PREDICTION OF SOLUBILITY OF AMINO ACIDS BASED ON COSMO CALCULAITION

by

Kaiyu Li

A dissertation submitted to Johns Hopkins University in conformity with the requirement for the degree of Master of Science in Engineering

Baltimore, Maryland

October 2019

Abstract

In order to maximize the concentration of amino acids in the culture, we need to obtain solubility of amino acid as a function of concentration of other components in the solution. This function can be obtained by calculating the activity coefficient along with solubility model. The activity coefficient of the amino acid can be calculated by UNIFAC. Due to the wide range of applications of UNIFAC, the prediction of the activity coefficient of amino acids is not very accurate. So we want to fit the parameters specific to amino acids based on the UNIFAC framework and existing solubility data. Due to the lack of solubility of amino acids in the multi-system, some interaction parameters are not available. COSMO is a widely used way to describe pairwise interactions in the solutions in the chemical industry. After suitable assumptions COSMO can calculate the pairwise interactions in the solutions, and largely reduce the complexion of quantum chemical calculation. In this paper, a method combining quantum chemistry and COSMO calculation is designed to accurately predict the solubility of amino acids in multi-component solutions in the

absence of parameters, as a supplement to experimental data.

Primary Reader and Advisor: Marc D. Donohue

Secondary Reader: Gregory Aranovich

Contents

List of Tables

List of Figures

1. Introduction

1.1 Introduction of phase behavior

Process intensification is increasing the demand for media with concentrated levels of nutrients needed to optimize cell growth and production of antibodies. However, many of these nutrients have limited solubility and precipitate during storage. To maximize the solubility of amino acid in the media, we need to get solubility of amino acid as a function of temperature and concentration of other component in the media environment.

For a long time, people have recognized the important role of thermodynamics in solving practical chemical problems. When applying thermodynamic principles, it is necessary to know many of the characteristics of the system, such as the PVT relationship of the fluid mixture, heat capacity, excess properties, equilibrium constants, and so on. They can certainly be obtained directly through experiments,

but as the production process becomes more complex, it is unrealistic to obtain all the data required by experimentation. In order to make better use of limited experimental data, we hope to use a small amount of experimental data to build models to predict a large number of unknown properties.

According to the phase state of the phase equilibrium, the phase equilibrium can be divided into the following categories: gas-liquid equilibrium, solid-liquid equilibrium and so on.

The amount of data in phase equilibrium is very large. As far as VLE is concerned, there are more than 100,000 data for binary systems. A large number of data collecting work began in the 1950s, and in the 1970s, not only data was sorted, but also the regression of data. With the development of computer technology, some large-scale databases have emerged. The most famous one is the Dortmund database, which involves the most substances and the most authoritative VLE data[1, 2].

To get the solubility function of amino acid, we are actually calculating a solid-liquid equilibrium. According to the principle of solid-liquid equilibrium, the solubility of

amino acids in water can generally be expressed as:

$$ln(\gamma_2 x_2) = \frac{\Delta H_t}{R}\left(\frac{1}{T_t} - \frac{1}{T}\right) - \frac{\Delta C_p}{R}\left(ln\frac{T_t}{T} - \frac{T_t}{T} + 1\right) - \frac{\Delta V}{RT}(p - p_t) \quad (1.1)$$

The activity coefficient of the amino acid can be calculated in many ways, such as EOS method, Group-contribution method and so on.

1.2 Introduction of original UNIFAC

The UNIFAC model is based on the group contribution method and is combined with the UNIFQUAC model. Fredenslund et al[3] derived the basic form of the UNIQUAC model, which expresses the activity coefficient as consisting of a combinational term and a residual term. The combinational term is related to the size and shape of the molecule, and is the contribution of entropy to the activity; the remaining term contains the energy parameter, which is degree contribution. The model can be expressed as:

$$ln\gamma_i = ln\gamma_{i(combinational)} + ln\gamma_{i(remain)} \quad (1.2)$$

Both parts in function 1.2 are based on the UNIQUAC equation.

The combinational term can be expressed as:

$$ln\gamma_{i(combinational)} = ln\frac{\phi_i}{x_i} + \frac{z}{2}q_i ln\frac{\theta_i}{\phi_i} + l_i - \frac{\phi_i}{x_i}\sum_{j=1}^{k} x_j l_j \quad (1.3)$$

In which:

$$l_i = \left(\frac{z}{2}\right)(r_i - q_i) - (r_i - 1) \quad (1.4)$$

Where, the number z is taken as 10; x_i is the mole fraction of the component i in the solution, and θ_i and φ_i are the surface area fraction and the volume fraction, respectively, which are obtained by:

$$\theta_i = \frac{q_i x_i}{\sum_j q_j x_j} \quad (1.5)$$

$$\phi_i = \frac{r_i x_i}{\sum_j r_j x_j} \quad (1.6)$$

Where q_i and r_i are the structural parameters of the pure component i, respectively, obtained by superimposing corresponding parameters of the respective groups constituting the component.

$$q_i = \sum v_k^{(i)} Q_k \quad (1.7)$$

$$r_i = \sum v_k^{(i)} R_k \quad (1.8)$$

The data required to calculate γ_i^C is the Q_k and R_k values of the groups involved. This microscopic parameter can be obtained from the Van der waals relationship given

by bondi[4].

The UNIFAC model assumes that the remainder is the synthesis of each group in the solution minus its contribution to the pure component. Residual activity coefficient γ_i^R can be expressed:

$$ln\gamma_i^R = \sum_{i=1}^{m} v_k^{(i)} \left[ln\Gamma_k - ln\Gamma_k^{(i)}\right] \quad (1.9)$$

$$ln\Gamma_k = Q_k \left[1 - ln\left(\sum_{j=1}^{m} \theta_j \psi_{jk}\right) - \sum_{j=1}^{m} \left(\frac{\theta_j \psi_{kj}}{\sum_{n=1}^{m} \theta_n \psi_{nj}}\right)\right] \quad (1.10)$$

There are two types of UNIFAC models that are relatively successful and widely used: Larsen and Dortmund UNIFAC.

1.3 Introduction of UNIFAC（LASERN）

In 1987, Larsen et al. improved the combination of the original UNIFAC model as follows:

$$ln\gamma_i^c = \left(\frac{\phi}{x_i}\right) + 1 - \frac{\phi_i}{x_i} \quad (1.11)$$

Corrected volume fraction can be expressed:

$$\phi_i = \frac{x_i r_i^{2/3}}{\sum_j x_j r_j^{2/3}} \quad (1.12)$$

Where x_i is the mole fraction of the i component in the solution; v_j is the number of

groups j in the molecular i component. Where ψ_{jk} and ψ_{kj} become the interaction parameters of the group. Group interaction parameter processing as a function of temperature:

$$\psi_{jk} = \exp\left(-\frac{a_{jk}}{T}\right) \tag{1.13}$$

$$\psi_{kj} = \exp\left(-\frac{a_{kj}}{T}\right) \tag{1.14}$$

The UNIFAC (LARSEN) model modifies the combination and improves the prediction accuracy of the non-stacking system; the temperature-dependent parameters are introduced to better predict the activity coefficient[5]. The temperature range is extended to 550-600K, and the estimation accuracy is improved compared with the original UNIFAC model. Since the UNIFAC (LARSEN) model parameters are small, the model is only used for simple compound systems such as alcohols, ethers, etc. The model can estimate only 60% of the original UNIFAC model and UNIFAC (Dortmund). UNIFAC (LARSEN) is relatively inferior to water, acid and macromolecular systems. Nor can it describe the effects of heterogeneity.

1.4 Introduction of UNIFAC（Dortmund）

In 1987, Weidlich and Gmehling corrected the combination by the method of research. The model expression of UNIFAC(Dortmund) is:

$$ln\gamma_i^C = 1 - \phi_i^{'} + ln\phi_i^{'} - 5q_i\left(1 - \frac{\phi_i}{\theta_i} + ln\left(\frac{\phi_i}{\theta_i}\right)\right) \quad (1.15)$$

in function 1.15:

$$\phi_i^{'} = \frac{r_i^{3/4}}{\sum_j x_j r_j^{3/4}} \quad (1.16)$$

$$\phi_i = \frac{r_i}{\sum_j x_j r_j} \quad (1.17)$$

The remainder is the same as the rest of the original UNIFAC model.

The UNIFAC (Dortmund) model modifies the combinational part, introduces a function of temperature, can better describe the temperature dependence, is more reliable for non-polar systems, and has significantly better precision than the original UNIFAC model. It is superior to the original UNIFAC and UNIFAC (Larsen) for estimating non-electrolyte systems. Joh et al. estimated the solid-liquid equilibrium of the 325 ternary system with an average relative deviation of 1.71%. The UNIFAC (Dortmund) model are more accurate than other models, and the average deviation

is 40%-60% smaller than other models.

1.5 Development and application of UNIFAC

The interaction parameters in the UNIFAC model depend on the interaction between the groups, and the application of the model is limited by the interaction parameters of the groups. The interaction parameters of the groups in the original UNIFAC model are not a function of temperature and therefore the applicable temperature range is small. In recent years, the DDBT database has been widely collected and has become the basis for the development of the parameters of the UNIFAC model[6].

The original UNIFAC model group was developed from the 18 main groups that were first published in 1975 to the 67 main groups today. The UNIFAC DORTMUND model introduced 45 main groups and 85 sub-groups in 1993. By 2004, it had expanded to 82 main groups[7-9], which greatly expanding the range of predictable components.

1.6 Introduction of COSMO

In the chemical industry design process, phase equilibrium data of the real mixture is

very important. Experiments are the most direct and reliable means of obtaining phase equilibrium data, but in the design calculation of the process, it is not realistic to measure all of the data. Experiments are the most direct and reliable means of obtaining phase equilibrium data, but in the design calculation of the process, it is not realistic to measure the data. To this end, chemists and engineers have conducted extensive exploration and developed many prediction methods, such as: state equation method, group contribution method, QSAR/QSPR method, molecular structure method (MONTECARLO simulation, molecular dynamics method, quantum chemical method). and many more.

The above method can deal with the interaction between molecules from different angles, and can accurately predict the phase balance within its application range, and to some extent meet the requirements of phase equilibrium data in the design.

Many predictions of thermodynamic properties are based on the group contribution method. The group contribution method assumes that the influence of a particular group on a physical property is independent of other groups in the molecule, and

the contribution of a single group is additive. This type of method reduces the number of parameters required to describe a multi-component mixture and can be used to predict the properties of a new compound by rationally defining the group and then fitting the group contribution constants with a large amount of experimental data.

Since the group contribution method is based on experiments and has a certain theoretical basis, the prediction method developed based on this principle has a more accurate budgetary result, and the calculation process is simple and convenient, and is widely used in modern industrial production. The most widely used group contribution method is the UNIFAC method and its revised versions UNIFAC (Dortmund) and UNIFAC (Larsen). The group contribution method has been applied to predict gas-liquid equilibrium, liquid-liquid equilibrium, solid-liquid equilibrium, activity coefficient, excess enthalpy, etc. However, the group contribution method is not applicable to the mixture including new functional groups. For the new system which is endless in the modernization research, the lack of experimental data leads

to a serious lack of interaction parameters of the group, which limits the application range of the method.

With the development of quantum chemistry and computer technology, especially DFT. The application of theory, even for molecules containing about 40 atoms, can be theoretically calculated to obtain higher quality molecular geometry and properties. The reaction enthalpy of industrially relevant compounds with chemical measurement level accuracy can also be calculated by efficient combined DFT configuration optimization and more advanced single point energy calculations. Forecasting by calculations is becoming more and more the focus of chemical research.

Although great progress has been made in the field of quantum chemistry calculations, quantum chemistry developed by theoretical chemists is generally only applicable to molecules in vacuum or in thin gases, and where intermolecular interactions are negligible. While many industrial-related chemical processes and most biochemical processes usually occur in liquid or multiphase systems, quantum

mechanical methods are not applicable. In the condensed matter system, the weak intermolecular interaction (Van der Waals force) and so on are very important, and the DFT theory cannot be accurately described. Therefore, it is still difficult to solve the phase equilibrium problem in practice by using only quantum chemistry. Similarly, Monte Carlo simulations, molecular dynamics simulations, etc. have also made great progress, and some of them are gradually replacing classical thermodynamics, but each has its own limitations.

The solvation thermodynamics method developed by Klamt et al[10-12]. The real solvent "partially solves the above problem. COSMO-RS characterizes the interaction between molecules by the surface shielding charge density calculated by COSMO.

COSMO is a continuous medium solvation model[13, 14] in which the dielectric constant of a continuous medium is set to infinity (ideal conductor[15]), which limits the shielding charge to the interface, so that there is no electric field between the molecule and the solvent, and there is no charge in the conductor. On the basis of COSMO, combined with statistical mechanic methods, Klamt et al. developed COSMO-RS for

quantitative calculation of solvation phenomena. COSMO-RS predicts the phase equilibrium data of a multivariate system by quantum chemical calculations of individual molecules. COSMO-RS uses the concept of local composition and can calculate the pure components and their chemical potential in the mixture. When calculating the chemical potential, consider the interaction between molecules. The model decomposes the molecules into fragments of equal surface area, and the concept of intermolecular interactions is based on the physical view of the interaction of surface fragments. The difference in energy between the two segments in the real system and the ideal conductor is measured by the net shielding charge density σ and σ'of the surface segments in contact with each other.

$$E_{misfit}(\sigma,\sigma') = \alpha_{effe_{misfit}}(\sigma,\sigma') = a_{eff}\frac{a'}{2}(\sigma+\sigma')^2 \qquad (1.18)$$

a is the effective contact surface area, a' is the energy factor, which can be calculated by electrostatic theory. If a strong polar compound is present, the hydrogen bond interaction E_{hb} should also be considered. Klamt and Eckert also proposed the expression of hydrogen contribution E_{hb}:

$$E_{hb}(\sigma,\sigma') = \alpha_{effCC_{hb}}(T) \times min\{0, max[0, \sigma_{acc} - \sigma_{hb}]min[0, \sigma_{don} + \sigma_{hb}]\} \quad (1.19)$$

in function above:

$$\sigma_{acc} = max[\sigma,\sigma'] \quad (1.20)$$

$$\sigma_{don} = max[\sigma,\sigma'] \quad (1.21)$$

The total fragment interaction energy is:

$$e(\sigma,\sigma') = E_{mf}(\sigma,\sigma') + E_{hb}(\sigma,\sigma') \quad (1.22)$$

The fragment chemical potential can be calculated by:

$$\mu_s(\sigma) = -RTln\left[\int P_s(\sigma')exp\left(\frac{\mu_s(\sigma')-\alpha_{eff}e(\sigma,\sigma')}{RT}\right)d\sigma'\right] \quad (1.23)$$

P(σ) is a very important concept in COSMO-RS for the molecular surface to shield the charge density distribution.

The molecular surface charge density required for COSMO-RS calculations is generated by quantum chemical calculations.

The σ profile is then obtained by charge averaging. The σ of the numerator i is defined as:

$$p_i(\sigma) = \frac{A_i(\sigma)}{A_i} = \frac{n_i(\sigma)}{n_i} \quad (1.24)$$

Wherein the $A_i(\sigma)$-type charge density is the total surface area of all fragments of sigma; A_i is the total hole surface area; $n_i(\sigma)$ is the number of fragments with a charge density of σ, and nt is the total number of fragments. COSMO-RS uses the following average step to get σ:

$$p_i(\sigma_n)_{COSMO-RS(OI)} = \frac{1}{3}\sum_{n-1}^{n+1} p_i(\sigma_n) \tag{1.25}$$

For the mixture, the discriminal p_s of the fragment with the shielding charge density σ was found, which was obtained by weighted average of the σ of the pure component i and its mole fraction in the system.

$$p_s(\sigma) = \frac{\sum_i x_i A_i p_i(\sigma)}{\sum_i x_i A_i} \tag{1.26}$$

The activity coefficient of the remaining part is

$$\gamma_s^j = exp\left(\frac{\mu_s^j - \mu_l^j}{RT}\right) \tag{1.27}$$

The activity coefficient of the combined part is similar to UNIQUAC.

A small number of adjustable universal constants are required for the application of COSMO-RS, some of which are derived from experimental measurements and others from experimental data regression. In theory, it is only necessary to fit the above

constants from the finite experimental data, and the obtained constants can be applied to the calculation of the properties of the new substances, and the reactive intermediates and the properties of the transition states which cannot be measured by experiments can be calculated. This is an advantage over its contribution to the group. The advantage of COSMO-RS is that, theoretically, only one COSMO calculation is required for each compound, and efficient separation of isomers is possible, and the proximity effect can also be taken into account.

The procedure for calculating the true mixture behavior using the COSMO-RS method is as follows: the quantitative calculation can simultaneously obtain the surface area Ai of the molecule and the total cavity mentioned Vi, the above information can be used to calculate the combined part of the activity coefficient; the fragment live can be obtained by solving the self-consistent equation The degree factor, the remainder of the activity coefficient of the substance in the mixture, can be derived from the fragment activity coefficient. Detailed steps and equations for applying COSMO-RS and the parameters used can be found in references[16, 17].

As a method for describing the thermodynamic properties of fluid phases, COSMO-RS has developed rapidly due to the lack of experimental data. The COSMO paper elaborated by Klamt has been consumed more than 1,000 times. COSMO-RS can be used to predict gas-liquid equilibrium, liquid-liquid equilibrium, solid-liquid equilibrium[18], vapor pressure of pure components and mixtures, partition coefficient[19], adsorption equilibrium[20, 21], solubility[22, 23], evaporation enthalpy[24, 25], Pka[26], and viscosity[27]. The application system of COSMO-RS has also been extended to ionic liquids, polymer solutions, surface-active micelles, biofilms and the like[28].

Molecular sigma is required for the use of COSMO-RS. The COSMO method for calculating sigma is embedded in different quantum chemistry calculation software such as Gaussian, Turbomol, MOPAC, DMol3, GAMEss and so on.

1.7 Purpose and design

In a quantitative sense, the binding group contribution method and the solubility model can predict the solubility of organic matter in solution. By using the solubility data of amino acids to narrow the fit range, fitting the UNIFAC parameters specific to

amino acids can further improve the accuracy of solubility prediction. Since the solubility data of available amino acids in multi-component solutions is very small, this paper designs a solubility prediction method combining quantum mechanical calculation and COSMO method to predict the solubility of amino acids as a supplement to the missing experimental data.

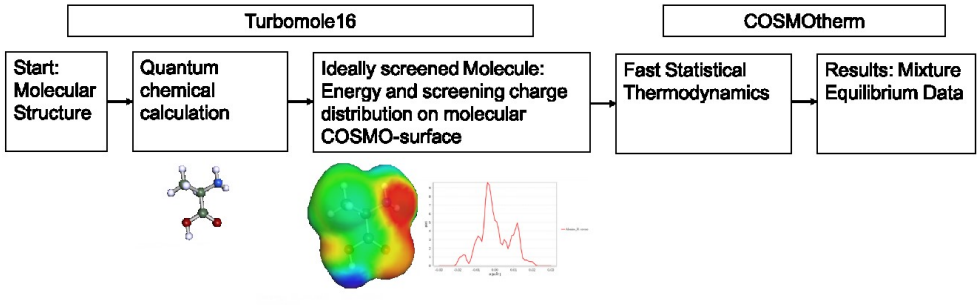

Fig 1. Flowchart of prediction with COSMO

2.Experiment

2.1 Quantum mechanical calculation with Gaussian

Gaussian is a general purpose computational chemistry software package initially released in 1970 by John Pople[29] and his research group at Carnegie Mellon University as Gaussian 70. It has been continuously updated since then. Gaussian 16 is the latest version of the Gaussian series of electronic structure programs, used by chemists, chemical engineers, biochemists, physicists and other scientists worldwide. Gaussian 16 provides a wide-ranging suite of the most advanced modeling capabilities available. You can use it to investigate the real-world chemical problems that interest you, in all of their complexity, even on modest computer hardware. The energy change going from the gas phase to solution is known as the solvation energy of a molecule. It can be computed for the same compound with several solvents in order to understand its relative solubility in different environments. The predicted free energy can also be used to predict reaction energies in solution. The SMD method is

an SCRF-based solvation model from Truhlar and coworkers. It was parametrized specifically to predict free energies of solvation, and includes different values for the non-electrostatic terms.

Calculating COSMO file with Gaussian has three steps: 1. Configuration optimization; 2. PCM and DFT calculation, in order to get the σ and COSMO surface required for COSMO calculation; 3. COSMO file can be read by COSMOtherm. Calculate solubility of amino acids. The interface of Gaussian is shown in figure 2. All the commands need to be compiled to input. For the purpose of doing a COSMO calculation, here I choose B3LYP 6-311G mode calculating mode. Under this mode Gaussian generate a COSMO file which contains all COSMO surface and C profile for next step.

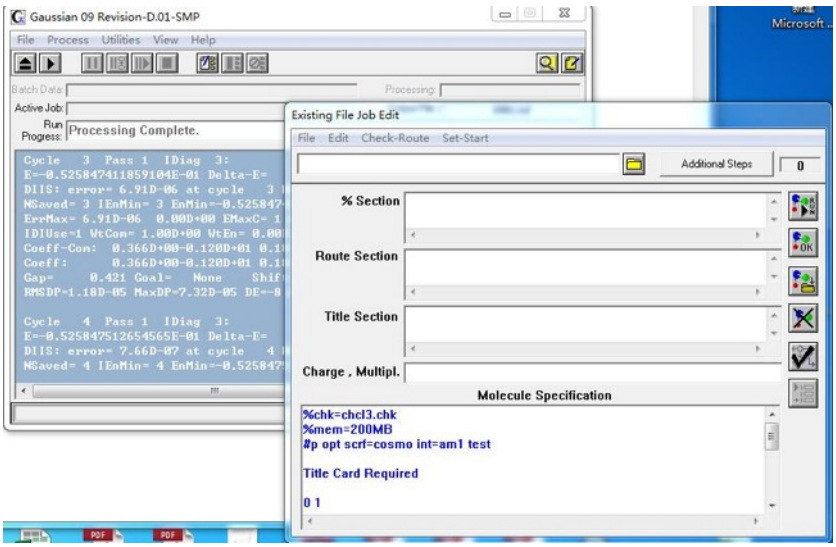

Fig 2 Interface of Gaussian

Since amino acid molecules has dozens of atoms, the atomic matrix that needs to be input is very large, and it is easy to generate errors in the input process, resulting in optimization errors. Compared to inconvenience of Gaussian, Turbomole has a graphic interface, it is much efficient and easy to use. I choose Turbomole to do quantum chemical calculation.

2.2 Quantum mechanical calculation with Turbomole

Turbomole was developed in 1987 and turned into a mature program system under the control of Reinhart Ahlrichs and his collaborators[30]. Turbomole can perform a large-scale quantum chemical simulations of molecules, clusters, and later periodic solids. Gaussian basis sets are used in Turbomole. The functionality of the program concentrates extensively on the electronic structure methods with effective cost-performance characteristics such as density functional theory[31], second–order Møller-Plesset[32, 33] and coupled cluster theory. Aside from energies and structures, an assortment of optical, electrical, and magnetic properties are available from analytical

energy derivative for electronic ground and excited states. However, up to the year 2000, Turbomole was only limited to the calculation of molecules in gas phase, thus, COSMO has been implemented in the Turbomole in a cooperative initiative of BASF AG and Bayer AG. Turbomole version 6.5 releasing in the year 2013, comes with post-Kohn-Sham calculations within the random-phase approximation. Turbomole also comes with another significant additions including nonadiabatic molecular dynamics, ultra-efficient higher order CC methods, new density functionals and periodic calculations. TmoleX is available as a graphical user interface for Turbomole allowing the user to perform the entire workflow of a quantum chemical investigation ranging from building of an initial structure to the interpretation of the results.

I use Turbomole16 to do quantum chemical computation under its SVP mode, 3D structure after optimization can be seen in figure 3:

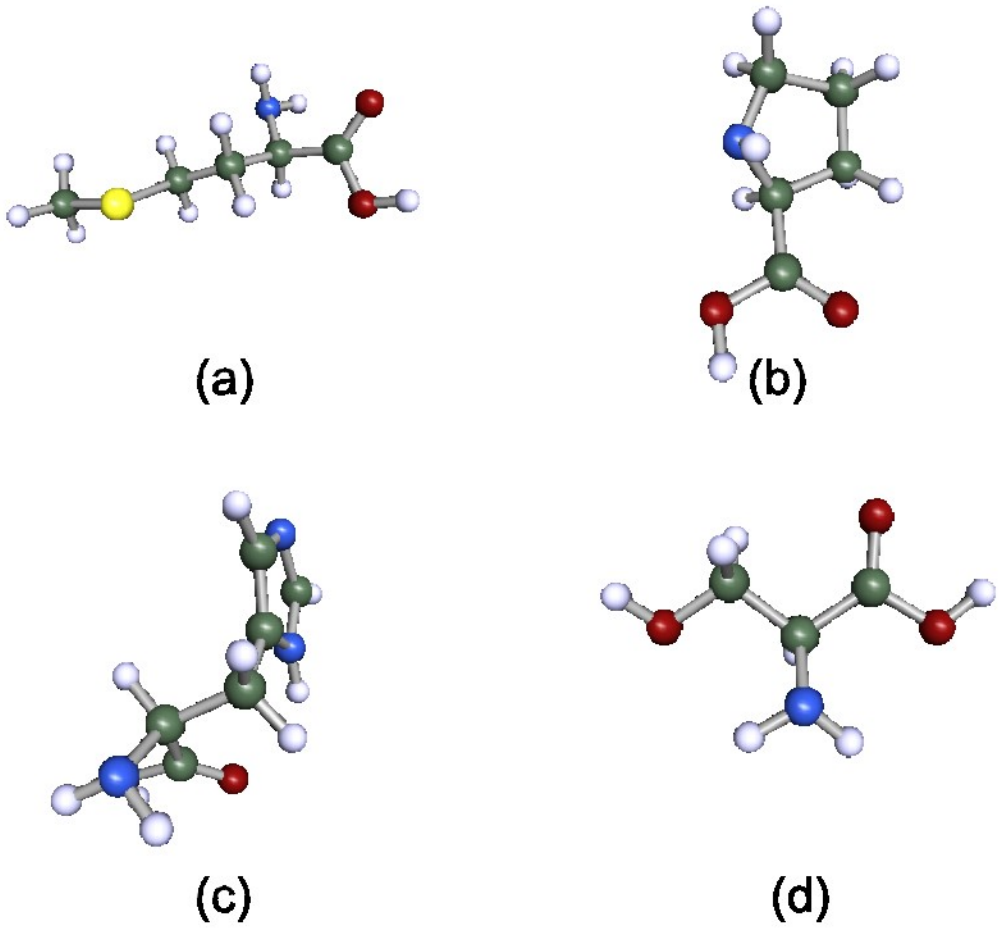

Fig 3. 3D structures of 4 amino acids after optimization: (a) Methionine (b) Proline (c) Histidine (d) Serine

The second step in MD/MC is the reduction of the real quantum chemical system to an ensemble of pair-wise interacting spheres, having certain interaction parameters which are derived from the initial QC step. Instead, in COSMO-RS we represent the system by surface pieces, having interaction parameters from QC. After calculation

with Turbomole, COSMO surface of amino acids can be shown in figure 4:

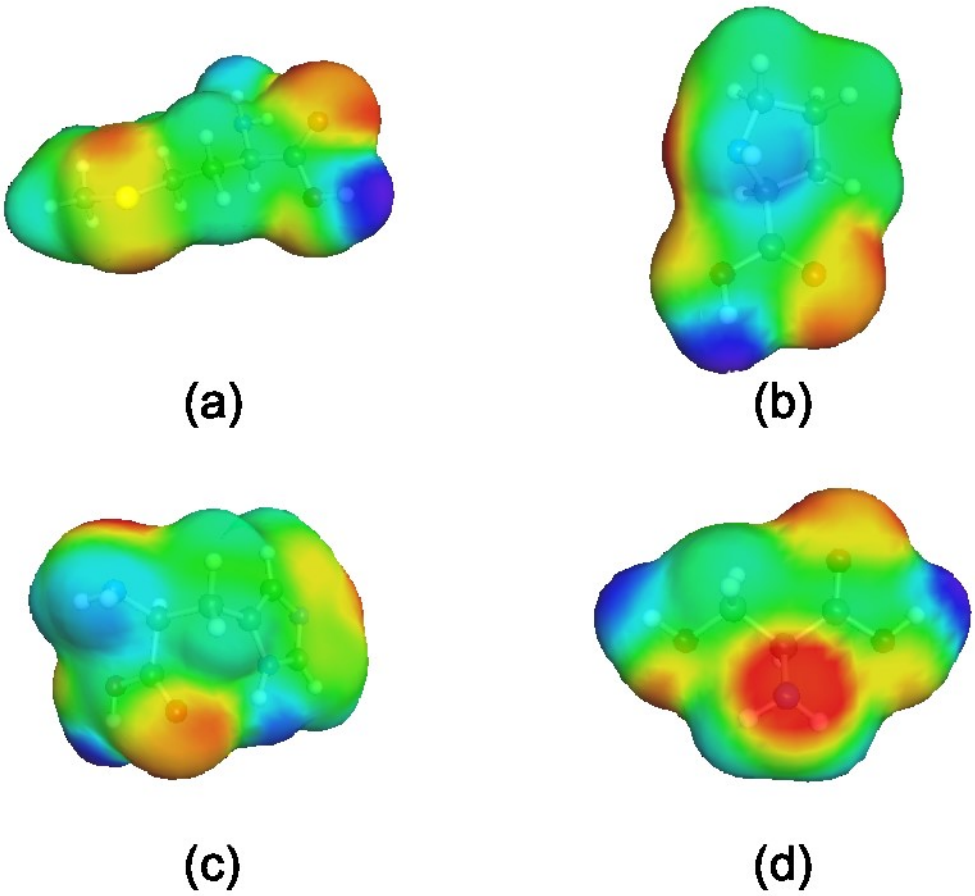

Fig 4. COSMO surface of 4 amino acids: (a) Methionine (b) Proline (c) Histidine (d) Serine

As a result of a MC calculation we do not only yield the total energy of X in its self-consistent state in the conductor, but we also gain the polarization charge density σ, which the conductor places on the cavity in order to screen the electric field of the

molecule. σ profile of amino acids can be seen in figure 5:

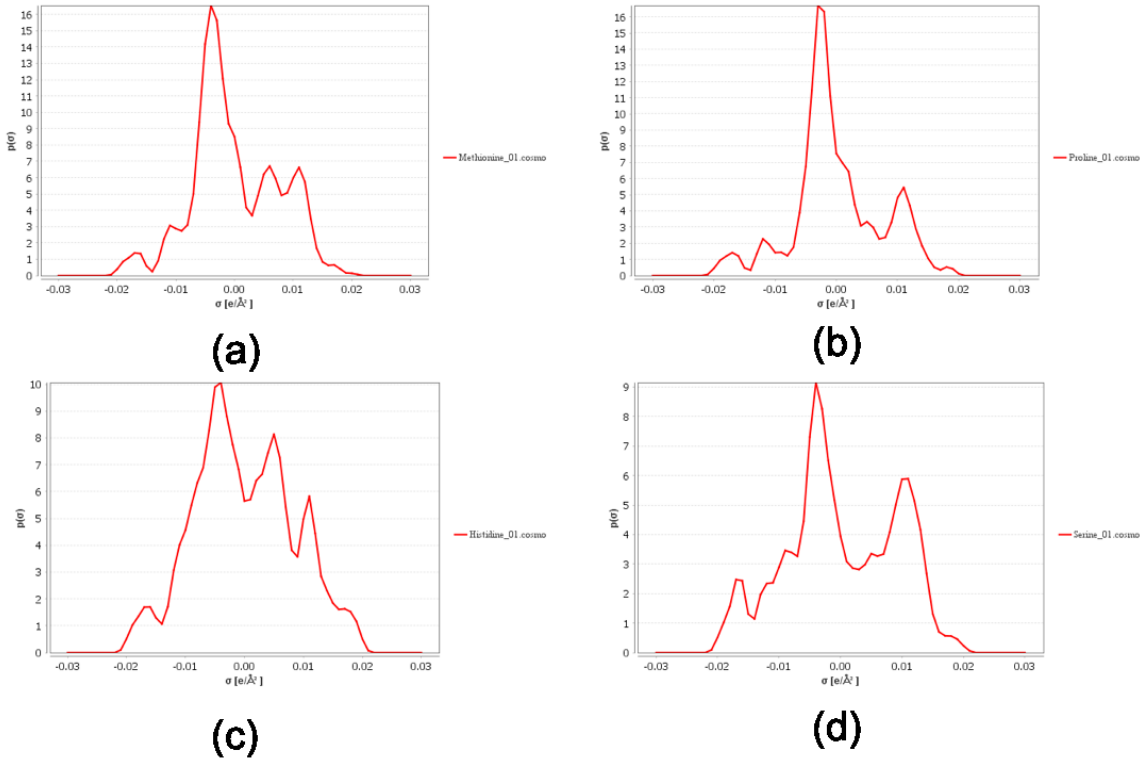

Fig 5. σ potential of 4 amino acids: (a) Methionine (b) Proline (c) Histidine (d) Serine

By finishing these 3 steps, a COSMO file is generated and can be read by COSMOtherm.

2.3 Solubility Prediction by COSMOtherm

COSMOtherm is the universal tool for predictive property calculation of liquids, and combines quantum chemistry and thermodynamics in a unique fashion. It calculates

the chemical potential of almost any molecule in almost any pure or mixed liquid at variable temperature, i.e. it predicts how happy a molecule is in a certain liquid environment. This is the key for the prediction of a multitude of properties required in industrial applications or academic research, including solubility, partitioning, vapor pressure, and complete phase diagrams. In contrast to several other available methods COSMOtherm is able to predict properties as function of concentration and temperature by applying thermodynamically consistent equations. COSMOtherm is the first publicly available and most advanced implementation of the COSMO-RS theory, which was published by A. Klamt at Bayer in 1995. He started COSMOlogic a few years later to focus on the advancement of COSMO-RS and COSMOtherm. COSMOtherm has found widespread use in many industrial branches related to chemistry, pharmaceutics, consumer goods or fragrances.

By reading the COSMO file generated by quantum mechanical computational software. COSMOtherm can use the sigma profile to calculate the solubility.

First I use COSMOtherm to calculate solubility of 16 binary systems, which contain

only 1 amino acid and water from 298K to 323K, to check the accuracy of COSMO calculation.

Table 1. Comparison between predicted and experimental data for binary systems

Amino Acid		298K	303K	323K	AAD
Alanine	exp	166.5	175.7	217.9	
	pred	180.3	195.7	268.8	0.14
Arginine	exp	196.0	237.0	434.5	
	pred	240.3	293.7	497.7	0.20
Asparagine	exp	5.0	6.0	12.0	
	pred	5.1	5.9	11.8	0.02
Glutamic Acid	exp	8.6	10.2	21.9	
	pred	9.4	11.3	22.9	0.08
Glutamine	exp	42.5	50.7	87.1	

	pred	47.0	56.8	99.6	0.12
Glycine	exp	249.9	275.9	391.0	
	pred	284.2	325.5	523.3	0.22
Histidine	exp	43.6	48.6	67.6	
	pred	52.8	56.7	75.2	0.16
Leucine	exp	24.3	24.9	28.9	
	pred	31.1	31.9	36.4	0.27
Lycine	exp	246.6	273.2	367.2	
	pred	284.3	287.5	298.2	0.13
Methionine	exp	51.4	56.2	75.2	
	pred	62.3	69.3	99.0	0.25
Proline	exp	1623.0	1703.0	2067.0	
	pred	1923.4	1984.3	2304.3	0.16
Serine	exp	422.0	476.0	687.0	
	pred	448.0	495.6	705.5	0.04

Taurine	exp	57.3	62.3	82.6	
	pred	60.5	69.7	99.7	0.13
Tryptophan	exp	11.4	12.5	17.1	
	pred	14.2	14.6	16.9	0.14
Tyrosine	exp	0.5	0.5	1.1	
	pred	0.5	0.7	1.3	0.19
Valine	exp	58.1	59.7	65.5	
	pred	65.7	66.2	69.1	0.10

Unit (g/L)

From table 1 we can see the errors are pretty small, which verifies that COSMO model has a good performance on predicting solubility of binary systems.

And I calculated ternary systems which have solubility data to verify the performance of COSMO on multiple systems. Here I chose 2 ternary systems and calculate solubility of each solute at 303K and 323K.

Table 2. Comparison between predicted and experimental data for ternary

systems at 303K

303K	Serine	Alanine(exp)	Alanine(pred)	AAD
	0.00	175.70	195.70	
	20.00	179.70	197.30	
	40.00	183.80	199.70	
				0.10
	Alanine	Serine(exp)	Serine(pred)	
	0.00	476.00	495.60	
	40.00	486.00	498.30	
	80.00	496.00	501.30	
				0.03
	Leucine	Alanine(exp)	Alanine(pred)	
	0.00	175.70	195.70	
	90.00	173.60	195.50	
	130.00	171.30	195.40	

				0.13
	Alanine	Leucine(exp)	Leucine(pred)	
	0.00	175.70	195.70	
	50.00	155.30	175.30	
	100.00	137.40	156.30	
				0.13

Unit (g/L)

Table 3. Comparison between predicted and experimental data for ternary systems at 323K

323K	Serine	Alanine(exp)	Alanine(pred)	AAD
	0.00	217.90	268.80	
	20.00	222.30	271.10	
	40.00	226.90	273.40	
				0.22
	Alanine	Serine(exp)	Serine(pred)	

0.00	687.00	705.50	
40.00	703.50	702.30	
80.00	719.30	698.40	
			0.02

Leucine	Alanine(exp)	Alanine(pred)	
0.00	289.00	217.90	
90.00	287.00	216.40	
130.00	285.00	216.10	
			0.24

Alanine	Leucine(exp)	Leucine(pred)	
0.00	289.00	217.90	
50.00	274.60	193.20	
100.00	259.20	172.10	
			0.29

Unit (g/L)

From table 2 and 3 we can see that COSMO can accurately predict the solubility of amino acids in the ternary system. And it has higher accuracy at lower temperatures. In addition, COSMO can also predict the solubility of multi-systems, which are combinations of multiple amino acids and water.

3. Conclusion

For solution systems containing multiple amino acids, this paper proposes a solubility prediction method based on COSMO model. The full text mainly draws the following conclusions:

1.The method of predicting amino acid solubility based on COSMO was verified, and the $\boldsymbol{\sigma}$ profile and COSMO surface of 16 common amino acids were obtained. Prediction of solutions containing these 20 amino acids can be made directly with COSMOtherm.

2. The accuracy of predicting amino acid solubility using the COSMO model was verified by comparison with experimental data. The method is more accurate at lower temperatures.

Reference

1. Gmehling, J.; Onken, U.; Behrens, D.; Eckermann, R., *Vapor-liquid equilibrium data collection: Aqueous-organic systems*. Dechema Frankfurt, Germany: 1977; Vol. 1.

2. Marsh, K. N.; Niamskul, P.; Gmehling, J.; Bölts, R. J. F. P. E., Review of thermophysical property measurements on mixtures containing MTBE, TAME, and other ethers with non-polar solvents. **1999,** *156* (1-2), 207-227.

3. Prausnitz, J. M.; Lichtenthaler, R. N.; de Azevedo, E. G., *Molecular thermodynamics of fluid-phase equilibria*. Pearson Education: 1998.

4. Gubbins, K. E.; Quirke, N., *Molecular simulation and industrial applications: methods, examples, and prospects*. Taylor & Francis: 1996; Vol. 1.

5. Larsen, B. L.; Rasmussen, P.; Fredenslund, A. J. I.; research, e. c., A modified UNIFAC group-contribution model for prediction of phase equilibria and heats of mixing. **1987,** *26* (11), 2274-2286.

6. Onken, U.; Rarey-Nies, J.; Gmehling, J. J. I. J. o. T., The Dortmund Data Bank: A

computerized system for retrieval, correlation, and prediction of thermodynamic properties of mixtures. **1989,** *10* (3), 739-747.

7. Gmehling, J.; Li, J.; Schiller, M. J. I.; Research, E. C., A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. **1993,** *32* (1), 178-193.

8. Gmehling, J.; Lohmann, J.; Jakob, A.; Li, J.; Joh, R. J. I.; research, e. c., A modified UNIFAC (Dortmund) model. 3. Revision and extension. **1998,** *37* (12), 4876-4882.

9. Gmehling, J.; Wittig, R.; Lohmann, J.; Joh, R. J. I.; research, e. c., A modified UNIFAC (Dortmund) model. 4. Revision and extension. **2002,** *41* (6), 1678-1688.

10. Klamt, A. J. T. J. o. P. C., Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. **1995,** *99* (7), 2224-2235.

11. Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. J. T. J. o. P. C. A., Refinement and parametrization of COSMO-RS. **1998,** *102* (26), 5074-5085.

12. Klamt, A.; Eckert, F. J. F. P. E., COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. **2000,** *172* (1), 43-72.

13. Tomasi, J.; Cammi, R.; Mennucci, B. J. I. j. o. q. c., Medium effects on the properties of chemical systems: An overview of recent formulations in the polarizable continuum model (PCM). **1999,** *75* (4-5), 783-803.

14. Tomasi, J.; Mennucci, B.; Cammi, R. J. C. r., Quantum mechanical continuum solvation models. **2005,** *105* (8), 2999-3094.

15. Klamt, A.; Schüürmann, G. J. J. o. t. C. S., Perkin Transactions 2, COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. **1993,** (5), 799-805.

16. Lin, S.-T.; Sandler, S. I. J. I.; research, e. c., A priori phase equilibrium prediction from a segment contribution solvation model. **2002,** *41* (5), 899-913.

17. Grensemann, H.; Gmehling, J. J. I.; research, e. c., Performance of a conductor-like screening model for real solvents model in comparison to classical group contribution methods. **2005,** *44* (5), 1610-1624.

18. Klamt, A.; Eckert, F.; Hornig, M.; Beck, M. E.; Bürger, T. J. J. o. c. c., Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. **2002,** *23* (2), 275-281.

19. Klamt, A.; Eckert, F.; Diedenhofen, M. J. E. T.; Journal, C. A. I., Prediction of soil sorption coefficients with a conductor-like screening model for real solvents. **2002,** *21* (12), 2562-2566.

20. Mehler, C.; Klamt, A.; Peukert, W. J. A. j., Use of COSMO-RS for the prediction of adsorption equilibria. **2002,** *48* (5), 1093-1099.

21. Schürer, G.; Peukert, W. J. A., Prediction of adsorption equilibria from physical properties of the pure components. **2005,** *11* (1), 43-47.

22. Kolář, P.; Nakata, H.; Shen, J.-W.; Tsuboi, A.; Suzuki, H.; Ue, M. J. F. p. e., Prediction of gas solubility in battery formulations. **2005,** *228*, 59-66.

23. Ikeda, H.; Chiba, K.; Kanou, A.; Hirayama, N. J. C.; bulletin, p., Prediction of solubility of drugs by conductor-like screening model for real solvents. **2005,** *53* (2), 253-255.

24. Lin, S.-T.; Chang, J.; Wang, S.; Goddard, W. A.; Sandler, S. I. J. T. J. o. P. C. A.,

Prediction of vapor pressures and enthalpies of vaporization using a COSMO solvation model. **2004,** *108* (36), 7429-7439.

25. Emel'yanenko, V. N.; Verevkin, S. P.; Heintz, A. J. J. o. t. A. C. S., The gaseous enthalpy of formation of the ionic liquid 1-butyl-3-methylimidazolium dicyanamide from combustion calorimetry, vapor pressure measurements, and ab initio calculations. **2007,** *129* (13), 3930-3937.

26. Klamt, A.; Eckert, F.; Diedenhofen, M.; Beck, M. E. J. T. J. o. P. C. A., First principles calculations of aqueous p K a values for organic and inorganic acids using COSMO− RS reveal an inconsistency in the slope of the p K a scale. **2003,** *107* (44), 9380-9386.

27. Bosse, D.; Bart, H.-J. J. I.; research, e. c., Viscosity calculations on the basis of Eyring's absolute reaction rate theory and COSMOSPACE. **2005,** *44* (22), 8428-8435.

28. Klamt, A., *COSMO-RS: from quantum chemistry to fluid phase thermodynamics and drug design*. Elsevier: 2005.

29. Hehre, W.; Lathan, W.; Ditchfield, R.; Newton, M.; Pople, J., Gaussian 70

(Quantum Chemistry Program Exchange. Program: 1970.

30. Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. J. C. P. L., Electronic structure calculations on workstation computers: The program system turbomole. **1989,** *162* (3), 165-169.

31. Von Arnim, M.; Ahlrichs, R. J. J. o. c. c., Performance of parallel TURBOMOLE for density functional calculations. **1998,** *19* (15), 1746-1757.

32. Gerenkamp, M.; Grimme, S. J. C. p. l., Spin-component scaled second-order Møller–Plesset perturbation theory for the calculation of molecular geometries and harmonic vibrational frequencies. **2004,** *392* (1-3), 229-235.

33. Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C.; Eckert, F. J. P. C. C. P., COSMO Implementation in TURBOMOLE: Extension of an efficient quantum chemical code towards liquid systems. **2000,** *2* (10), 2187-2193.

Vita

Kaiyu Li

kli41@jhu.edu | +1-4438396057

30 West Biddle Street Baltimore, MD 21201 US

Date of Birth: 01/12/1991

Location of Birth: China

ACADEMIC EXPERIENCE

12/2017-present	*Dr. Marc Donohue's Research Group*
12/2014-08/2017	*State Key Laboratory of Heavy Oil Processing*