.

Kamlet, M.J., Abboud, J.L.M., and Taft, R.W., *Prog. Phys. Org. Chem.*, 13 (**1981**), 485.

Kamlet, M.J., and Taft, R.W., *Acta Chim. Scand. Ser. B*, 39 (**1985**), 611.

Kamlet, M.J., Doherty, R.M., Abboud, J.-M., Abraham, M.H., Marcus, Y., and Taft,

R.W., *CHEMTECH*, 16 (**1986**), 566.

Kamlet, M.J., Doherty, R.M., Abraham, M.H., Marcus, Y., and Taft, R.W., *J. Phys. Chem.*, 92 (**1988**), 5244.

Kantola, A., Villar, H. O., and Loew, G. H., *J. Comput. Chem.*, 12 (**1991**), 681.

Kauzmann, W., *Adv. Protein Chem.*, 14 (**1959**), 37.

Kearsley, S.K. and Smith, G.M., *Tet. Comput. Method*, 3 (**1992**), 615.

Kellogg, G.E., Kier, L.B. and Hall, L.H., *J.Comput-Aided Molec. Design*, (**1997**) submitted for publication.

Kellogg, G.E. and Abraham, D.J., *J. Mol. Graphics*, 10 (**1992**), 212.

Khaledi, M.G., *Biochromatography*, 3 (**1987**), 20.

Khaledi, M.G., *Anal. Chem.*, 60 (**1988**), 876.

Khaledi, M.G. and Breyer, E.D., *Anal. Chem.*, 61 (**1989**), 1040.

Kier, L.B. and Hall, L.H., (**1992**) in: "*Advances in Drug. Design*", vol. 22, Testa, B. (Ed.), Academic Press.

Kim, K.H., *Quant. Struct.-Act. Relat*, 12 (**1993**), 232.

Klebe, G., Abraham, U., and Mietzner, T., *J. Med. Chem.* 37 (**1994**), 4130.

Kleinbaum, D.G. Kupper, L.L. and Muller, K.E., (**1988**), "*Applied Regression Analysis and Other Multivariables Methods*", PWS-KENT, Boston.

Klir, G.J. and Yuan, B., (**1995**) "*Fuzzy Sets and Fuzzy Logic, Theory and Applications*", Prentice Hall PTR, Upper Saddle River, New Jersey.

Klopman, G., Li, J.-Y., Wang, S. and Dimayuga, M., *J. Chem. Inf. Comput. Sci.*, 34 (**1994**), 752.

Klopman, G., Namboodiri, K., and Schochet, M., *J. Comput. Chem.*, 6 (**1985**), 28.

Klopman, G. and Iroff, L.D., *J. Comput. Chem.*, 2 (**1981**), 157.

Koenig, S.H., Hallenga, K. and Shporer, M., *Proc. Natl. Acad. Sci. USA*, 72 (**1975**), 2667.

Kollman, P.A., *Curr. Opin Struct. Biol.*, 4 (**1994**), 240.

Kollman, P.A., (**1981**) in: "*Chemical Applications of Atomic and Molecular Electrostatic Potential*", Politzer, P. and Truhlar, D.G. (Eds.), Plenum, New York.

Kollman, P.A., *Acc. Chem. Res.*, 10 (**1977**), 365.

Laaksonen, A. and Stilbs, P., *Mol. Phys.*, 74 (**1991**), 747.

Langridge, R., Ferrin, T.E., Kuntz, I.D., and Connolly, M.L., *Science*, 211 (**1981**), 661.

Lee, B. and Richards, F. M., *J. Mol. Biol.*, 55 (**1971**), 379.

Lee, S.H. and Rossky, P.J., *J. Chem. Phys.*, 100 (**1994**), 3334.

Lee, C.Y., McCammon, J.A. and Rossky, P.J., *J. Chem. Phys.*, 80 (**1984**), 4448.

Lemieux, R.U., *Acc. Chem. Res.*, 29 (**1996**), 373.

Leo, A., Hansch, C. and Elkins, D., *Chem. Revs.*, 71 (**1971**), 525.

Lindman, B., Olsson, U. and Söderman, O., (**1994**) *in: "Dynamics of Solutions and Fluid Mixtures by NMR, 8. Surfactant Solutions: Aggregation Phenomena and Microheterogeneity"*, Delpuech, J.-J. (Ed.), John Wiley and Sons.

Liu, H. and Shi, Y., *J. Comput. Chem.*, 15 (**1994**), 1311.

Loew, G.H., Villar, H.O., and Alkorta, I., *Pharmaceutical Research*, 10 (**1993**), 475.

Luo, X., Arteca, G.A., Mezey, P.G., and Zhang, C.-H., *J. Organomet. Chem.*, 444 (**1993**), 131.

Luque, F.J., Gadre, S.R., Bhadane, P.K., and Orozco, M., *Chem. Phys. Lett.*, 232 (**1995**), 509.

Lybrand, T.P., (**1990**) in: "*Reviews in Computational Chemistry*", vol 1., Lipkowitz, K.B. and Boyd, D.B. (Eds.), VCH, New York.

Madden, P. and Kivelson, P. *J. Chem. Phys.*, 56 (**1984**), 1057.

Marcelja, S., Mitchell, D.J., Ninham, B.W. and Sculley, M.J., *J. Chem. Soc. Faraday Trans.* II 73 (**1977**), 630.

Mardia, K.V., Kent, J.T. and Bibby, J.M., (**1979**) *"Multivariate Analysis"*, Academic Press, New York.

Marshall, G.R. and Cramer III, R.D., *Trends Pharm. Sci.*, 9 (**1988**), 285.

Masek, B.B., Marchant, A. and Matthew, J.B., *J. Med. Chem.*, 36 (**1993**), 1230.

McCammon, J.A. and Harvey, S.C., (**1987**) "*Dynamics of proteins and nucleic acids*", Cambridge University Press, Cambridge (UK).

McDonald, D.Q. and Still, W.C., *J. Am. Chem. Soc.*, 116 (**1994**), 11550.

McQuarrie, D.A., (**1976**) "*Statistical Mechanics*", Harper and Row, New York.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E., *J. Chem. Phys.*, 21 (**1953**), 1087.

Meyer, A.Y. and Richards, W.G., *J. Comput.-Aided Mol. Design*, 5 (**1991**), 427.

Meyer, K.H. and Hemmi, H., *Biochem. Z.*, 277 (**1935**), 39.

Mezey, P.G., *Struct. Chem.*, 6 (**1995 a**), 261.

Mezey, P.G., *J. Math. Chem.*, 18 (**1995 b**), 141.

Mezey, P.G., (**1993**) "*Shape in Chemistry: An Introduction to Molecular Shape and Topology*", VCH Publishers, New York.

Mezey, P.G., (**1990**) in: "*Reviews of Computational Chemistry*", vol.1, Lipkowitz, K.B. and Boyd, D.B. (Eds.), VCH, New York.

Mezey, P.G., *Int. J. Quantum Chem. QBS*, 12 (**1986**), 113.

Miertuš, S. and Moravek, R., *Coll. Czech-Slov. Chem. Commun.*, 55 (**1990**), 2430.

Miertuš, S. and Tomasi, J., *Chem. Phys.*, 65 (**1982**), 239.

Miertuš, S., Scrocco, E., and Tomasi, J., *Chem. Phys.*, 55 (**1981**), 117.

Mishra, P.C. and Kumar, A., (**1995**) in: "*Topics in Current Chemistry (174), Molecular Similarity II*", Sen, K. (Ed.), Springer-Verlag, New York.

Miller, K.W., Hammond, L., and Porter, E.G., *Chem. Phys. Lipids*, 20 (**1977**), 229.

Mishra, P.C. and Kumar, A., (**1995**) in: "*Topics in Current Chemistry, 174, Molecular Similarity II*", Sen, K. (Ed.), Springer, New York.

Morokuma K. and Kitaura, K., (**1981**) in: "*Chemical Applications of Atomic and Molecular Electrostatic Potential*", Politzer, P. and Truhlar, D.G. (Eds.), Plenum, New York.

Murphy, M.M., Schullek, J.R., Gordon, E.M., and Gallop, M.A., *J. Am. Chem. Soc.*, 117 (**1995**), 7029.

Murray, J.S. and Politzer, P., *J. Org. Chem.*, 56 (**1991 a**), 6715.

Murray, J.S. and Politzer, P., *J. Org. Chem.*, 56 (**1991 b**), 3734.

Murray, J.S., Ranganathan, S. and Politzer, P., *J. Org. Chem.*, 56 (**1991 c**), 3734.

Murray, J.S. and Politzer, P., *J. Chem. Res. (Synop.)*, 1992 (**1992**), 110.

Murray, J.S., Lane, P., Brinck, T., and Politzer, P., *J. Phys. Chem.*, 97 (**1993 a**), 5144.

Murray, J.S., Lane, P., Brinck, T., Paulsen, K., Grice, M.E. and Politzer, P., *J. Phys. Chem.*, 97 (**1993 b**), 9369.

Murray, P., Brinck, T., and Politzer, P., *J. Phys. Chem.*, 97 (**1993 c**), 13807.

Murray, J.S., Brinck, T., Lane, P., Paulsen, K., and Politzer, P., *J. Mol. Strct. (Theochem)*, 307 (**1993 d**), 55.

Náray-Szabó, G. and Nagy, P., *Int. J. Quantum. Chem.*, 35 (**1989**), 215.

Náray-Szabó, G., *J. Mol. Graph.*, 7 (**1989 a**), 76.

Náray-Szabó, G., *Int. J. Quantum. Chem.*: *quantum Biology Symposium*, 16 (**1989 b**), 87.

Náray-Szabó, G., *J. Mol. Struct. (Theochem)*, 138 (**1986**), 197.

Némethy, G. and Scheraga, H. A., *J. Chem. Phys.*, 36 (**1962**), 3401.

Nernst, W., *Z. Phys. Chem.*, 8 (**1891**), 110.

Northrup, S.H. and McCammon, J.A., *Biopolymers*, 19 (**1980**), 1001.

Ooi, T., Oobatake, M., Némethy, G., and Scheraga, H. A., *Proc. Natl. Acad. Sci. USA*, 84 (**1987**), 3086.

Pascual-Ahuir, J.-L., Silla, E., Tomasi, J., and Bonaccorsi, R., *J. Comput. Chem.*, 8 (**1987**), 778.

Pastor, M. Cruciani, G. and Clementi, S., *J. Med. Chem.*, 14 (**1997**), 1455.

Pathak, R.K. and Gadre, S.R., *J. Chem. Phys.*, 93 (**1990**), 1770.

Pauling, L., (**1960**) "*The Nature of the Chemical Bond*", Cornell University, Ithaca.

Pearlman, R.S., (**1986**) in: "*Partition Coefficient Determination and Estimation*", Dunn III, W.J., Block, J.H., and Pearlman, R.S. (Eds.), Pergamon Press, New York.

Piculell, L., *J. Chem. Soc., Faraday Trans.* 82 (**1986**), 387.

Piculell, L. and Halle, B., *J. Chem. Soc., Faraday Trans.* 82 (**1986**), 401.

Platt, D.E. and Silverman, B.D., *J. Comp. Chem.*, 17 (**1996**), 358.

Pleiss, M.A. and Grunewald, G.L., *J. Med. Chem.*, 26 (**1983**), 1760.

Politzer, P., (**1981**) in: "*Chemical Applications of Atomic and Molecular Electrostatic Potential*", Politzer, P. and Truhlar, D.G. (Eds.), Plenum, New York.

Politzer, P. and Daiker, K.C., (**1981**) in: "*The Force Concept in Chemistry*," Chapter 6, Deb, B.M. (Ed.), Van Nostrand Reinhold, New York.

Politzer, P., Parr, R.G., and Murphy, D.R., *J. Chem. Phys.*, 79 (**1983**), 3859.

Politzer, P. and Murray, J.S., (**1991**) in "*Reviews in Computational Chemistry*", vol. 2, Lipkowitz, K.B. and Boyd, D.B. (Eds.), VCH, New York.

Politzer, P., Lane, P., Murray, J.S., and Brinck, T., *J. Phys. Chem.*, 96 (**1992**), 7938.

Politzer, P., Murray, J.S., Lane, P., and Brinck, T., *J. Phys. Chem.*, 97 (**1993**), 729.

Politzer, P., Murray, J.S., Concha, M.C., and Brinck, T., *J. Mol. Struct. (Theochem)*, 281 (**1993**), 107

Purdum, J., (**1989**) "*C Programmer's Toolkit*", Que Corporation, USA.

Rekker, R.F., (**1977**) "*The Hydrophobic Fragment Constants*", Elsevier, New York.

Rekker, R.F. and de Kort, H.M., *Eur. J. Med. Chem.*, 14 (**1979**), 479.

Richards, F.M., *Annu. Rev. Biophys. Bioeng.*, 6 (**1977**), 151.

Richards, W.G., (**1983**) "*Quantum Pharmacology*", 2nd ed., Butterworth, London.

Romano, S. and Singer, K. *Mol. Phys.*, 37 (**1979**), 1765.

Rotschild, W.G., (**1984**) "*Dynamics of Molecular Liquids*", John Wiley and Sons, New York.

Rossini, F.D., (**1950**) "*Chemical Thermodynamics*", Wiley, New York.

Rozas, I., Du, Q., and Arteca, G.A., *J. Mol. Graphics*, 13 (**1995**), 98.

Rozas, I. and Arteca, G.A., *Can. J. Chem.*, 70 (**1992**), 2296.

Rozas, I., Arteca, G.A., and Mezey, P.G., *Int. J. Quantum Chem., QBS*, 18 (**1991**), 269.

Scrocco, E. and Tomasi, J., *Topics Curr. Chem.*, 42 (**1973**), 95.

Sen, K.D. and Politzer, P., *J. Chem. Phys.*, 90 (**1989 a**), 4370

Sen, K.D. and Politzer, P., *J. Chem. Phys.*, 91 (**1989 b**), 5123.

Sharp, K.A., Nicholls, A., Fine, R.F., and Honig, B., *Science* 252 (**1991**), 106.

Shinoda, K. and Hildebrand, J.H., *J. Phys. Chem.*, 62 (**1958**), 292.

Singh, U.C. and Kollman, P.A., (**1982**) *QCPE Bull.*, 117, prog. no. 446.

Singh, U.C., Weiner, P.K., Caldwell, J., and Kollman, P.A., (**1986**) "*Amber 3.0*", University of California, San Francisco.

Skipper, T.N., *Chem. Phys. Lett.*, 207 (**1993**), 424.

Smith, D.E. and Haymet, A.D.J., *J. Chem. Phys.*, 98 (**1993**), 6445.

Smith, J.C. and Karplus, M., *J. Am. Chem. Soc.*, 114 (**1992**), 805.

Steele, W.A., (**1976**) in: "*Advances in Chemical Physics*", vol. XXXIV, Prigogine, I. and Rice, S.A. (Eds.), John Wiley and Sons, New York.

Stillinger, F.H. and Rahman, A., *J. Chem. Phys.*, 60 (**1974**), 1545.

Suresh, L. and Walz, J.Y., *J. Colloid and Interface Science*, 183 (**1996**), 199.

Taft, R.W., Abboud, J.-M., Kamlet, M.J., and Abraham, M.H., *J. Solut. Chem.*, 14 (**1985**), 153.

Tanaka, H. and Nakanishi, K., *J. Chem. Phys.*, 95 (**1991**), 3719.

Tanford, C., (**1961**) "*Physical Chemistry of Macromolecules*", Chap. 4, John Wiley, New York.

Tanford, C., (**1973**) "*The Hydrophobic Effect*", Wiley, New York.

Tasi, G. and Pálinkó, I., (**1995**), in: "*Topics in Current Chemistry, 174, Molecular Similarity II*", Sen, K. (Ed.), Springer, New York.

Terasawa, S., Itsuki, H. and Arakawa, H., *J. Phys. Chem.*, 79 (**1975**), 2345.

Theorell, H., Yonetani, T., and Sjöberg, B., *Acta Chim. Scand.*, 23 (**1969**), 255.

Tokarski, J.S. and Hopfinger, A.J., *J. Med. Chem.* 37 (**1994**), 3639.

Tolf, B.R., Dahlbom, R., Åkeson, Å. and Theorell, *Acta Pharm. Suec.*, 22 (**1985**), 147.

Tolf, B.R., Siddiqui, Dahlbom, R., Åkeson, Å. and Theorell, H., *Eur. J. Med.Chem.*, 17 (**1982**), 395.

Tolf, B.R., Plechaczek, J., Dahlbom, R., Theorell, H., Akeson, A. and Lundquist, G., *Acta Chem. Scand. B*, 33 (**1979**), 483.

Tomasi, J., (**1981**), in: "*Chemical Applications of Atomic and Molecular Electrostatic Potentials*", Politzer, P. and Truhlar, D.G. (Eds.), Plenum Press, New York and London.

Tomlinson, E., *J. Chromatogr.*, 113 (**1975**), 1.

Treiner, C., *J. Colloid Interface Sci.*, 93 (**1986**), 101.

Treiner, C. and Chattopadhyay, A.K., *J. Colloid Interface Sci.*, 109 (**1986**), 101.

van Gunsteren, W.F. and Berendsen, H.J.C., *Angew. Chem. (Int. Engl. Ed.)*, 29 (**1990**), 992.

Vásquez, M., Némethy, G., and Scheraga, H.A., *Macromolecules*, 16 (**1983**), 1043.

Ventura, O.S., Lledòs, A., Bonnaccorsi, R., Bertran, I., and Tomasi, J., *Theor. Chim. (Berlin)* 72 (**1987**), 175.

Verloop, A., Hoogenstraaten, W. and Tipker, J., (**1976**) in: "*Drug Design*", vol. 7, Ariens, J. (Ed.), Academic Press, New York, p165.

Verwey, E.J.W. and Overbeek, J.Th.G. (**1948**), "*Theory of Stability of Lyophobic Colloids*", Elsevier, Amsterdam.

Viswanadhan, V.N., Ghose, A.K., Revankar, G.R., and Robins, R.K., *J. Chem. Inf. Comput. Sci.*, 29 (**1989**), 163.

Viswanadhan, V.N., Reddy, M.R., Bacquet, R.J., and Erion, M.D., *J. Comput Chem.*, 14 (**1993**), 1019.

Walker, P.D. and Mezey, P.G., *J. Am. Chem. Soc.*, 116 (**1994**), 12022.

Walker, P.D. and Mezey, P.G., *J. Am. Chem. Soc.*, 115 (**1993**), 12423.

Waller, C.L. and Kellogg, G.E., *Network Science*, 2 (1) (**1996**), http://www.awod.com/netsci/Science/Compchem/feature10.html.

Waller, C.L. and Marshall G.R., *J. Med. Chem.*, 36 (**1993**), 2390.

Waller, C.L. Evans, M.V. and McKinney, J.D., (**1995**), *Drug Metab. Disp.*, in press.

Wallqvist, A., *Chem. Phys. Lett.*, 165 (**1989**), 437.

Wallqvist, A. and Berne, B.J., *Chem. Phys. Lett.*, 145 (**1988**), 26.

Walters, D.E., *Network Science*, 3 (1) (**1997**), http://www.awod.com/netsci/Science/Compchem/feature03.html.

Watanabe, Y. and Mitsui, Y., (**1981**) in: "*101st Annual Meeting of Pharmaceutical Society of Japan*", p. 198.

Walters, D.E., *Network Science*, **3** (1) (**1997**), http://www.awod.com/netsci/Science/Compchem/feature03.html.

Wei, D. and Patey, G.N., *J. Chem. Phys.*, 91 (**1989**), 7113.

Weiner, S.T., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P., *J. Am Chem., Soc.*, 106 (**1984**), 765.

Weiner, S.T., Kollman, P.A., Nguyen, D.T., and Case, D.A., *J. Comput. Chem.*, 7 (**1987**), 230.

Weinstein, H., *Int. J. Quant. Chem. Quant. Biol. Symp.*, 2 (**1975**), 59.

Weinstein, H., (**1981**) in: "*Chemical Application of Electrostatic Potentials*", Politzer, P. and Truhlar, D.G. (Eds.), Plenum, New York, p. 309.

Widom, B., *J. Chem. Phys.*, 39 (**1963**), 2808.

Williams, G., *Chem. Soc. Rev.*, 89 (1978), 7.

Williams, R.J.P., Eur. J. Biochem, 183 (**1989**), 479.

Wilson, E.B., Decius, J.C. and Cross, P.C., (**1955**) "*Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*", McGraw-Hill, New York.

Wong, M.W., Frisch, M. J., and Wiberg, K. B., *J. Am. Chem. Soc.*, 113 (**1991**), 4776.

Worth, B. and Reid, E., *J. Am. Chem. Soc.*, 38 (**1916**), 2316.

Appendix: Programs Used in Thesis Research

Five FORTRAN programs are used in my thesis research.

1. Gaussian 92: Calculate the MEP on the grid points of molecular surface.
2. MS.FOR: Build the molecular van der Waals surface established by a set of grid points.
3. MEPG92.FOR: Make command files for Gaussian 92 using the information of molecular surface calculated by MS.FOR.
4. CUT.FOR: Cut the MEP information from the output files by Gaussian 92, and make a data file for MEPMLP.FOR.
5. MEPMLP.FOR: Calculate the empirical and heuristic molecular lipophilicity potentials (EMLP and HMLP).

Program Gaussian 92 is a quantum chemical software package written by Gaussian Inc. [Frisch *et al.* 1992]. Program MS.FOR is a software package for molecular surfaces presented by Connolly [1983 a, b, 1985]. I made some changes in MS.FOR for my HMLP calculations. Program MEPG92.FOR, CUT.FOR, and MEPMLP.FOR are written by the author. Program MEPMLP.FOR is used for the calculations in Chapters 3 and 5. In the calculations of Chapter 4, optimizations of screening functions and parameters, a different version of MEPMLP.FOR, called FUNCTION.FOR, is used. The last three programs are appended below.

1. **MEPMLP.FOR**

```
C     F77 PROGRAM: MEPMLP.FOR                JAN. 1, 1996
C     Version 3 (For Chapter 3 and Chapter 5 of thesis. Chapter 4 uses a different version.)
C     University of Saskatchewan
C     Qishi Du
C
C     Program is for the calculations of the lipophilic potential of a molecule, using
C     two methods:
C     1). Empirical Molecular Lipophilicity Potential (EMLP)
C        Suggested by Audry et al.(E. Audry, J.P.
C        Dubost, J.C. Colleter, and P. Dallet, (1986) Eur. J. Med.
C        Chem. 21, 71).
C        Two data sets are used in this method provided by Crippen's group:
C            old data set (A.K. Ghose and G.M. Crippen, (1986) J. Comput. Chem. 7, 565)
C            new data set (V.N. Viswanadhan, A.K. Ghose, G.R. Revankar, and R.K.
C                Robins, (1989) J. Chem. Inf. Comput. Sci., 29,163).
C     2). Heuristic Molecular Lipophilicity Potential (HMLP)
C        Suggested by author in my thesis research.
C
C     Two input data files are used in this program.
C     1). MLPTYP.DAT: Molecular geometry established by nuclear cartesian coordinates,
C            and atomic types used in EMLP.
C     2). MEPXYZ.DAT: Molecular surface established by a set of cartesian
C            coordinates of grid points on molecular surface, MEP, and area of
C            surface elements.
C
C     Two output files are produced in this program.
C     1). MEPXYZ.DAT contains the following information at every grid point:
C            X  Y  Z  MEP  MLP(Theory)  MLP(Empirical)  AREA No. of Atom
C
C     2). MEPMLP.OUT holds the statistical information of MEP and MLP on the
C        molecular surface.
C
      CHARACTER*4 ATYPE(200),DATASET*23
      COMMON ATYPE
      DIMENSION X(200),Y(200),Z(200),X2(8000),Y2(8000),Z2(8000)
      DIMENSION AMEP(8000),AREA(8000),IAT(8000),AMLP1(8000)
```

```
      DIMENSION AMLP2(8000),ATOME(200),ATOML(200),MNAT(200)
      DIMENSION TEMP(2000),ATOMB(200),ATOMA(200)

      OPEN(UNIT=31,FILE='MLPTYP.DAT',STATUS='OLD')
      OPEN(UNIT=32,FILE='MEPXYZ.DAT',STATUS='OLD',
     :  FORM='UNFORMATTED')
      OPEN(UNIT=61,FILE='MEPMLP.OUT',STATUS='UNKNOWN')
      OPEN(UNIT=62,FILE='MEPMLP.DAT',STATUS='UNKNOWN')

C     MLPTYP.DAT is as following (free format):
C     line1: NATOM  NPOINT  KEY1
C     linex: X(I)   Y(I)   Z(I)  ATYPE(I)

C     KEY1=1: new data set of atomic lipophilicity parameters
C          =2: old data set of atomic lipophilicity parameters
C     ATYPE(i) is character data for atom i (A4). The lipophylic
C     atomic symbol of atom i can be found in reference: Ghose and
C     Crippen J. COMPUT. CHEM. Vol. 7 565-577 (1986).

      READ(31,*) NATOM,NP,KEY1
      DO 10 I=1,NATOM
10    READ(31,*) X(I),Y(I),Z(I),ATYPE(I)

C     MEPXYZ.DAT CONTAINS (free format):
C     X2(I)  Y2(I)  Z2(I)  AMEP(I)  AREA(I)  IAT(I)
C     MEP in atomic unit,
C     Coordinates in Angstrom,
C     Area in square Angstrom.
C     IAT(I) IS INDEX NUMBERING THE ATOMIC SPHERE ON WHICH THE
C     POINT LIVES.
C
      SCALE=627.525138775
      FACTOR=0.52917706
      SFACTOR=FACTOR*FACTOR

      DO 20 I=1,NP
20    READ(32) X2(I),Y2(I),Z2(I),AMEP(I),AREA(I),IAT(I)
C     AMEP(I)=AMEP(I)*SCALE
```

```
      CLOSE(UNIT=31)
      CLOSE(UNIT=32)

C     *****
C     PART 1: MOLECULAR MEP DISTRIBUTIONS
C     *****

      WRITE(*,*) '***** Part 1: MOLECULAR MEP DISTRIBUTIONS'
      CALL MINIMA (NP,AMEP,PMI,PMA,IMI,IMA)

      WRITE(61,*) '***** PART 1: MOLECULAR MEP DISTRIBUTION *****'
      WRITE(61,*)
      WRITE(61,*) '(MEP IN ATOMIC UNIT, COORDINATES IN ANGSTROM)'
      WRITE(61,*)
      IJ=IAT(IMI)
      WRITE(61,*)' MIN. MEP     X       Y       Z   ',
     :      '   ATOM     NO.'
      WRITE(61,1111) PMI,X2(IMI),Y2(IMI),Z2(IMI),ATYPE(IJ),IMI
      WRITE(61,*)
      IJ=IAT(IMA)
      WRITE(61,*)' MAX. MEP     X       Y       Z   ',
     :      '   ATOM     NO.'
      WRITE(61,1111) PMA,X2(IMA),Y2(IMA),Z2(IMA),ATYPE(IJ),IMA
      WRITE(61,*)
1111  FORMAT(1X,4F10.5,5X,A4,I8)

C     MNAT(I):  Total points on Atom I
C     ATOME(I): Total MEP on Atom I

      WRITE(61,*)'MEP DISTRIBUTIONS ON ATOMS'
      WRITE(61,*)'ATOM','      ','ATOMIC MEP','    ','MINIMUM',
     &      '      ','MAXIMUM','    ','POINTS'

      DO 15 I=1,NATOM
      MNAT(I)=0
15    ATOME(I)=0.0
      DO 25 I=1,NP
```

```
      IJ=IAT(I)
      MNAT(IJ)=MNAT(IJ)+1
      ATOME(IJ)=ATOME(IJ)+AMEP(I)
25    CONTINUE

C     Minimum and Maximum MEP on Atoms

      IJ=0
      DO 60 I=1,NATOM
      ITOL=MNAT(I)
      DO 70 J=1,ITOL
70    TEMP(J)=AMEP(IJ+J)
      CALL MINIMA(ITOL,TEMP,PMI,PMA,IMI,IMA)
      WRITE(61,1112) ATYPE(IAT(IJ+1)),ATOME(I),PMI,PMA,ITOL
      IJ=IJ+ITOL
60    CONTINUE
1112  FORMAT(1X,A4,'  ',3F14.6,I9)

C     *****
C     PART 2: MEP SURFACE-DESCRIPTORS
C     *****

      WRITE(*,*) '***** PART 2: MEP SURFACE-DESCRIPTORS'
      WRITE(61,*)
      WRITE(61,*)'***** PART 2: MEP SURFACE-DESCRIPTORS'
      MP=0
      MN=0
      VPM=0.0
      VNM=0.0
      SP=0.0
      SN=0.0
      BP=0.0
      BN=0.0

      DO 50 I=1,NP
      AMEP(I)=AMEP(I)*SCALE
      IF (AMEP(I).LT.0.0) THEN
            MN=MN+1
```

```
                VNM=VNM+AMEP(I)
                SN=SN+AREA(I)
                BN=BN+AMEP(I)*AREA(I)
        ELSE
C               IF (AMEP(I).EQ.0.0) GOTO 50
                MP=MP+1
                VPM=VPM+AMEP(I)
                SP=SP+AREA(I)
                BP=BP+AMEP(I)*AREA(I)
        ENDIF
        IJ=IAT(I)
        ATOME(IJ)=ATOME(IJ)+AMEP(I)
50      CONTINUE
        ST=SP+SN

        WRITE(61,*)
        WRITE(61,*)'SURFACE-MEP DESCRIPTORS ON MOLECULE:'
        WRITE(61,*)'POSITIVE AREA   NEGATIVE AREA    TOTAL AREA'
        WRITE(61,1113) SP,SN,ST
        WRITE(61,1114) SP/SFACTOR,SN/SFACTOR,ST/SFACTOR
        WRITE(61,*)
        WRITE(61,*)'(+)SURFACE MEP  (-)SURFACE MEP     TOTAL '
        WRITE(61,1115) BP,BN,BP+BN
        WRITE(61,1116) BP/SCALE/SFACTOR,BN/SCALE/SFACTOR,
     :          (BP+BN)/SCALE/SFACTOR
        WRITE(61,*)
        WRITE(61,*)'POSITIVE POINTS=',MP,'  NEGATIVE POINTS=',MN
        WRITE(61,*)
1113    FORMAT(1X,F12.6,'  ',F12.6,'   ',F12.6,'  (A^2)')
1114    FORMAT(1X,F12.6,'  ',F12.6,'   ',F12.6,'  (A.U.^2)')
1115    FORMAT(1X,F12.6,'  ',F12.6,'   ',F12.6,'  (kCal.A^2)')
1116    FORMAT(1X,F12.6,'  ',F12.6,'   ',F12.6,'  (A.U.^2)')
        WRITE(61,*) 'ATOMIC SURFACE-MEP DESCRIPTORS'
        WRITE(61,*) '(A^2 and kCal/mol)'
        WRITE(61,*)'ATOM','   ',' S+ ',' ',' S- ',' ',
     &     ' STA ',' ',' B+ ',' ',' B- ',' ',
     &     '  BTA ',' POINTS'
        IJ=0
```

```
      DO 65 I=1,NATOM
      SP=0.0
      SN=0.0
      BP=0.0
      BN=0.0
      ITOL=MNAT(I)
      DO 75 J=1,ITOL
      IF (AMEP(IJ+J).LE.0.0) THEN
      SN=SN+AREA(IJ+J)
      BN=BN+AREA(IJ+J)*AMEP(IJ+J)
      ELSE
      SP=SP+AREA(IJ+J)
      BP=BP+AMEP(IJ+J)*AREA(IJ+J)
      ENDIF
75    CONTINUE
      STA=SP+SN
      BTA=BP+BN
      WRITE(61,1117) ATYPE(IAT(IJ+1)),SP,SN,STA,BP,BN,
     &          BTA,MNAT(I)
65    IJ=IJ+ITOL
1117  FORMAT(1X,A4,3F10.5,3F11.5,I8)
      WRITE(61,*)

      WRITE(61,*) '(ATOMIC UNIT (A.U.))'
      WRITE(61,*)'ATOM',' ',' S+ ',' ',' S- ',' ',
     &    ' STA ',' ',' B+ ',' ',' B- ',' ',
     &    ' BTA ',' POINTS'
      IJ=0
      DO 35 I=1,NATOM
      SP=0.0
      SN=0.0
      BP=0.0
      BN=0.0
      ITOL=MNAT(I)
      DO 45 J=1,ITOL
      IF (AMEP(IJ+J).LE.0.0) THEN
      SN=SN+AREA(IJ+J)/SFACTOR
      BN=BN+AREA(IJ+J)*AMEP(IJ+J)/SCALE/SFACTOR
```

```
      ELSE
      SP=SP+AREA(IJ+J)/SFACTOR
      BP=BP+AMEP(IJ+J)*AREA(IJ+J)/SFACTOR/SCALE
      ENDIF
45    CONTINUE
      STA=SP+SN
      BTA=BP+BN
      WRITE(61,1117) ATYPE(IAT(IJ+1)),SP,SN,STA,BP,BN,
     &          BTA,MNAT(I)
35    IJ=IJ+ITOL
      WRITE(61,*)

C     *****
C     PART 3: POLIZTER ANALISIS
C     *****

      WRITE(*,*) '***** PART 3: POLITZER ANALYSIS'
      WRITE(61,*) '***** PART 3: POLITZER ANALYSIS'
      VSM=(VPM+VNM)/NP
      IF (MP.GT.0) VPM=VPM/MP
      IF (MN.GT.0) VNM=VNM/MN
      PAI=0
      SGMP=0.0
      SGMN=0.0
      DO 80 I=1,NP
      PAI=PAI+ABS(AMEP(I)-VSM)
      IF (AMEP(I).GT.0.0) THEN
        SGMP=SGMP+(AMEP(I)-VPM)*(AMEP(I)-VPM)
      ELSE
        SGMN=SGMN+(AMEP(I)-VNM)*(AMEP(I)-VNM)
      ENDIF
80    CONTINUE
      PAI=PAI/NP
      IF (MP.GT.0) SGMP=SGMP/MP
      IF (MN.GT.0) SGMN=SGMN/MN
      SGMT=SGMP+SGMN
      GAMA=(SGMP*SGMN)/(SGMT*SGMT)
      WRITE(61,*)
```

```
      WRITE(61,*)'POLITZER STATISTIC QUANLITIES:'
      WRITE(61,*)
      WRITE(61,1118)
1118  FORMAT(1X,'    PAI    ','Average MEP ',' Average +MEP ',
     :   ' Average -MEP ')
      WRITE(61,1119) PAI,VSM,VPM,VNM
1119  FORMAT(1X,F11.6,F12.6,F13.6,F13.6,' (KCAL/MOL)')
      WRITE(61,1120) PAI/SCALE,VSM/SCALE,VPM/SCALE,VNM/SCALE
1120  FORMAT(1X,F11.6,F12.6,F13.6,F13.6,' (A.U./MOL)')
      WRITE(61,*)
      WRITE(61,1121)
1121  FORMAT(1X,'  +STD. DEV. ',' -STD. DEV. ',' TOTAL STD. DEV. ',
     :   ' BALANCE COEF. ')
      WRITE(61,1122) SGMP,SGMN,SGMT,GAMA
1122  FORMAT(1X,F13.6,F12.6,F17.6,F13.6,' (KCAL/MOL)^2')
      WRITE(61,1123) SGMP/SCALE/SCALE,SGMN/SCALE/SCALE,
     :        SGMT/SCALE/SCALE,GAMA
1123  FORMAT(1X,F13.6,F12.6,F17.6,F13.6,' (A.U./MOL)^2')
      WRITE(61,*)

C     *****
C     PART 4: MOLECULAR LIPOPHILICITY POTENTIAL
C     *****

      WRITE(61,*)
      WRITE(*,*)'***** PART 4: MOLECULAR LIPOPHILICITY (MLP) *****'
      WRITE(61,*)'***** PART 4: MOLECULAR LIPOPHILICITY (MLP) *****'
      WRITE(*,*)'EMPIRICAL CALCULATION'

C     MLP1 (Empirical method)

      DO 90 I=1,NP
      ALP=0.0
      DO 100 J=1,NATOM
      D=((X2(I)-X(J))**2+(Y2(I)-Y(J))**2+(Z2(I)-Z(J))**2)**0.5
C     D=D/FACTOR
      IF (KEY1.EQ.1) ALP=ALP+FINDEX1(J)/(1.+D)
      IF (KEY1.EQ.2) ALP=ALP+FINDEX2(J)/(1.+D)
```

```
100     CONTINUE
        AMLP1(I)=ALP
90      CONTINUE

C       Minimum and Maximum MLP of empirical calculation
        CALL MINIMA (NP,AMLP1,PMI,PMA,IMI,IMA)

        WRITE(61,*)
        IF (KEY1.EQ.1) THEN
        DATASET='(NEW DATA SET, NO UNIT)'
        ELSE
        DATASET='(OLD DATA SET, NO UNIT)'
        ENDIF
        WRITE(61,*)'****** EMPIRICAL MLP1 ',DATASET,' ******'
        WRITE(61,*)
        IJ=IAT(IMI)
        WRITE(61,*)' MIN. MLP1     X       Y       Z   ',
     :          '   ATOM     NO.'
        WRITE(61,1111) PMI,X2(IMI),Y2(IMI),Z2(IMI),ATYPE(IJ),IMI
        WRITE(61,*)
        IJ=IAT(IMA)
        WRITE(61,*)' MAX. MLP1     X       Y       Z   ',
     :          '   ATOM     NO.'
        WRITE(61,1111) PMA,X2(IMA),Y2(IMA),Z2(IMA),ATYPE(IJ),IMA
        WRITE(61,*)

C       Minimum and Maximum MLP on Atoms
        WRITE(61,*)
        WRITE(61,*)'ATOM','      ','ATOMIC  MLP1','    ','MINIMUM',
     &          '      ','MAXIMUM','     ','POINTS'
        DO 130 I=1,NATOM
130     ATOML(I)=0.0

C       ATOML(I): Total Surface-MLP on Atom I
        IJ=0
        DO 150 I=1,NATOM
        SUM=0.0
        ITOL=MNAT(I)
```

```
      DO 140 J=1,ITOL
      SUM=SUM+AMLP1(IJ+J)*AREA(IJ+J)
140   CONTINUE
C     Change the area unit to atomic unit.
      ATOML(I)=SUM/FACTOR/FACTOR
      IJ=IJ+ITOL
150   CONTINUE

      IJ=0
      DO 160 I=1,NATOM
      ITOL=MNAT(I)
      DO 170 J=1,ITOL
170   TEMP(J)=AMLP1(IJ+J)
      CALL MINIMA(ITOL,TEMP,PMI,PMA,IMI,IMA)
      WRITE(61,1112) ATYPE(IAT(IJ+1)),ATOML(I),PMI,PMA,ITOL
      IJ=IJ+ITOL
160   CONTINUE

      PLIP=0.0
      HYD=0.0
      DO 180 I=1,NATOM
      IF (ATOML(I).LE.0) THEN
      HYD=HYD+ATOML(I)
      ELSE
      PLIP=PLIP+ATOML(I)
      ENDIF
180   CONTINUE
      WRITE(61,*)
      WRITE(61,1128) PLIP
      WRITE(61,1129) HYD
1128  FORMAT (1X,'Molecular Lipophilicity Index:  LIP=',F15.8)
1129  FORMAT (1X,'Molecular Hydrophilicity Index: HYD=',F15.8)

      WRITE(*,*)'SEMITHEORETICAL CALCULATION'
C     SEMITHEORETICAL CALCULATION OF LIPOPHILICITY POTENTIAL ON
C     MOLECULAR SURFACE.
C
      WRITE(61,*)
```

```
      WRITE(61,*)'****** SEMITHEORETICAL MLP2 ******'
      DO 200 I=1,NATOM
      ATOMB(I)=0.0
200   ATOMA(I)=0.0

      DO 210 I=1,NP
      IJ=IAT(I)
      AMEP(I)=AMEP(I)/SCALE
      AREA(I)=AREA(I)/SFACTOR
      ATOMB(IJ)=ATOMB(IJ)+AMEP(I)*AREA(I)
      ATOMA(IJ)=ATOMA(IJ)+AREA(I)
210   CONTINUE
      DO 220 I=1,NATOM
      ATOMB(I)=ATOMB(I)
C     ATOMB(I)=ATOMB(I)/ATOMA(I)
220   CONTINUE

      EXPN=2.50
      KEY3=1
      IF (KEY3.EQ.2) GOTO 500
      DO 240 I=1,NP
      SUM=0.0
      DO 250 J=1,NATOM
      IF (IAT(I).NE.J) THEN
      D=((X2(I)-X(J))**2+(Y2(I)-Y(J))**2+(Z2(I)-Z(J))**2)**0.5
      D=D/FACTOR
      D=D**EXPN
      SUM=SUM+ATOMB(J)/D
      ENDIF
250   CONTINUE
      AMLP2(I)=AMEP(I)*SUM
240   CONTINUE
      GOTO 600

500   DO 540 I=1,NP
      SUM=0.0
      DO 550 J=1,NP
      IF (IAT(J).NE.IAT(I)) THEN
```

```
      D=((X2(I)-X2(J))**2+(Y2(I)-Y2(J))**2+(Z2(I)-Z2(J))**2)**0.5
      D=D/FACTOR
      D=D**EXPN
      IF(D.EQ.0.0) write(*,*) i,j
      SUM=SUM+AMEP(J)*AREA(J)/D
      ENDIF
550   CONTINUE
      AMLP2(I)=AMEP(I)*SUM
540   CONTINUE

600   WRITE(61,*)
      CALL MINIMA (NP,AMLP2,PMI,PMA,IMI,IMA)
      IJ=IAT(IMI)
      WRITE(61,*)' MIN. MLP2    X       Y       Z   ',
     :      '  ATOM     NO.'
      WRITE(61,1111) PMI,X2(IMI),Y2(IMI),Z2(IMI),ATYPE(IJ),IMI
      WRITE(61,*)
      IJ=IAT(IMA)
      WRITE(61,*)' MAX. MLP2    X       Y       Z   ',
     :      '  ATOM     NO.'
      WRITE(61,1111) PMA,X2(IMA),Y2(IMA),Z2(IMA),ATYPE(IJ),IMA

      IJ=0
      DO 280 I=1,NATOM
      SUM=0.0
      ITOL=MNAT(I)
      DO 260 J=1,ITOL
      SUM=SUM+AMLP2(IJ+J)*AREA(IJ+J)
260   CONTINUE
C     Change the area unit to atomic unit.
      ATOML(I)=SUM
      IJ=IJ+ITOL
280   CONTINUE

      WRITE(61,*)
      WRITE(61,*)'ATOM',' ','MLP INDEX',' ','MINIMUM',
     &      '  ','MAXIMUM',' ','POINTS'
      IJ=0
```

```
      DO 300 I=1,NATOM
      ITOL=MNAT(I)
      DO 310 J=1,ITOL
310   TEMP(J)=AMLP2(IJ+J)
      CALL MINIMA(ITOL,TEMP,PMI,PMA,IMI,IMA)
      WRITE(61,1112) ATYPE(IAT(IJ+1)),ATOML(I),PMI,PMA,ITOL
      IJ=IJ+ITOL
300   CONTINUE

      PLIP=0.0
      HYD=0.0
      DO 350 I=1,NATOM
      IF (ATOML(I).LE.0) THEN
      HYD=HYD+ATOML(I)
      ELSE
      PLIP=PLIP+ATOML(I)
      ENDIF
350   CONTINUE
      WRITE(61,*)
      WRITE(61,1128) PLIP
      WRITE(61,1129) HYD
      WRITE(61,*)

      WRITE(62,*) '    X       Y       Z       MEP ',
     &        '  MLP1     MLP2     AREA  No.'
      DO 400 I=1,NP
400   WRITE(62,4444) X2(I),Y2(I),Z2(I),AMEP(I),AMLP1(I),AMLP2(I),
     &        AREA(I),IAT(I)
4444  FORMAT(1X,3F11.6,2F10.6,F11.6,F10.6,I4)
      CLOSE(UNIT=61)
      CLOSE(UNIT=62)
      WRITE(*,*) 'PROGRAM TERMINATES NORMALLY'
      END

      FUNCTION FINDEX2(J)
C     *****
C     OLD DATA SET (1986)
C     *****
```

```
      CHARACTER*4 ATYPE(1000)
      COMMON ATYPE

C     C IN: CH3R,CH4 (R: ANY GROUP LINKED THROUGH CARBON)
      IF (ATYPE(J).EQ.'C001') f=-0.6327
C     C IN: CH2R2
      IF (ATYPE(J).EQ.'C002') f=-0.3998
C     C IN: CHR3
      IF (ATYPE(J).EQ.'C003') f=-0.2793
C     C IN: CR4
      IF (ATYPE(J).EQ.'C004') f=0.2202
C     C IN: CH3X (X: O,S,N, AND HALOGEN)
      IF (ATYPE(J).EQ.'C005') f=-1.1461
C     C IN: CH2RX
      IF (ATYPE(J).EQ.'C006') f=-0.9481
C     C IN: CH2X2
      IF (ATYPE(J).EQ.'C007') f=0.2394
C     C IN: CHR2X
      IF (ATYPE(J).EQ.'C008') f=-0.9463
C     C IN: CHRX2
      IF (ATYPE(J).EQ.'C009') f=0.5822
C     C IN: CHX3
      IF (ATYPE(J).EQ.'C010') f=0.7245
C     C IN: CR3X
      IF (ATYPE(J).EQ.'C011') f=-1.0777
C     C IN: CR2X2
      IF (ATYPE(J).EQ.'C012') f=1.1220
C     C IN: CRX3
      IF (ATYPE(J).EQ.'C013') f=0.6278
C     C IN: CX4
      IF (ATYPE(J).EQ.'C014') f=1.2558
C     C IN: =CH2 (=: DOUBLE BOND)
      IF (ATYPE(J).EQ.'C015') f=-0.2633
C     C IN: =CHR
      IF (ATYPE(J).EQ.'C016') f=-0.0460
C     C IN:=CR2
      IF (ATYPE(J).EQ.'C017') f=0.3496
C     C IN: =CHX
```

```
      IF (ATYPE(J).EQ.'C018') f=-0.3053
C     C IN: =CRX
      IF (ATYPE(J).EQ.'C019') f=-0.4451
C     C IN: =CX2
      IF (ATYPE(J).EQ.'C020') f=-0.1915
C     C IN: %CH (%: TRIPLE BOND)
      IF (ATYPE(J).EQ.'C021') f=0.1785
C     C IN: %CR,R=C=R
      IF (ATYPE(J).EQ.'C022') f=0.1541
C     C IN: R--CH--R (--: AROMATIC BONDS AS IN BENZENE OR
C         DELOCALIZED BOND AS THE N-O BOND IN NITRO GROUP)
      IF (ATYPE(J).EQ.'C024') f=-0.0548
C     C IN: R--CR--R
      IF (ATYPE(J).EQ.'C025') f=0.3345
C     C IN: R--CX--R
      IF (ATYPE(J).EQ.'C026') f=-0.1153
C     C IN: R--CH--X
      IF (ATYPE(J).EQ.'C027') f=0.0219
C     C IN: R--CR--X
      IF (ATYPE(J).EQ.'C028') f=0.2093
C     C IN: R--CX--X
      IF (ATYPE(J).EQ.'C029') f=-0.1378
C     C IN: X--CH--X
      IF (ATYPE(J).EQ.'C030') f=-0.2686
C     C IN: X--CR--X
      IF (ATYPE(J).EQ.'C031') f=0.7376
C     C IN: X--CX--X
      IF (ATYPE(J).EQ.'C032') f=0.0339
C     C IN: R--CH...X (...: AROMATIC "SINGLE" BOND AS THE
C         C-N BOND IN PYRROLE)
      IF (ATYPE(J).EQ.'C033') f=0.0230
C     C IN: R--CR...X
      IF (ATYPE(J).EQ.'C034') f=0.2455
C     C IN: R--CX...X
      IF (ATYPE(J).EQ.'C035') f=-0.1883
C     C IN: AL-CH=X (AL: ALIPHATIC GROUP)
      IF (ATYPE(J).EQ.'C036') f=0.7853
C     C IN: AR-CH=X (AR: AROMATIC GROUP)
```

```
      IF (ATYPE(J).EQ.'C037') f=0.1682
C     C IN: AL-C(=X)-AL
      IF (ATYPE(J).EQ.'C038') f=-0.4349
C     C IN: AR-C(=X)-R
      IF (ATYPE(J).EQ.'C039') f=-0.2392
C     C IN: R-C(=X)-X, R-C%X, X=C=X
      IF (ATYPE(J).EQ.'C040') f=-0.1703
C     C IN: X-C(=X)-X
      IF (ATYPE(J).EQ.'C041') f=0.0340
C     C IN: X--CH...X
      IF (ATYPE(J).EQ.'C042') f=-0.7231
C     C IN: X--CR...X
      IF (ATYPE(J).EQ.'C043') f=0.2256
C     C IN: X--CX...X
      IF (ATYPE(J).EQ.'C044') f=-0.2692

C     THE FIRST NUMBER REPRESENTS HYBRIDIZATION AND THE SECOND
C     ITS FORMAL OXIDATION NUMBER
C     H ATTACHED TO: C(SP3,0)
      IF (ATYPE(J).EQ.'H046') f=0.4307
C     H ATTACHED TO: C(SP3,1), C(SP2,0)
      IF (ATYPE(J).EQ.'H047') f=0.3722
C     H ATTACHED TO: C(SP3,2), C(SP2,1), C(SP,0)
      IF (ATYPE(J).EQ.'H048') f=0.0065
C     H ATTACHED TO: C(SP3,3), C(SP2,2), C(SP2,3), C(SP,3)
      IF (ATYPE(J).EQ.'H049') f=-0.2232
C     H ATTACHED TO: HETEROATOM
      IF (ATYPE(J).EQ.'H050') f=-0.3703
C     H ATTACHED TO: ALPHA-C
      IF (ATYPE(J).EQ.'H051') f=0.2421

C     O IN: ALCOHOL
      IF (ATYPE(J).EQ.'O056') f=-0.0517
C     O IN: PHENOL, ENOL, CARBOXYL OH
      IF (ATYPE(J).EQ.'O057') f=0.5212
C     O IN: =O
      IF (ATYPE(J).EQ.'O058') f=-0.1729
C     O IN: AL-O-AL
```

```
      IF (ATYPE(J).EQ.'O059') f=0.0407
C     O IN: AL-O-AR, AR2O, R...O...R, R-O-C=X
      IF (ATYPE(J).EQ.'O060') f=0.3410
C     O IN: --O (AS IN NITRO, =N-OXIDES)
      IF (ATYPE(J).EQ.'O061') f=1.8020

C     N IN: AL-NH2
      IF (ATYPE(J).EQ.'N066') f=0.2658
C     N IN: AL2NH
      IF (ATYPE(J).EQ.'N067') f=0.2817
C     N IN: AL3N
      IF (ATYPE(J).EQ.'N068') f=0.3990
C     N IN: AR-NH2, X-NH2
      IF (ATYPE(J).EQ.'N069') f=0.4442
C     N IN: AR-NH-AL
      IF (ATYPE(J).EQ.'N070') f=1.0841

C     N IN: AR-NAL2
      IF (ATYPE(J).EQ.'N071') f=0.6632
C     N IN: RCO-N<, >N-X=X
      IF (ATYPE(J).EQ.'N072') f=0.1414
C     N IN: AR2NH, AR3N, AR2N-AL, R...N...R (PYRROLE TYPE STRUCTURE)
      IF (ATYPE(J).EQ.'N073') f=0.3493
C     N IN: R%N, R=N-
      IF (ATYPE(J).EQ.'N074') f=-0.1201
C     N IN: R--N--R (PYRIDINE TYPE STRUCTURE), R--N--X
      IF (ATYPE(J).EQ.'N075') f=0.1757
C     N IN: AR-NO2, R--N(--R)--O (PYRIDINE-N-OXIDE TYPE), RO-NO2
      IF (ATYPE(J).EQ.'N076') f=-3.1516
C     N IN: AL-NO2
      IF (ATYPE(J).EQ.'N077') f=-3.3332
C     N IN: AR-N=X, X-N=X
      IF (ATYPE(J).EQ.'N078') f=0.1709

C     F ATTACHED TO: C(SP3,1)
      IF (ATYPE(J).EQ.'F081') f=0.4649
C     F ATTACHED TO: C(SP3,2)
      IF (ATYPE(J).EQ.'F082') f=-0.1701
```

```
C       F ATTACHED TO: C(SP3,3)
        IF (ATYPE(J).EQ.'F083') f=0.1172
C       F ATTACHED TO: C(SP2,1)
        IF (ATYPE(J).EQ.'F084') f=0.6035
C       F ATTACHED TO: C(SP2,2-4), C(SP,1), C(SP,4), X
        IF (ATYPE(J).EQ.'F085') f=0.4752

C       CL ATTACHED TO: C(SP3,1)
        IF (ATYPE(J).EQ.'CL86') f=1.0723
C       CL ATTACHED TO: C(SP3,2)
        IF (ATYPE(J).EQ.'CL87') f=0.3027
C       CL ATTACHED TO: C(SP3,3)
        IF (ATYPE(J).EQ.'CL88') f=0.4108
C       CL ATTACHED TO: C(SP2,1)
        IF (ATYPE(J).EQ.'CL89') f=1.0278
C       CL ATTACHED TO: C(SP2,2-4), C(SP,1), C(SP,4), X
        IF (ATYPE(J).EQ.'CL90') f=0.6972

C       BR ATTACHED TO: C(SP3,1)
        IF (ATYPE(J).EQ.'BR91') f=1.0966
C       BR ATTACHED TO: C(SP3,2)
        IF (ATYPE(J).EQ.'BR92') f=0.4292
C       BR ATTACHED TO: C(SP2,1)
        IF (ATYPE(J).EQ.'BR94') f=1.3224
C       BR ATTACHED TO: C(SP2,2-4), C(SP,1), C(SP,4), X
        IF (ATYPE(J).EQ.'BR95') f=0.9987

C       I ATTACHED TO: C(SP3,1)
        IF (ATYPE(J).EQ.'I096') f=1.4334
C       I ATTACHED TO: C(SP2,1)
        IF (ATYPE(J).EQ.'I099') f=1.8282
C       I ATTACHED TO; C(SP2,2-4), C(SP,1), C(SP,4), X
        IF (ATYPE(J).EQ.'I100') f=1.0735

C       S IN: R-SH
        IF (ATYPE(J).EQ.'S106') f=1.0152
C       S IN: R2S, RS-SR
        IF (ATYPE(J).EQ.'S107') f=1.0339
```

```
C       S IN: R=S
        IF (ATYPE(J).EQ.'S108') f=0.0727
C       S IN: R-SO-R
        IF (ATYPE(J).EQ.'S109') f=-0.3332
C       S IN: R-SO2-R
        IF (ATYPE(J).EQ.'S110') f=-0.1005
        FINDEX2=f

        RETURN
        END

        FUNCTION FINDEX1(K)
C       *****
C       NEW DATA SET (1989)
C       *****
        REAL FINDEX
        COMMON ATYPE
        CHARACTER*4 ATYPE(1000)
C       "THE LIPOPHYLIC ATOMIC SYMBOL CORRESPONDING TO EACH ATOM C
        ACCORDING
C       THE PAPER OF VISWANADHAN, GHOSE, REVANKAR AND ROBINS J. CHEM. INF.
C       COMPUT SCI. VOL.29 163-172 (1989)"
C
C       C IN: CH3R,CH4 (R: ANY GROUP LINKED THROUGH CARBON)
        IF (ATYPE(K).EQ.'C001') FINDEX=-0.6771
C       C IN: CH2R2
        IF (ATYPE(K).EQ.'C002') FINDEX=-0.4873
C       C IN: CHR3
        IF (ATYPE(K).EQ.'C003') FINDEX=-0.3633
C       C IN: CR4
        IF (ATYPE(K).EQ.'C004') FINDEX=-0.1366
C       C IN: CH3X (X: O,S,N, AND HALOGEN)
        IF (ATYPE(K).EQ.'C005') FINDEX=-1.0824
C       C IN: CH2RX
        IF (ATYPE(K).EQ.'C006') FINDEX=-0.8370
C       C IN: CH2X2
        IF (ATYPE(K).EQ.'C007') FINDEX=-0.6015
C       C IN: CHR2X
```

```
      IF (ATYPE(K).EQ.'C008') FINDEX=-0.5210
C     C IN: CHRX2
      IF (ATYPE(K).EQ.'C009') FINDEX=-0.4042
C     C IN: CHX3
      IF (ATYPE(K).EQ.'C010') FINDEX=0.3651
C     C IN: CR3X
      IF (ATYPE(K).EQ.'C011') FINDEX=-0.5399
C     C IN: CR2X2
      IF (ATYPE(K).EQ.'C012') FINDEX=0.4011
C     C IN: CRX3
      IF (ATYPE(K).EQ.'C013') FINDEX=0.2263
C     C IN: CX4
      IF (ATYPE(K).EQ.'C014') FINDEX=0.8282
C     C IN: =CH2 (=: DOUBLE BOND)
      IF (ATYPE(K).EQ.'C015') FINDEX=-0.1053
C     C IN: =CHR
      IF (ATYPE(K).EQ.'C016') FINDEX=-0.0681
C     C IN:=CR2
      IF (ATYPE(K).EQ.'C017') FINDEX=-0.2287
C     C IN: =CHX
      IF (ATYPE(K).EQ.'C018') FINDEX=-0.3665
C     C IN: =CRX
      IF (ATYPE(K).EQ.'C019') FINDEX=-0.9188
C     C IN: =CX2
      IF (ATYPE(K).EQ.'C020') FINDEX=-0.0082
C     C IN: %CH (%: TRIPLE BOND)
      IF (ATYPE(K).EQ.'C021') FINDEX=-0.1047
C     C IN: %CR,R=C=R
      IF (ATYPE(K).EQ.'C022') FINDEX=0.1513
C     C IN: R--CH--R (--: AROMATIC BONDS AS IN BENZENE OR
C         DELOCALIZED BOND AS THE N-O BOND IN NITRO GROUP)
      IF (ATYPE(K).EQ.'C024') FINDEX=0.0068
C     C IN: R--CR--R
      IF (ATYPE(K).EQ.'C025') FINDEX=0.1600
C     C IN: R--CX--R
      IF (ATYPE(K).EQ.'C026') FINDEX=-0.1033
C     C IN: R--CH--X
      IF (ATYPE(K).EQ.'C027') FINDEX=0.0598
```

```
C       C IN: R--CR--X
        IF (ATYPE(K).EQ.'C028') FINDEX=0.1290
C       C IN: R--CX--X
        IF (ATYPE(K).EQ.'C029') FINDEX=0.1652
C       C IN: X--CH--X
        IF (ATYPE(K).EQ.'C030') FINDEX=0.2975
C       C IN: X--CR--X

      IF (ATYPE(K).EQ.'C031') FINDEX=0.9421
C       C IN: X--CX--X
        IF (ATYPE(K).EQ.'C032') FINDEX=0.2074
C       C IN: R--CH...X (...: AROMATIC "SINGLE" BOND AS THE
C           C-N BOND IN PYRROLE)
        IF (ATYPE(K).EQ.'C033') FINDEX=-0.1774
C       C IN: R--CR...X
        IF (ATYPE(K).EQ.'C034') FINDEX=-0.2782
C       C IN: R--CX...X
        IF (ATYPE(K).EQ.'C035') FINDEX=-0.3630
C       C IN: AL-CH=X (AL: ALIPHATIC GROUP)
        IF (ATYPE(K).EQ.'C036') FINDEX=-0.0321
C       C IN: AR-CH=X (AR: AROMATIC GROUP)
        IF (ATYPE(K).EQ.'C037') FINDEX=0.3568
C       C IN: AL-C(=X)-AL
        IF (ATYPE(K).EQ.'C038') FINDEX=0.8255
C       C IN: AR-C(=X)-R
        IF (ATYPE(K).EQ.'C039') FINDEX=-0.1116
C       C IN: R-C(=X)-X, R-C%X, X=C=X
        IF (ATYPE(K).EQ.'C040') FINDEX=0.0709
C       C IN: X-C(=X)-X
        IF (ATYPE(K).EQ.'C041') FINDEX=0.4571
C       C IN: X--CH...X
        IF (ATYPE(K).EQ.'C042') FINDEX=-0.1316
C       C IN: X--CR...X
        IF (ATYPE(K).EQ.'C043') FINDEX=0.0498
C       C IN: X--CX...X
        IF (ATYPE(K).EQ.'C044') FINDEX=0.1847

C       THE FIRST NUMBER REPRESENTS HYBRIDIZATION AND THE SECOND
```

```
C       ITS FORMAL OXIDATION NUMBER. THE FORMAL OXIDATION NUMBER OF A
C       CARBON ATOM=SUM FORMAL BOND ORDERS WITH ELECTRONEGATIVE C
        ATOMS.
C       H ATTACHED TO: C(SP3,0) HAVING NO X ATTACHED TO NEXT C
        IF (ATYPE(K).EQ.'H046') FINDEX=0.4418
C       H ATTACHED TO: C(SP3,1), C(SP2,0)
        IF (ATYPE(K).EQ.'H047') FINDEX=0.3343
C       H ATTACHED TO: C(SP3,2), C(SP2,1), C(SP,0)
        IF (ATYPE(K).EQ.'H048') FINDEX=0.3161
C       H ATTACHED TO: C(SP3,3), C(SP2,2), C(SP2,3), C(SP,3)
        IF (ATYPE(K).EQ.'H049') FINDEX=-0.1488
C       H ATTACHED TO: HETEROATOM
        IF (ATYPE(K).EQ.'H050') FINDEX=-0.3260
C       H ATTACHED TO: ALPHA-C (MAY BE DEFINED AS A C ATTACHED
C       THROUGH A SINGLE BOND WITH -C=X, -C%X, -C--X
        IF (ATYPE(K).EQ.'H051') FINDEX=0.2099
C       H ATTACHED TO: C(SP3,0) HAVING 1 X ATTACHED TO NEXT C
        IF (ATYPE(K).EQ.'H052') FINDEX=0.3695
C       H ATTACHED TO: C(SP3,0) HAVING 2 X ATTACHED TO NEXT C
        IF (ATYPE(K).EQ.'H053') FINDEX=0.2697
C       H ATTACHED TO: C(SP3,0) HAVING 3 X ATTACHED TO NEXT C
        IF (ATYPE(K).EQ.'H054') FINDEX=0.3647

C       O IN: ALCOHOL
        IF (ATYPE(K).EQ.'O056') FINDEX=0.1402

C       O IN: PHENOL, ENOL, CARBOXYL OH
        IF (ATYPE(K).EQ.'O057') FINDEX=0.4860
C       O IN: =O
        IF (ATYPE(K).EQ.'O058') FINDEX=-0.3514
C       O IN: AL-O-AL
        IF (ATYPE(K).EQ.'O059') FINDEX=0.1720
C       O IN: AL-O-AR, AR2O, R...O...R, R-O-C=X
        IF (ATYPE(K).EQ.'O060') FINDEX=0.2712
C       O IN: --O (AS IN NITRO, =N-OXIDES)
        IF (ATYPE(K).EQ.'O061') FINDEX=1.5810

C       Se IN: ANY-SE-ANY
```

```
      IF (ATYPE(K).EQ.'SE64') FINDEX=0.1473

C     N IN: AL-NH2
      IF (ATYPE(K).EQ.'N066') FINDEX=0.1187
C     N IN: AL2NH
      IF (ATYPE(K).EQ.'N067') FINDEX=0.2805
C     N IN: AL3N
      IF (ATYPE(K).EQ.'N068') FINDEX=0.3954
C     N IN: AR-NH2, X-NH2
      IF (ATYPE(K).EQ.'N069') FINDEX=-0.3132
C     N IN: AR-NH-AL
      IF (ATYPE(K).EQ.'N070') FINDEX=0.4238
C     N IN: AR-NAL2
      IF (ATYPE(K).EQ.'N071') FINDEX=0.8678
C     N IN: RCO-N<, >N-X=X
      IF (ATYPE(K).EQ.'N072') FINDEX=-0.0528
C     N IN: AR2NH, AR3N, AR2N-AL, R...N...R (PYRROLE TYPE STRUCTURE)
      IF (ATYPE(K).EQ.'N073') FINDEX=0.4198
C     N IN: R%N, R=N-
      IF (ATYPE(K).EQ.'N074') FINDEX=0.1461
C     N IN: R--N--R (PYRIDINE TYPE STRUCTURE), R--N--X
      IF (ATYPE(K).EQ.'N075') FINDEX=-0.1106
C     N IN: AR-NO2, R--N(--R)--O (PYRIDINE-N-OXIDE TYPE), RO-NO2
      IF (ATYPE(K).EQ.'N076') FINDEX=-2.7640
C     N IN: AL-NO2
      IF (ATYPE(K).EQ.'N077') FINDEX=-2.7919
C     N IN: AR-N=X, X-N=X
      IF (ATYPE(K).EQ.'N078') FINDEX=0.5721

C     F ATTACHED TO: C(SP3,1)
      IF (ATYPE(K).EQ.'F081') FINDEX=0.4174
C     F ATTACHED TO: C(SP3,2)
      IF (ATYPE(K).EQ.'F082') FINDEX=0.2167
C     F ATTACHED TO: C(SP3,3)
      IF (ATYPE(K).EQ.'F083') FINDEX=0.2792
C     F ATTACHED TO: C(SP2,1)
      IF (ATYPE(K).EQ.'F084') FINDEX=0.5839
C     F ATTACHED TO: C(SP2,2-4), C(SP,1), C(SP,4), X
```

```
      IF (ATYPE(K).EQ.'F085') FINDEX=0.3425

C     CL ATTACHED TO: C(SP3,1)
      IF (ATYPE(K).EQ.'CL86') FINDEX=0.9609
C     CL ATTACHED TO: C(SP3,2)
      IF (ATYPE(K).EQ.'CL87') FINDEX=0.5594
C     CL ATTACHED TO: C(SP3,3)
      IF (ATYPE(K).EQ.'CL88') FINDEX=0.4656
C     CL ATTACHED TO: C(SP2,1)
      IF (ATYPE(K).EQ.'CL89') FINDEX=0.9624
C     CL ATTACHED TO: C(SP2,2-4), C(SP,1), C(SP,4), X
      IF (ATYPE(K).EQ.'CL90') FINDEX=0.6345

C     BR ATTACHED TO: C(SP3,1)
      IF (ATYPE(K).EQ.'BR91') FINDEX=1.0242
C     BR ATTACHED TO: C(SP3,2)
      IF (ATYPE(K).EQ.'BR92') FINDEX=0.4374
C     BR ATTACHED TO: C(SP3,3)
      IF (ATYPE(K).EQ.'BR93') FINDEX=0.4332
C     BR ATTACHED TO: C(SP2,1)
      IF (ATYPE(K).EQ.'BR94') FINDEX=1.2362
C     BR ATTACHED TO: C(SP2,2-4), C(SP,1), C(SP,4), X
      IF (ATYPE(K).EQ.'BR95') FINDEX=0.9351

C     I ATTACHED TO: C(SP3,1)
      IF (ATYPE(K).EQ.'I096') FINDEX=1.4350
C     I ATTACHED TO: C(SP2,1)
      IF (ATYPE(K).EQ.'I099') FINDEX=1.7018
C     I ATTACHED TO; C(SP2,2-4), C(SP,1), C(SP,4), X
      IF (ATYPE(K).EQ.'I100') FINDEX=0.9336

C     S IN: R-SH
      IF (ATYPE(K).EQ.'S106') FINDEX=0.7268
C     S IN: R2S, RS-SR
      IF (ATYPE(K).EQ.'S107') FINDEX=0.6145
C     S IN: R=S
      IF (ATYPE(K).EQ.'S108') FINDEX=0.3828
C     S IN: R-SO-R
```

```
      IF (ATYPE(K).EQ.'S109') FINDEX=-0.1708
C     S IN: R-SO2-R
      IF (ATYPE(K).EQ.'S110') FINDEX=0.3717

C     P IN: R3-P=X
      IF (ATYPE(K).EQ.'P116') FINDEX=-1.6251
C     P IN: X3-P=X (PHOSPHATE)
      IF (ATYPE(K).EQ.'P117') FINDEX=0.3308
C     P IN: C-P(X)2=X (PHOSPHONATE)
      IF (ATYPE(K).EQ.'P120') FINDEX=0.0236
      FINDEX1=FINDEX

      RETURN
      END

      SUBROUTINE MINIMA(N,X,XMI,XMA,IMI,IMA)
      DIMENSION X(8000)
      IF(X(1).LT.X(2))GO TO 1
      XMA=X(1)
      IMA=1
      XMI=X(2)
      IMI=2
      GO TO 2
1     XMI=X(1)
      IMI=1
      XMA=X(2)
      IMA=2
2     DO 3 J=3,N
      IF(X(J).GT.XMI) GOTO 10
      XMI=X(J)
      IMI=J
      GO TO 3
10    IF(X(J).LT.XMA) GO TO 3
      XMA=X(J)
      IMA=J
3     CONTINUE
      RETURN
      END
```

2. MEPG92.FOR

```
C       F77 PROGRAM: MEPG92.FOR              September 20, 1995
C       U. of S., Qishi Du
C
C       To make a series of command files for G92 to calculate MEP. It needs two
C       input files.
C        1) AA.DAT: contains the command lines for GAUSS 92.
C        2) MSNEW.ALV: Contains the molecular surface (coordinates
C                 of points on the surface X  Y  Z) produced
C                 by MSNEW program.
C
C       Output File: MEPx.COM (x=0,1,2.....)
C       A series of command file for G92 to calculate MEP on a
C       molecular surface.

        CHARACTER*78 INPUT1,INPUT2,OUTPUT
        CHARACTER*72 LINE1,LINE2
        CHARACTER*8  MEP(20)

        DATA MEP/'MEP0.COM','MEP1.COM','MEP2.COM','MEP3.COM','MEP4.COM',
     1          'MEP5.COM','MEP6.COM','MEP7.COM','MEP8.COM','MEP9.COM',
     2          'MEPA.COM','MEPB.COM','MEPC.COM','MEPD.COM','MEPE.COM',
     3          'MEPF.COM','MEPG.COM','MEPH.COM','MEPI.COM','MEPJ.COM'/

C       READ(5,*) INPUT1
C       READ(5,*) INPUT2
C       READ(5,*) OUTPUT
C       WRITE(*,*) INPUT1,' ',INPUT2,' ',OUTPUT

        OPEN(UNIT=31,FILE='AA.DAT',STATUS='OLD')
        OPEN(UNIT=32,FILE='MSNEW.ALV',STATUS='OLD')
        OPEN(UNIT=60,FILE=MEP(1),STATUS='UNKNOWN')

        I=0
        MN=0
300     CONTINUE
        I=I+1
```

```
200     READ(31,111,END=100) LINE1
        IF (I.EQ.1) WRITE(*,111) LINE1
        WRITE(60,111) LINE1
111     FORMAT(A72)
        GOTO 200
100     CONTINUE

        WRITE(*,*)'I=',I,'   ',MEP(I)
        J=0
500     READ(32,*,END=400) X,Y,Z,AREA,IATOM
        WRITE(60,*) X,Y,Z
        MN=MN+1
        J=J+1
        IF (J.EQ.900) THEN
                WRITE(60,*) ' '
                CLOSE(UNIT=31)
                CLOSE(UNIT=60)
                OPEN(UNIT=31,FILE='AA.DAT',STATUS='OLD')
                OPEN(UNIT=60,FILE=MEP(I+1),STATUS='UNKNOWN')
                GO TO 300
        ENDIF
        GOTO 500
400     CONTINUE
        WRITE(*,*) 'TOTAL LINE: MN=',MN
        CLOSE(UNIT=31)
        CLOSE(UNIT=32)
        CLOSE(UNIT=60)
        END
```

3. **CUT.FOR**

```
C       F77 PROGRAM: CUT.FOR                September 20, 1995
C       U. of S., Qishi Du
C       Changed at Jan. 1997.
C       The output file MEPXYZ.DAT is changed to the unformatted file in order
C       to save disk space.
C
C       To make a DATA file from MEPx .log for program MEPMLP.FOR. It needs
C       three input files.
C       1). The first input file is a series of output files of Gaussian 92: MEPx.log.
C       2). The second input file contains the information of the position to be cut
C          off: LINE.DAT.
C       3). The third input file holds all data on the molecular surface: MSNEW.ALV.
C          It contains coordinates of grid points on the molecular surface, which is
C          produced by MSNEW program.
C
C       Program produces a data file: MEPXYZ.DAT, containing all MEP values on the
C       molecular surface, and coordinates of grid points on the molecular surface.
C
C        INPUT FILE1: MEP0.log, MEP1.log, MEP2.log.....(Output .log
C             files of GAUSS 92)
C        INPUT FILE2: LINE.DAT (Contains information for cut off)
C        INPUT FILE3: MSNEW.ALV (free format)
C             X(I)  Y(I)  Z(I)  AREA(I)  IATOM(I)
C        Temporary FILE: MEP.DAT (a work file)
C
C        OUTPUT: MLPXYZ.DAT (free format)
C        After Dec 1996, it was changed into an unformatted file.
C             X(I)  Y(I)  Z(I)  MEP(I)  AREA(I)  IATOM(I)

      CHARACTER*78 LINE1,LINE2,LINE3
      CHARACTER*8  MEP(20)
      DIMENSION X(8000),Y(8000),Z(8000),AMEP(8000),AREA(8000),
     1     IATOM(8000)

      DATA MEP/'MEP0.log','MEP1.log','MEP2.log','MEP3.log','MEP4.log',
     1     'MEP5.log','MEP6.log','MEP7.log','MEP8.log','MEP9.log',
```

```
     2       'MEPA.log','MEPB.log','MEPC.log','MEPD.log','MEPE.log',
     3       'MEPF.log','MEPG.log','MEPH.log','MEPI.log','MEPJ.log'/

C         READ(5,*) INPUT1
C         READ(5,*) INPUT2
C         READ(5,*) OUTPUT
C         WRITE(*,*) INPUT1,'  ',INPUT2,'  ',OUTPUT

          OPEN(UNIT=31,FILE='LINE.DAT',STATUS='OLD')
          OPEN(UNIT=32,FILE=MEP(1),STATUS='OLD')
          OPEN(UNIT=60,FILE='MEP.DAT',STATUS='UNKNOWN')

          READ(31,111) LINE1
          WRITE(*,111) LINE1
          READ(31,111) LINE2
          WRITE(*,111) LINE2
          CLOSE(UNIT=31)

          MN=0
          I=1
300       CONTINUE
          WRITE(*,*) 'I=',I,'   ',MEP(I)
200       READ(32,111,END=1000) LINE3
C         WRITE(*,111) LINE3
          IF (LINE3.NE.LINE1) GOTO 200
111       FORMAT(A78)

          DO 100 J=1,900
          READ(32,111,END=1000) LINE3
C         WRITE(*,*) LINE3
          IF (LINE3.EQ.LINE2) GOTO 1000
          MN=MN+1
          WRITE(60,111) LINE3
100       CONTINUE

          CLOSE(UNIT=32)
          I=I+1
          OPEN(UNIT=32,FILE=MEP(I),STATUS='OLD')
```

```
      GOTO 300

1000  CONTINUE
      WRITE(*,*) 'TOTAL LINE: MN=',MN
      CLOSE(UNIT=32)
      CLOSE(UNIT=60)

      OPEN(UNIT=31,FILE='MEP.DAT',STATUS='OLD')
      OPEN(UNIT=32,FILE='MSNEW.ALV',STATUS='OLD')
      ISUM=0
      IJ=1
      DO 400 I=1,MN
      READ(31,*) JUNK,AMEP(I)
      READ(32,*) X(I),Y(I),Z(I),AREA(I),IATOM(I)
      IF (IATOM(I).EQ.IJ) THEN
            ISUM=ISUM+1
      ELSE
            WRITE(*,*) 'ATOM ',IJ,ISUM
            IJ=IJ+1
            ISUM=1
      ENDIF
400   CONTINUE
      WRITE(*,*) 'ATOM ',IJ,ISUM
C     WRITE(*,333) I,X(I),Y(I),Z(I),AMEP(I),AREA(I),IATOM(I)
333   FORMAT(1X,I5,5F12.6,I5)
      CLOSE(UNIT=31)
      CLOSE(UNIT=32)

      OPEN(UNIT=60,FILE='MEPXYZ.DAT',STATUS='UNKNOWN',
     :  FORM='UNFORMATTED')
      DO 600 I=1,MN
600   WRITE(60) X(I),Y(I),Z(I),AMEP(I),AREA(I),IATOM(I)
      CLOSE(UNIT=60)

      END
```

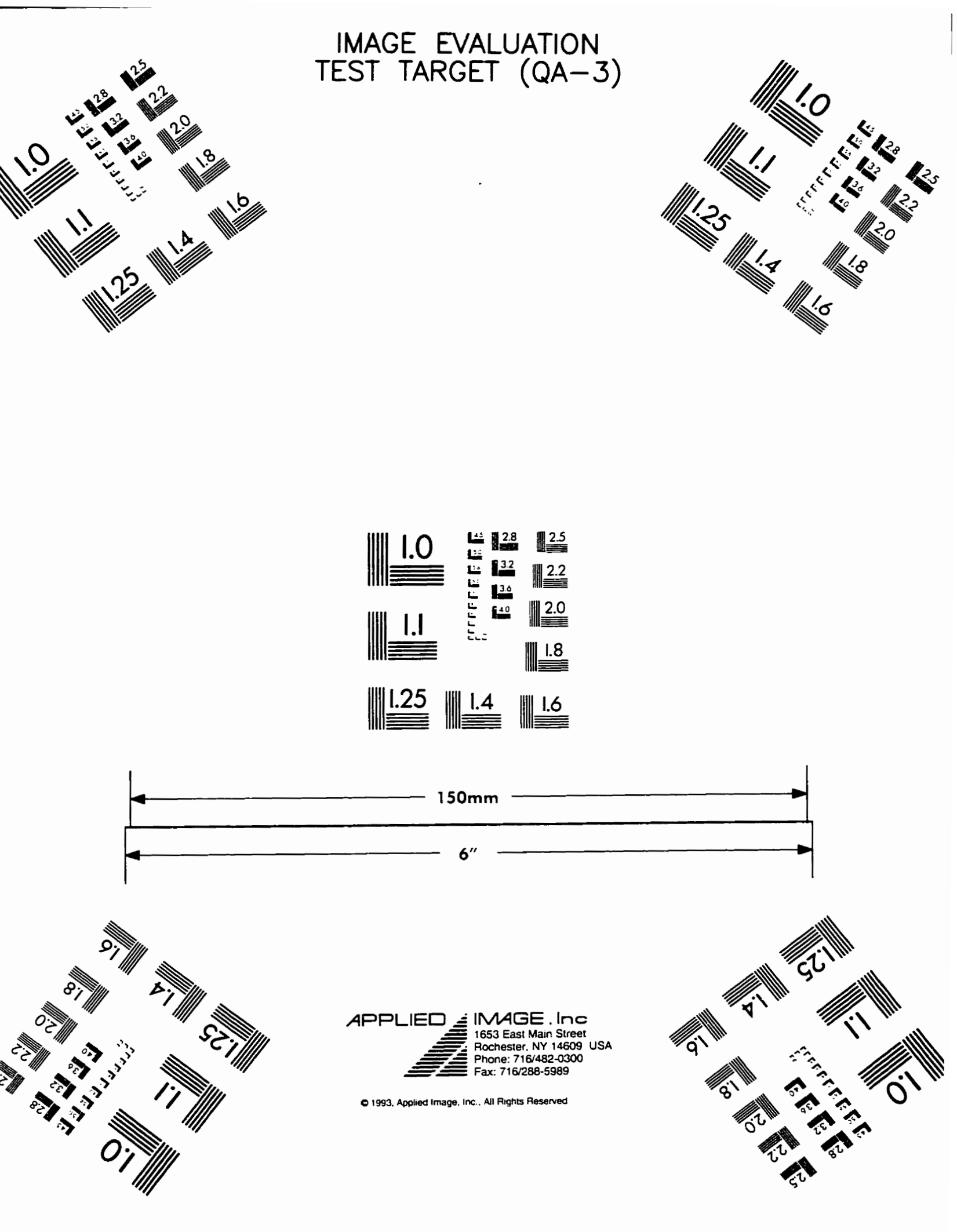


IMAGE EVALUATION
TEST TARGET (QA-3)
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
150mm
6"
APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989
© 1993, Applied Image, Inc., All Rights Reserved