I&EC research
Industrial & Engineering Chemistry Research

Article

A Critical Evaluation on the Performance of COSMO-SAC Models for Vapor-Liquid and Liquid-Liquid Equilibrium Predictions based on Different Quantum Chemical Calculations

Wei-Lin Chen, Chieh-Ming Hsieh, Li Yang, Chan-Chia Hsu, and Shiang-Tai Lin

Ind. Eng. Chem. Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.iecr.6b02345 • Publication Date (Web): 12 Aug 2016

ACS Publications

Industrial & Engineering Chemistry Research is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

A Critical Evaluation on the Performance of COSMO-SAC Models for Vapor-Liquid and Liquid-Liquid Equilibrium Predictions based on Different Quantum Chemical Calculations

Wei-Lin Chen[a], Chieh-Ming Hsieh[b], Li Yang[ac], Chan-Chia Hsu[d], and Shiang-Tai Lin[a*]

[a]Department of Chemical Engineering, National Taiwan University, Taipei 10617, Taiwan

[b]Department of Chemical and Materials Engineering, National Central University, Jhongli 32001, Taiwan

[c]Key Laboratory of Green Chemical Process of Ministry of Education, Key Laboratory of Novel Reactor and Green Chemical Technology of Hubei Province, School of Chemical Engineering and Pharmacy, Wuhan Institute of Technology, Wuhan 430073, China

[d]LCY Chemical Corporation, Taipei, Taiwan

Corresponding author. Tel.:+886 2 3366 1369; fax: +886 2 2362 3040. E-mail address:stlin@ntu.edu.tw (S.-T. Lin).

Keywords: COSMO-SAC model, activity coefficient, vapor-liquid equilibrium, liquid-liquid equilibrium, octanol-water partition coefficient, density functional theory, basis set.

Abstract: The performance of two versions of the COSMO-SAC activity coefficient model is carefully examined based on 8 sets of quantum chemical computations (VWN-BP/DNP, b3lyp/6-31G(d,p), b3lyp/6-31G(2d,p), b3lyp/6-31+G(d,p), b3lyp/6-311G(d,p), wb97xd/6-31G(d,p), wb97xd/6-31G(2d,p), wb97xd/6-31+G(d,p)) and one semi-empirical calculations (PM6). Furthermore, the effect of molecular geometry is examined based on equilibrium structures determined both in vacuum, representing nonpolar environment, and in conductor, representing highly polar environment. The model parameters are re-optimized for each quantum chemical calculation method and the performance is evaluated using a large set of database covering the vapor-liquid equilibrium (VLE), liquid-liquid equilibrium (LLE), infinite dilution activity coefficient (IDAC) of binary mixtures, and octanol-water partition coefficient (K_{ow}) (containing over 22 thousand data points). It is found that the original COSMO-SAC model is sensitive to the quantum chemical method used; whereas the revised COSMO-SAC model is not. For the original COSMO-SAC, a method that gives higher molecular polarity often results in a better prediction accuracy. The modifications introduced in the revised COSMO-SAC model not only improves the accuracy but also allows for the use of a lower quality quantum computational theory without much loss of its accuracy.

1. Introduction

The COSMO-SAC model[1] is a predictive method for liquid phase nonideality based on results of quantum mechanical calculations. Based on the same physical picture of the pioneering COSMO-RS model,[2] the COSMO-SAC resolves the thermodynamic inconsistency in the original COSMO-RS model[1] and has been successfully applied in many fields, such as vapor-liquid equilibrium, liquid-liquid equilibrium, drug solubility, partition coefficient, ionic liquid screening, etc.[3-8] The COSMO-based methods are in some ways similar to the popular group contribution method for activity coefficient, e.g., the UNIFAC model.[9] In UNIFAC the molecular activity coefficient is calculated as sum of contributions from functional groups with group interaction parameters obtained from regression to a large set of experimental data.[10] Instead of functional group interactions, the COSMO-based methods consider interactions through contacting molecular surfaces in the liquid phase. The strength of such surface interactions are quantified using apparent screening charges on the molecular surface when the molecule is solvated by a perfect conductor. By considering the molecule as a collection of surface segments, each possessing a certain amount of screening charges, the molecular activity coefficient is calculated as sum of contributions from the surface segments. The outstanding merit of the surface segment approach (as compared to functional group approach) is that the segment interactions can be determined based on the charge density on the interacting segments.[2] Hence, there is no species or functional group dependent parameters involved in the COSMO-based methods.

While the UNIFAC model determines the functional group contributions based on the UNIQUAC activity co-

efficient model,[11] the COSMO-SAC model determines the segment coefficient from a self-consistent segment activity coefficient equation. It has been shown[12] that local composition models based on Boltzmann distribution, such as UNIQUAC, does not satisfy the Flemr condition,[13] i.e., the number of neighboring pairs should be the same for either of the pairing species. The self-consistent equation for segment activity coefficient of COSMO-SAC indeed satisfies the Flemr condition. The COSMO-SAC model has a sound theoretical basis in the detailed neighboring pairs, and therefore, is often found to provide consistent accuracy for both intermediate and infinite dilution conditions.[12]

The distribution of screening charges on the molecular surface, also known as the σ-profile, provides the most important chemical information of a chemical of interest. The surface charges can be calculated based on the fact that the net electric potential at the boundary (molecular surface) between the solute (molecule of interest) and the solvent (conductor) is zero.[14, 15] Therefore, the σ-profile reflects the electronic structure and electrostatic characters of the molecule (molecular fingerprint). There are many software packages available for providing the surface screening charge densities of a solvated molecule based on highly accurate quantum chemical treatment of the solute.[16-19]

The calculation for the screening charges, referred to as QM/COSMO calculations, is also the most time-consuming step in the COSMO-SAC calculation for the activity coefficient. The QM/COSMO calculation may take a few hours for a molecule of about 50 atoms on a modern PC while the activity coefficient calculations takes only milliseconds. Therefore, it is sensible to collect the calculated results for later use in the mixture property calculations. For example, the VT-database[20] from Liu's group at Virginia Tech contains such data for more than 1400 compounds calculated using DMol3.[18]

With the advances in computational chemistry, there are more accurate theoretical methods developed and larger basis sets becomes available. It is therefore an interesting question whether the COSMO-SAC model can be improved with the use of higher level quantum theories and larger basis sets. Mu and Gmehling[21] examined the performance of COSMO-SAC and COSMO-RS on predicting the infinite dilution activity coefficient (IDAC) and the vapor-liquid equilibrium (VLE) using the screening charge distribution from BP/TZVP and B3LYP/6-311G(d,p). They found that B3LYP/6-311G(d,p) provides more accurate predictions compared to BP/TZVP. However, in their study, the model parameters were taken from the original paper based on the VT-database (determined using VWN-BP/DNP) and were not optimized for each QM/COSMO calculation. Franke and Hannebauer[22] examined the performance dependence of COSMO-RS on the QM/COSMO calculations based on infinite dilution activity coefficient (IDAC). They found that re-optimization of the model parameters for different quantum chemical methods significantly improves the prediction accuracy. They suggested that the basis set of double-zeta in valance and generalized gradient approximation (GGA) functionals are the minimum requirement. More recently Paulechaka et al.[23] propose a variant of COSMO-SAC model and re-parameterize it based on their σ–profile database containing 897 molecules using the quantum calculation method in b3lyp/6-311G(d,p).

Despite of these works, there have been no systematic study on how the use of a different QM/COSMO method influences performance of the COSMO-SAC model in various kinds of phase behaviors, VLE, LLE, IDAC, etc. The purpose of the present work is to address this important question and provide an optimal set of model parameters for some of the commonly used QM/COSMO methods.

2. Theory

In the COSMO-SAC model[1, 3] the activity cient, $\gamma_{i/S}$, for solute *i* in solution *S* is determined from the solvation free energy[24]

$$\ln \gamma_{i/S} = \frac{\Delta G_{i/S}^{*res} - \Delta G_{i/i}^{*res}}{RT} + ln\gamma_{i/S}^{SG} \quad (1)$$

where $\Delta G_{i/S}^{*res}$ is the restoring free energy, $\gamma_{i/S}^{SG}$ is the Staverman-Guggenheim combinatorial term relating to molecular size and shape differences between the species. Superscript * is used to emphasize that the solute is not allowed to move in the liquid phase, according to the definition of solvation process by Ben-Naim.[25]

The restoring energy accounts for the non-ideality as a result of different molecular interactions. This is calculated from the surface screening charges when the molecule is in close contact. The ideal screening charges, q*, are calculated from the first-principle solvation calculation in a perfect conductor (infinite dielectric constant). A semi-theoretical charge-averaging process [see Supporting Information for details] is applied to obtain the averaged charges, q. The distribution of the charge density, the averaged charge on a segment divided by its area, is quantified as σ-profile, $P_i(\sigma)$ (probability of finding a surface segment with charge density σ). In other words, $P_i(\sigma)$ is the ratio of surface area with charge density of σ for molecule *i* [$A_i(\sigma)$] to the total molecular surface area (A_i). For a mixture, the σ-profile is determined from the area weighted average of the pure components

$$P_s(\sigma) = \frac{\sum_i x_i A_i P_i(\sigma)}{\sum_i x_i A_i} \quad (2)$$

The original version of COSMO-SAC categorized segments into two types, hydrogen-bonding ($A_i^{hb}(\sigma)$) and non- hydrogen bonding ($A_i^{nhb}(\sigma)$),where $A_i^{hb}(\sigma)$ is the surface whose charge density exceed a cutoff value, σ_{hb}=0.0084 e/Å. The σ-profile then is

$$P_i(\sigma) = P_i^{nhb}(\sigma) + P_i^{hb}(\sigma) = A_i^{nhb}(\sigma)/A_i + A_i^{hb}(\sigma)/A_i \quad (3)$$

The restoring free energy can be obtained from the summation of all the segment contributions

$$\frac{\Delta G_{i/S}^{*res}}{RT} = \frac{A_i}{a_{eff}} \sum_s^{nhb,hb} \sum_{\sigma_m} P_i^s(\sigma_m^s) \ln \Gamma_j^s(\sigma_m^s) \quad (4)$$

where a_{eff} is the standard segment surface; subscript j can either be the pure liquid (i) or a mixture (S); the superscript s and t can be either nhb or hb; Γ_j^s, the activity coefficient of segment j with charge density σ_m, can be determined by the σ-profile of the fluid and the segment exchange energy (ΔW):

$$\ln \Gamma_j^t(\sigma_m^t) = -ln\left\{\sum_s^{hb,nhb} \sum_{\sigma_n} P_j^s(\sigma_n^s) \Gamma_j^t(\sigma_n^s) \exp\left(\frac{-\Delta W(\sigma_m^t, \sigma_n^s)}{kT}\right)\right\} \quad (5)$$

$$\Delta W(\sigma_m^t, \sigma_n^s) = C_{ES}(\sigma_m^t + \sigma_n^s)^2 + C_{hb} \max[0, \sigma_{acc} - \sigma_{hb}] \min[0, \sigma_{don} + \sigma_{hb}] \quad (6)$$

where C_{ES} is the theoretical value of electrostatic interaction parameter; $C_{hb}(\sigma_m^t, \sigma_n^s)$ is the hydrogen bonding parameter; σ_{acc} and σ_{don} are the larger and smaller values of σ_m and σ_n.

$$C_{ES} = \frac{f_{pol} \times 0.3 \times {a_{eff}}^{3/2}}{2 \times \epsilon_0} \quad (7)$$

$$C_{hb}(\sigma_m^t, \sigma_n^s) = \begin{cases} C_{hb} & if\ s = t = hb\ and\ \sigma_m^t \cdot \sigma_n^s < 0 \\ 0 & otherwise \end{cases} \quad (8)$$

where the f_{pol} is the polarization factor.

For a better description on the interaction between segments, Hsieh et al.[3] proposed a revised COSMO-SAC model. Two major modifications were introduced. First, the electrostatic interaction is made to be temperature dependent.

$$C_{ES} = A_{ES} + \frac{B_{ES}}{T^2} \quad (9)$$

Secondly, the hydrogen bonding surfaces from hydroxyl and non-hydroxyl groups are differentiated.

$$P_i^{hb}(\sigma) = \left[{P_i}^{OH}(\sigma) + {P_i}^{OT}(\sigma)\right] \times P^{HB}(\sigma) \quad (10)$$

where the OH represents the surface on the hydroxyl group and the OT represents the other hydrogen bonding surface. The $P^{HB}(\sigma) = 1 - exp\left(\frac{\sigma}{2\,{\sigma_0}^2}\right)$ represents the possibility of forming a hydrogen bond for a hb segment and σ_0 is 0.007 e/Å^2.[26] Therefore σ-profile for nhb segment becomes $P_i^{nhb}(\sigma) + P_i^{hb}(\sigma) \times (1 - P^{HB}(\sigma))$.

The hydrogen bonding interactions are modified as

$$\Delta W(\sigma_m^t, \sigma_n^s) = C_{ES}(\sigma_m^t + \sigma_n^s)^2 - C_{hb}(\sigma_m^t, \sigma_n^s)(\sigma_m^t - \sigma_n^s)^2 \quad (11)$$

with the interaction parameters for different hydrogen bonding interactions

$$C_{hb}(\sigma_m^t, \sigma_n^s) = \begin{cases} C_{OH-OH} & if\ s = t = OH\ and\ \sigma_m^t \times \sigma_n^s < 0 \\ C_{OT-OT} & if\ s = t = OT\ and\ \sigma_m^t \times \sigma_n^s < 0 \\ C_{OH-OT} & if\ s = OH, t = OH\ and\ \sigma_m^t \times \sigma_n^s < 0 \\ 0 & otherwise \end{cases} \quad (12)$$

3. Computational Details

The screening charge density distribution on molecular surface is determined for the following QM/COSMO methods

1. VWN-BP/DNP
2. b3lyp/6-31G(d,p)
3. b3lyp/6-31G(2d,p)
4. b3lyp/6-31+G(d,p)
5. b3lyp/6-311G(d,p)
6. wb97xd/6-31G(d,p)
7. wb97xd/6-31G(2d,p)
8. wb97xd/6-31+G(d,p)
9. b3lyp/6-31G(d,p)-cosmo
10. PM6

The VWN-BP/DNP results are taken from the VT-database.[20] In this database, the equilibrium molecular geometry is first determined in vacuum, and the COSMO calculations are then performed without further geometry optimization. The data for all other methods (2-10) are determined using quantum chemistry package G09.[19] Two density functional theories, b3lyp and wb97xd, are selected for comparison. B3lyp[27] is a widely used global hybrid GGA functional.[16, 17] Wb97xd is a recently developed hybrid GGA functional with corrections to long-range electrostatic interactions and exchange-dispersion interactions.[28, 29] Four basis sets (6-31G(d,p), 6-31G(2d,p), 6-31+G(d,p) and 6-311G(d,p)) are used with each of the density functional to examine the influence of polarization, diffuse, and valence-split functions. In all these calculations (2-8), the molecular geometry is determined in vacuum, with the results in the VT-database taken as initial guess. The COSMO calculations are performed without further geometry optimization. To examine the influence of molecular geometry, we also determined the equilibrium molecular geometry in conductor using b3lyp/6-31G(d,p). This approach is denoted as b3lyp/6-31G(d,p)-cosmo (9). Finally, the semi-empirical method, PM6 (10), is also considered for comparison.

In order to test the performance of COSMO-SAC model among different quantum calculation methods thoroughly, the re-optimization of the parameters in COSMO-SAC model is necessary. First the polarization factor f_{pol} is determined based on internal consistency for the averaging process. The calculation details can be found in literature.[30] Second the parameters in original COSMO-SAC model, a_{eff} and C_{hb}, are fitted to vapor-liquid equilibrium (VLE) data for 1118 binary mixtures. Third the five parameters in revised COSMO-SAC model were optimized in the following order: the A_{ES}, B_{ES} were fitted to LLE of 71 non-hydrogen bond binary mixtures; $C_{OH\text{-}OH}$, $C_{OT\text{-}OT}$ and $C_{OH\text{-}OT}$ were fitted to VLE of 309, 60 and 190 binary mixtures, respectively. The objective functions used in the optimization are as follows.

obj of VLE =

$$\frac{1}{M}\left[\sum_{i=1}^{M}\left(y_i^{calc}-y_i^{expt}\right)^2\right]^{1/2}+\frac{1}{M}\left[\sum_{i=1}^{M}\left(\frac{P^{calc}-P^{expt}}{P^{expt}}\right)^2\right]^{1/2} \quad (13)$$

$$\textbf{obj of LLE}=\frac{1}{N}\left[\sum_{i=1}^{N}\left(x_i^{calc}-x_i^{expt}\right)^2\right]^{1/2} \quad (14)$$

where M is the number of VLE data points; N is the number of LLE data points; the superscripts *calc* and *expt* indicated the calculated and experimental values, respectively. All the VLE and LLE data were retrieved from DECHEMA Chemistry Data Series.[31, 32] The values of the optimal parameters for COSMO-SAC and revised COSMO-SAC model are given in Tables 1 and 2.

The infinite dilution activity coefficient (IDAC, taken from DECHEMA Chemistry Data Series)[33] and the octanol-water partition coefficient (K_{ow}, taken from CRC handbook)[34] are also examined for in this work. The K_{ow} is determined from the IDAC in water-rich phase (which is nearly pure water) and octanol-rich phase (containing approximately 0.725 mole fraction of octanol), respectively,[4]

$$logK_{ow,i}=log\left(\frac{8.37\gamma_i^{W,\infty}}{55.5\gamma_i^{O,\infty}}\right) \quad (15)$$

4. Results and Discussions

4.1. Comparison of the molecular properties among different quantum calculations. Basic molecular properties, such as surface area, volume, and dipole moment, are first examined for a total of 1405 molecules. Figure 1 compares the surface area and molecular volume from different quantum calculation approaches to the values from VWN-BP/DNP, based on which the original COSMO-SAC model was developed. In general, the surface area from G09 is higher than that from DMol3 while the molecular volume from G09 is lower than that from DMol3. Note that the atomic radii used are the same in all these calculations. Generally the slight differences observed here are a result of the differences in equilibrium geometry and the differences in the algorithm for constructing the molecular surface in the two programs. For example, the two apparent outlyers, diethylpentane (S1(a) which depict in Fig. S1(a)) and 1,2-dibromoethane (S1(b) which depict in Fig. S1(b)), are caused by differences in the equilibrium bond lengths.

The dipole moment calculated from the screening charges (as given in eq. 16) is a good indicator for the polarity of the molecule,

$$D=\left|\sum_{i=1}^{n}q_i\underline{r_i}\right| \quad (16)$$

where n is the total number of segments on a molecule, q_i and r_i are the screening charge and position of segment i. Fig. 2(a) compares D calculated from VWN-BP/DNP and b3lyp/6-31G(d,p). Although most of the results are similar from these two methods (data points on the diagonal line), there are quite a few noticeable differences. For example, there are several cases where the molecule is nonpolar (D ~ 0) based on VWN-BP/DNP but is polar based on b3lyp/6-31G(d,p) (data points on the y-axis). Such differences come from the fact that the molecular geometry in VT-database may not be the lowest energy one, but the one that produces the best vapor pressure.[20] Fig. S1(c) and (d) illustrates examples of the different geometries where the energy of the geometry in b3lyp/6-31G(d,p) is lower than that in VT-database.

Fig. 2(b), (c) and (d) compare the dipole moment of other DFT methods with respect to b3lyp/6-31G(d,p). The dipole moment shows fairly good similarities among these DFT methods, except for those based on PM6. In general, the polarity of a molecule calculated using different methods falls in the following order: b3lyp/6-31+G(d,p) > b3lyp/6-31G(d,p)-cosmo > b3lyp/6-31G(d,p) ~ b3lyp/6-311G(d,p) ~ wb97xd/6-31G(d,p) > b3lyp/6-31G(2d,p). The scatter plot in Fig. 2(c) shows that the change either in functional [b3lyp vs. wb97xd] or in split valance [6-31G(d,p) vs. 6-311G(d,p)] has little effect on the dipole moment. However, the inclusion of diffuse functions [6-31+G(d,p)] enhances the molecular dipole, while the use of additional polarization functions [6-31G(2d,p)] reduces the polarity. In addition, geometry optimization in conductor [b3lyp/6-31G(d,p)-cosmo] generally results in a greater molecular polarity. The scattered points in Fig. 2(d) imply that the equilibrium structure in conductor is different from that in vacuum. For example, inositol and malathion have greater exposed hydrogen bonding surface area and higher dipole values based on b3lyp/6-31G(d,p)-cosmo compared to those from b3lyp/6-31G(d,p) (6.46 D vs. 0.55 D for inositol and 9.33 D vs 5.89 D for malathion).

Some points on the x-axis of Fig. 2(d) indicate that the optimal structure of b3lyp/6-31G(d,p)-cosmo is symmetric and thus has zero dipole (see, for example, succinic acid in Fig. S2). The remarkable differences between PM6 and b3lyp/6-31G(d,p) (Fig. 2(d)) implies that the semi-empirical method is less reliable for determining the screening charge distribution.

The influences of different QM/COSMO calculations on the dipole moment also reflect on the σ-profile. Fig. 3 compares the σ-profile among b3lyp/6-31G(d,p), wb97xd/6-31G(d,p) and b3lyp/6-311G(d,p). The result again shows that the use of functional [wb97xd] and split valance [6-311G(d,p)] has little effects on the electronic properties. However the influence in diffuse function [6-31+G(d,p)] and polarization function [6-31G(2d,p)] can be easily seen for polar chemicals such as ethyl acetate, acetonitrile, ethanol and water, shown in Fig. 4. The extreme values in the σ-profile shows the following order: 6-31+G(d,p) > 6-31G(d,p) > 6-31G(2d,p) which agrees with the results from dipole moment discussed earlier. We also found that the σ-profile from PM6 is quite different from other DFT approaches. It again implies that the quality of screening charge distribution from semi-empirical approach may be less reliable than other DFT approaches.

4.2. Performance of original COSMO-SAC model based on different QM/COSMO methods. Table 3 summarizes the prediction accuracy among different quantum approaches using parameters optimized for each method (Table 1). For VLE, the collected experi-

mental data contain 1222 binary mixtures (15491 data points with a temperature range from 207.92 K to 553.15 K); for LLE, 261 binary mixtures (4041 data points with a temperature range from 255.35 K to 458.15 K); IDAC, 2343 data points (238.2 K to 453.2 K), and K_{ow}, 291 species. The re-optimization based on this set of experiment data for VWN-BP/DNP can further improve the prediction accuracy by around 10% for VLE and LLE (VWN-BP/DNP (Original) vs. VWN-BP/DNP (Re-optimized)). Furthermore, the performance of the COSMO-SAC method depends strongly on the changes in σ-profiles. For example, the change of density functional from b3lyp to wb97xd does not improve the accuracy much compared to the same basis set because the σ-profile based on the two methods are quite similar. Also the use of split-valance b3lyp/6-311G(d,p) gives rise to similar results as that from b3lyp/6-31G(d,p) for the same reason. Note that the computational cost from wb97xd and 6-311G(d,p) are higher. Therefore, they are considered not necessary for the original COSMO-SAC model.

On the other hand, the inclusion of polarization function (b3lyp/6-31G(2d,p)), diffuse function (b3lyp/6-31+G(d,p)), or geometry determined in conductor (b3lyp/6-31G(d,p)-cosmo) has a more significant impact on the performance of COSMO-SAC. For VLE, IDAC and K_{ow}, the prediction accuracy are progressively improved in the order of b3lyp/6-31G(2d,p) < b3lyp/6-31G(d,p) < b3lyp/6-31G(d,p)-cosmo < b3lyp/6-31+G(d,p). The accuracy in predicting the liquid composition of LLE shows a reversed order: b3lyp/6-31G(2d,p) > b3lyp/6-31G(d,p) > b3lyp/6-31G(d,p)-cosmo > b3lyp/6-31+G(d,p). However, the number of computable points are significantly reduced for b3lyp/6-31G(2d,p). We found that b3lyp/6-31G(2d,p) generally predicts a lower value for the upper critical solution temperature (UCST). No separation occurs at high temperatures and thus fewer points can be computed. Since none of these methods shows enhancement in both accuracy and number of computable points, we consider it difficult to judge the performance based on the results of LLE calculations.

It is interesting to note that the performance consists with the calculated polarity as discuss in previous section. The method giving the highest polarity of a molecule shows the best result in VLE, IDAC and K_{ow} prediction. We thus suggest b3lyp/6-31+G(d,p) to be the best choice among the 10 different QM/COSMO methods.

The performance of COSMO-SAC based on PM6 in VLE, LLE, IDAC and K_{ow} are much worse than that based on other DFT calculations. The error in VLE are almost twice higher, the computable systems in LLE are reduced significantly, and the accuracy in IDAC and K_{ow} both are inferior compared to others. It's also worth mentioning that although mod-UNIFAC can provide very accurate result for VLE, its performance in LLE is similar to that from the COSMO-SAC model. Furthermore, there are 94, 25, 15 and 43 systems in VLE, LLE, IDAC and K_{ow} where the mod-UNIFAC model is not available due to the lack of necessary group interaction parameters.

One obvious deficiency of the original COSMO-SAC model is unavoidable undesired hydrogen bonding interactions. For example, the hydrogen in chloroform is highly polarized and is considered as hydrogen bonding donor in the original COSMO-SAC model. Therefore, there are hydrogen bonding interactions between chloroform and acceptors such as ketone or ether. From Fig. 5 it can be seen that the IDAC for such mixtures are underestimated, indicating too strong interaction between chloroform and the other component. However, when we purposely turn off the hydrogen bonding interactions (by setting C_{hb} = 0), the accuracy improve significantly. Such problems are resolved in the revised COSMO-SAC model where the hydrogen bonding interactions are determined based on the atom types, in additional to the surface charge densities.

4.3. Performance of revised COSMO-SAC model based on different QM/COSMO methods. Here the performance of the revised COSMO-SAC model are analyzed based on VWN-BP/DNP, b3lyp/6-31G(d,p), b3lyp/6-31G(2d,p), b3lyp/6-31+G(d,p), b3lyp/6-311G(d,p), b3lyp/6-31G(d,p)-cosmo and PM6. Table 4 summarizes the performance for various equilibrium calculations based on optimized model parameters for each method (Table 2). Unlike the original COSMO-SAC model where there is a significant dependence of performance on QM/COSMO method, the performance of the revised COSMO-SAC model from different QM/COSMO methods are quite similar. The results show that the two major changes introduced (temperature-dependent non-hydrogen bonding interaction and atom-type dependent hydrogen bonding interactions) not only improves the accuracy of the method, but also allows for the differences in the σ-profiles due to the use of QM/COSMO methods be absorbed in the additional model parameters. We consider this a merit of the revised COSMO-SAC since high accuracy predictions can be achieved without going to high level DFT theory and large basis sets (computationally demanding).

The σ-profiles determined from b3lyp/6-31G(d,p)-cosmo are usually more polarized (as the dipole moment shown in Fig. 2(d)) even when the equilibrium structures are similar. This has a significant implication on the calculation of logK_{ow}, as shown in Fig. 6. The high polarity of the molecules based on b3lyp/6-31G(d,p)-cosmo results in increase of K_{ow} for non-polar chemicals (in the positive region) and decrease of K_{ow} for polar species (mostly in the negative region) compared to those based on b3lyp/6-31G(d,p). Table 5 compares the performance in K_{ow} for species whose dipole moment changes by more than 1 D between vacuum and conductor geometries. The large differences in dipole moment imply that the equilibrium structures determined in vacuum and in conductor can be quite different. The results clear shows that K_{ow} is strongly affected by the structures (see Fig. S2 for the molecular geometry).

It is quite surprising that the semi-empirical PM6 provides rather accurate predictions for IDAC. This is mainly a result of the (possibly lucky) better description of infi-

nite dilution of water or alcohol in alkane. In Fig. 7(a), it can be seen that b3lyp/6-31G(d,p) overestimates the IDAC of alcohol and water in alkane while b3lyp/6-31+G(d,p) and PM6 surprisingly describe it much better. However for other mixtures, b3lyp/6-31G(d,p) underestimates and b3lyp/6-31G+(d,p) overestimates the experimental data, as shown in Fig. 7(b). Except for water/alcohol in alkanes, all the DFT methods show similar accuracy and PM6 behaves worse than other DFT methods (see Table 6), as seen for VLE, LLE and K_{ow}.

Although PM6 is less accurate in the revised COSMO-SAC model, it still has a great potential for large molecules because the other DFT calculation method require more computational effort. For example, the COSMO computational time for Methyloleate ($C_{19}H_{36}O_2$) based on b3lyp/6-31G(d,p) is about six times longer than that based on PM6 (see Table S5). In summary, the best choice we found in this work for the revised COSMO-SAC model is the b3lyp/6-31G(d,p)-cosmo.

Table 7 compares the performance of the original COSMO-SAC model and the revised COMOS-SAC model based on the same QM/COSMO method. It's obvious that the revised COSMO-SAC outperforms the original model in VLE and in most cases, in LLE, IDAC, and K_{ow} predictions. While the RMS in LLE predictions are similar from the two methods, the number of computable points are significantly greater. The detailed performance of COSMO-SAC model in IDAC is summarized in Table S6. The revised model is found to be less accurate than the original one in two situations. First, the IDAC of water or alcohol in alkane (see column IDAC^b in Table 7); Second, the VLE of type III systems (hydrogen bonding between unlike species, compare Tables S1 and S3 in Supporting Information). In both cases, the mixtures involve hydrogen bonding interactions. It is not clear why the revised model is less accurate in these two situations. It is possible that the modification of the hydrogen bonding interaction may be physically less rigor in the revised model.

5. Conclusions

The performance of the original and revised COSMO-SAC models based on different QM/COSMO methods are analyzed using re-optimized model parameters for each method. It is found that the original COSMO-SAC model is sensitive to the choice of QM/COSMO method, in particular the choice of basis set. The method giving the highest polarity of a chemical often results in a better performance (e.g., inclusion of diffuse functions or geometry optimization in conductor). On the contrary, the performance of the revised COSMO-SAC model is not very dependent on the choice of QM/COSMO method, provided that the model parameters are re-optimized. As a compromise between accuracy and computational effort, b3lyp/6-31G-cosmo is recommended as the best QM/COSMO method for both the original and revised COSMO-SAC models. The errors from the original COSMO-SAC model using semi-empirical PM6 is about twice of those based on DFT methods. The difference in performance from the revised COSMO-SAC model for PM6 and other DFT methods are significantly reduced (by about 50%). Therefore, while less accurate, the revised COSMO-SAC model based on PM6 can be a computationally viable approach for problems involving large molecules.

SUPPORTING INFORMATION

The Supporting Information is available free of charge on the ACS Publications website. The Supporting Information include the description for the charge-average process, molecular structures from different quantum calculation methods, detailed comparisons on VLE, LLE and K_{ow}.

Acknowledgement

This research was partially supported by the Ministry of Science and Technology of Taiwan (MOST 104-2221-E-002-186-MY3 and MOST 103-2218-E-008-003-MY2), National Taiwan University (NTU-CDP-105R7876), and LCY Chemical Corp. Dr. Li Yang would also like to acknowledge the support from the National Natural Science Foundation of China (No. 21406172). The computation resources from the National Center for High-Performance Computing of Taiwan and the Computing and Information Networking Center of the National Taiwan University are acknowledged.

Reference

1. Lin, S. T.; Sandler, S. I., A priori phase equilibrium prediction from a segment contribution solvation model. *Industrial & Engineering Chemistry Research* **2002,** 41, (5), 899-913.
2. Klamt, A., Conductor-Like Screening Model for Real Solvents - a New Approach to the Quantitative Calculation of Solvation Phenomena. *Journal of Physical Chemistry* **1995,** 99, (7), 2224-2235.
3. Hsieh, C.-M.; Sandler, S. I.; Lin, S.-T., Improvements of COSMO-SAC for vapor–liquid and liquid–liquid equilibrium predictions. *Fluid Phase Equilibria* **2010,** 297, (1), 90-97.
4. Hsieh, C.-M.; Wang, S.; Lin, S.-T.; Sandler, S. I., A Predictive Model for the Solubility and Octanol-Water Partition Coefficient of Pharmaceuticals. *Journal of Chemical and Engineering Data* **2011,** 56, (4), 936-945.
5. Shu, C. C.; Lin, S. T., Prediction of Drug Solubility in Mixed Solvent Systems Using the COSMO-SAC Activity Coefficient Model. *Industrial & Engineering Chemistry Research* **2011,** 50, (1), 142-147.
6. Wang, L.-H.; Lin, S.-T., A predictive method for the solubility of drug in supercritical carbon dioxide. *Journal of Supercritical Fluids* **2014,** 85, 81-88.
7. Lee, B. S.; Lin, S. T., Screening of ionic liquids for CO_2 capture using the COSMO-SAC model. *Chemical Engineering Science* **2015,** 121, 157-168.
8. Lin, S.-T.; Wang, L.-H.; Chen, W.-L.; Lai, P.-K.; Hsieh, C.-M., Prediction of miscibility gaps in water/ether mixtures using COSMO-SAC model. *Fluid Phase Equilibria* **2011,** 310, (1-2), 19-24.
9. Fredenslund, A.; Gmehling, J.; Michelsen, M. L.; Rasmussen, P.; Prausnitz, J. M., Computerized Design of Multicomponent Distillation Columns Using the UNIFAC Group Contribution Method for Calculation of Activity Coefficients. *Industrial & Engineering Chemistry Process Design and Development* **1977,** 16, (4), 450-462.

10. Gmehling, J.; Li, J. D.; Schiller, M., A MODIFIED UNIFAC MODEL .2. PRESENT PARAMETER MATRIX AND RESULTS FOR DIFFERENT THERMODYNAMIC PROPERTIES. *Industrial & Engineering Chemistry Research* **1993,** 32, (1), 178-193.
11. Abrams, D. S.; Prausnitz, J. M., Statistical Thermodynamics of Liquid-Mixtures - New Expression for Excess Gibbs Energy of Partly or Completely Miscible Systems. *Aiche Journal* **1975,** 21, (1), 116-128.
12. Lin, S. T.; Hsieh, M. K.; Hsieh, C. M.; Hsu, C. C., Towards the development of theoretically correct liquid activity coefficient models. *Journal of Chemical Thermodynamics* **2009,** 41, (10), 1145-1153.
13. Flemr, V., A Note on Excess Gibbs Energy Equations Based on Local Composition Concept. *Collection of Czechoslovak Chemical Communications* **1976,** 41, (11), 3347-3349.
14. Klamt, A.; Schuurmann, G., COSMO - a New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *Journal of the Chemical Society-Perkin Transactions 2* **1993**, (5), 799-805.
15. Lin, S. T.; Hsieh, C. M., Efficient and accurate solvation energy calculation from polarizable continuum models. *Journal of Chemical Physics* **2006,** 125, (12).
16. Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C., Electronic structure calculations on workstation computers: The program system turbomole. *Chemical Physics Letters* **1989,** 162, (3), 165-169.
17. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A., General atomic and molecular electronic structure system. *Journal of Computational Chemistry* **1993,** 14, (11), 1347-1363.
18. Delley, B., From molecules to solids with the DMol(3) approach. *Journal of Chemical Physics* **2000,** 113, (18), 7756-7764.
19. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J., Gaussian 09, Revision B.01. In Wallingford CT, 2009.
20. Mullins, E.; Oldland, R.; Liu, Y. A.; Wang, S.; Sandler, S. I.; Chen, C. C.; Zwolak, M.; Seavey, K. C., Sigma-profile database for using COSMO-based thermodynamic methods. *Industrial & Engineering Chemistry Research* **2006,** 45, (12), 4389-4415.
21. Mu, T. C.; Rarey, J.; Gmehling, J., Performance of COSMO-RS with sigma profiles from different model chemistries. *Industrial & Engineering Chemistry Research* **2007,** 46, (20), 6612-6629.
22. Franke, R.; Hannebauer, B., On the influence of basis sets and quantum chemical methods on the prediction accuracy of COSMO-RS. *Physical Chemistry Chemical Physics* **2011,** 13, (48), 21344-21350.
23. Paulechka, E.; Diky, V.; Kazakov, A.; Kroenlein, K.; Frenkel, M., Reparameterization of COSMO-SAC for Phase Equilibrium Properties Based on Critically Evaluated Data. *Journal of Chemical & Engineering Data* **2015,** 60, (12), 3554-3561.
24. Lin, S. T.; Sandler, S. I., Infinite dilution activity coefficients from ab initio solvation calculations. *Aiche Journal* **1999,** 45, (12), 2606-2618.
25. Ben-Naim, A., *Solvation thermodynamics*. Plenum Press: 1987.
26. Wang, S.; Sandler, S. I.; Chen, C.-C., Refinement of COSMO–SAC and the Applications. *Industrial & Engineering Chemistry Research* **2007,** 46, (22), 7275-7288.
27. Becke, A. D., DENSITY-FUNCTIONAL THERMOCHEMISTRY .3. THE ROLE OF EXACT EXCHANGE. *Journal of Chemical Physics* **1993,** 98, (7), 5648-5652.
28. Chai, J.-D.; Head-Gordon, M., Systematic optimization of long-range corrected hybrid density functionals. *Journal of Chemical Physics* **2008,** 128, (8), 084106.
29. Chai, J.-D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008,** 10, (44), 6615-6620.
30. Lin, S.-T.; Chang, J.; Wang, S.; Goddard, W. A.; Sandler, S. I., Prediction of Vapor Pressures and Enthalpies of Vaporization Using a COSMO Solvation Model. *The Journal of Physical Chemistry A* **2004,** 108, (36), 7429-7439.
31. Gmehling, J. r.; Onken, U.; Arlt, W., *Vapor-liquid equilibrium data collection*. Dechema: Frankfurt, 1977.
32. Sørensen, J. M.; Arlt, W., *Liquid-liquid equilibrium data collection*. DECHEMA: Frankfurt, 1979.
33. Tiegs, D., *Activity coefficients at infinite dilution*. DECHEMA: Frankfurt, 1986.
34. Haynes, W. M.; Lide, D. R.; Bruno, T. J., *CRC handbook of chemistry and physics : a ready-reference book of chemical and physical data*. 2015.
35. Gmehling, J.; Lohmann, J.; Jakob, A.; Li, J. D.; Joh, R., A modified UNIFAC (Dortmund) model. 3. Revision and extension. *Industrial & Engineering Chemistry Research* **1998,** 37, (12), 4876-4882.
36. Tetko, I. V.; Tanchuk, V. Y., Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *Journal of Chemical Information and Computer Sciences* **2002,** 42, (5), 1136-1145.
37. Wang, Y. L.; Xiao, J. W.; Suzek, T. O.; Zhang, J.; Wang, J. Y.; Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* **2009,** 37, W623-W633.

Table 1. Parameters for the original COSMO-SAC model1 based on different QM/COSMO methods

Parameter (unit) / Method	f_{pol} (-)	a_{eff} (Å²)	C_{hb} (kcal/mol Å⁴/e²)
VWN-BP/DNP (Original)[a]	0.6916	7.5000	92992
VWN-BP/DNP (Re-optimized)	0.6916	6.8036	53160
b3lyp/6-31G(d,p)	0.7285	6.6985	56891
b3lyp/6-31G(2d,p)	0.7167	6.5490	64916
b3lyp/6-31+G(d,p)	0.6992	5.8661	34380
b3lyp/6-311G(d,p)	0.7227	6.3460	50276
wb97xd/6-31G(d,p)	0.7344	6.5232	50060
wb97xd/6-31G(2d,p)	0.7223	6.3815	57439
wb97xd/6-31+G(d,p)	0.7064	5.7778	32878
b3lyp/6-31G(d,p)-cosmo	0.7217	6.4547	55946
PM6	0.7450	4.6641	23165

[a]the values for f_{pol} and C_{hb} are taken from literature[30] and are slightly different from the ones in the original paper[1]

Table 2. Parameters for the revised COSMO-SAC model[3] based on different QM/COSMO methods

Parameter (unit) / Method	a_{eff} (Å²)	A_{ES} (kcal/mol Å⁴/e²)	B_{ES} (kcal/mol Å⁴/e² K²)	C_{OH-OH} (kcal/mol Å⁴/e²)	C_{OT-OT} (kcal/mol Å⁴/e²)	C_{OH-OT} (kcal/mol Å⁴/e²)
VWN-BP/DNP (Original)	7.2500	6525.69	1.4859E+08	4013.78	932.31	3016.43
VWN-BP/DNP (Re-optimized)	6.8036	6466.02	1.5026E+08	4256.00	1032.11	3632.89
b3lyp/6-31G(d,p)	6.6985	6509.36	1.3817E+08	5211.64	1336.00	4475.46
b3lyp/6-31G(2d,p)	6.5490	6668.44	1.5301E+08	5970.49	1479.21	4919.38
b3lyp/6-31+G(d,p)	5.8661	4869.68	8.1138E+07	2487.23	1181.26	2272.73
b3lyp/6-311G(d,p)	6.3460	5822.07	1.3981E+08	4514.29	1283.33	3926.52
b3lyp/6-31G(d,p)-cosmo	6.4547	5501.09	1.8822E+08	4900.06	1378.62	4211.67
PM6	4.6641	3391.95	2.2536E+08	2419.32	1573.16	1645.85

Table 3. Prediction accuracy of the original COSMO-SAC model in vapor-liquid equilibrium, liquid-liquid Equilibrium, infinite dilution activity coefficients (IDAC) and octanol-water partition coefficient (K_{ow}).

Accuracy / Method	VLE		LLE		IDAC	K_{ow}
	AARD-P (%)[a]	AAD-y (%)[b]	RMS-x_i[c]	N_{sys}[d]	RMS-lnγ_i[e]	RMS-logK_{ow}[f]
VWN-BP/DNP (Original)	7.57	2.94	0.1225	202	0.964	0.879
VWN-BP/DNP (Re-optimized)	6.67	2.54	0.1017	215	0.695	0.513
b3lyp/6-31G(d,p)	6.58	2.49	0.1045	213	0.817	0.693
b3lyp/6-31G(2d,p)	7.02	2.64	0.1037	200	1.035	0.768
b3lyp/6-31+G(d,p)	5.79	2.32	0.1211	226	0.649	0.512
b3lyp/6-311G(d,p)	6.61	2.50	0.1016	216	0.761	0.624
wb97xd/6-31G(d,p)	6.94	2.61	0.1048	216	0.825	0.662
wb97xd/6-31G(2d,p)	6.92	2.59	0.1029	208	0.822	0.734
wb97xd/6-31+G(d,p)	6.12	2.41	0.1276	226	0.672	0.533
b3lyp/6-31G(d,p)-cosmo	6.51	2.48	0.1058	215	0.748	0.603
PM6	11.19	4.30	0.1651	175	1.079	0.784
mod-UNIFAC[35] [g]	2.58	1.26	0.1016	223	0.876	0.761
Number of systems	1222		261		2343	291

[a] $\mathrm{AARD}-P(\%) = \left|P_j^{exp} - P_j^{cal}\right| / P_j^{exp} \times 100\%$; [b] $\mathrm{AAD}-y(\%) = \left|y_{exp} - y_{cal}\right| / 1 \times 100\%$

[c] $\mathrm{RMS\text{-}xi} = \frac{1}{N_{sys}} \sum_{i=1}^{N_{sys}} \left[\frac{1}{N_i} \sum_{j=1}^{N_i} \left(x_j^{calc} - x_j^{expt}\right)^2\right]^{1/2}$ where N_{sys} is the number of systems and N_i is the number of data points in i-th system.

[d] N_{sys} represents the number of computable systems and the total number of experimental systems is 261.

[e] $\mathrm{RMS}-\ln\gamma_i = \left[\frac{1}{N} \sum_{i=1}^{N} \left(ln\gamma_i^{calc} - ln\gamma_i^{expt}\right)^2\right]^{\frac{1}{2}}$; [f] $\mathrm{RMS}-\log K_{ow}(\%) = \left[\frac{1}{N} \sum_{i=1}^{N} \left(\log K_{ow}^{calc} - \log K_{ow}^{expt}\right)^2\right]^{\frac{1}{2}}$

[g] there are 94, 25, 15 and 43 fewer systems that can be calculated for VLE, LLE, IDAC and K_{ow} because of missing parameters

Table 4. Prediction accuracy of the revised COSMO-SAC model in vapor-liquid equilibrium, liquid-liquid Equilibrium, infinite dilution activity coefficients (IDAC) and octanol-water partition coefficient (K_{ow}).

Accuracy / Method	VLE		LLE		IDAC	K_{ow}
	AARD-P (%)[a]	AAD-y (%)[b]	RMS-x_i[c]	N_{sys}[d]	RMS-lnγ_i[e]	RMS-logK_{ow}[f]
VWN-BP/DNP (Original)	5.78	2.26	0.0899	223	0.762	0.464
VWN-BP/DNP (Re-optimized)	5.30	2.19	0.1026	226	0.884	0.576
b3lyp/6-31G(d,p)	5.30	2.15	0.1027	223	0.977	0.511
b3lyp/6-31G(2d,p)	5.29	2.14	0.0996	217	0.958	0.507
b3lyp/6-31+G(d,p)	5.29	2.13	0.1036	227	0.751	0.665
b3lyp/6-311G(d,p)	5.36	2.18	0.1061	228	0.941	0.524
b3lyp/6-31G(d,p)-cosmo	5.21	2.13	0.1071	230	0.936	0.505
PM6	7.90	2.87	0.1338	223	0.749	0.788
mod-UNIFAC[35g]	2.58	1.26	0.1016	223	0.876	0.761
Number of system	1222		261		2343	291

[a] $\mathrm{AARD}-P(\%) = \left|P_j^{exp} - P_j^{cal}\right| / P_j^{exp} \times 100\%$; [b] $\mathrm{AAD}-y(\%) = \left|y_{exp} - y_{cal}\right| / 1 \times 100\%$

[c] $\mathrm{RMS\text{-}xi} = \frac{1}{N_{sys}} \sum_{i=1}^{N_{sys}} \left[\frac{1}{N_i} \sum_{j=1}^{N_i} \left(x_j^{calc} - x_j^{expt}\right)^2\right]^{1/2}$ where N_{sys} is the number of systems and N_i is the number of data points in i-th system.

[d] N_{sys} represent the number of computable systems and the total number of experimental systems is 261.

[e] $\mathrm{RMS}-\ln\gamma_i = \left[\frac{1}{N} \sum_{i=1}^{N} \left(ln\gamma_i^{calc} - ln\gamma_i^{expt}\right)^2\right]^{\frac{1}{2}}$, [f] $\mathrm{RMS}-\log K_{ow}(\%) = \left[\frac{1}{N} \sum_{i=1}^{N} \left(\log K_{ow}^{calc} - \log K_{\mathrm{ow}}^{expt}\right)^2\right]^{\frac{1}{2}}$

[g] there are 94, 25, 15 and 43 fewer systems that can be calculated for VLE, LLE, IDAC and Kow because of missing parameters

Table 5. Comparison of dipole moment (in Debye) and logK_{ow} prediected from revised COSMO-SAC

Compound	b3lyp/6-31G(d,p)		b3lyp/6-31G(d,p)-cosmo		logK_{ow} literature
	Dipole	logK_{ow}	Dipole	logK_{ow}	
Inositol	0.55	-0.66	6.46	-2.08	-2.59[a]
Malathion	5.89	3.86	9.33	3.36	2.36[b]
p-Phenylenediamine	0.01	0.31	3.06	0.39	-0.25[b]
Benzoyl peroxide	0.00	2.86	2.15	3.38	3.46[b]
Benzidine	0.80	2.06	2.68	2.25	1.34[b]
m-Phenylenediamine	1.58	0.46	3.25	0.56	-0.33[b]
Maleic acid	0.70	-0.76	2.32	-0.94	-0.48[b]
2,4-Diaminotoluene	3.12	0.98	1.52	1.07	0.14[b]
Sorbitol	2.02	-1.78	3.41	-2.34	-2.2[b]
p-Nitroaniline	10.06	0.64	11.29	0.54	1.39[b]
Diethyl succinate	1.56	1.37	0.53	1.28	1.25[a]
Methyl chloroacetate	2.26	0.48	1.25	0.28	0.63[b]

[a] the estimated value is determined from ALOGPS 2.1 program using the atom contribution method[36]

[b] the experimental data is retrieved from PubChem[37]

Table 6. Comparison of Prediction Accuracy for IDAC from the revised COSMO-SAC model based on different QM/COSMO methods and modified UNIFAC.

Method \ RMS-lnγ_i[a]	water or alcohols in alkanes	Other systems	Overall
VWN-BP/DNP (Original)	3.569	0.551	0.762
VWN-BP/DNP (Re-optimized)	4.738	0.537	0.884
b3lyp/6-31G(d,p)	5.522	0.533	0.977
b3lyp/6-31G(2d,p)	5.404	0.525	0.958
b3lyp/6-31+G(d,p)	3.348	0.568	0.751
b3lyp/6-311(d,p)	5.286	0.521	0.941
b3lyp/6-31G(d,p)-cosmo	5.349	0.497	0.936
PM6	1.291	0.732	0.749
mod-UNIFAC[35b]	1.446	0.858	0.876
Numbers of Data Points	52	2291	2343

[a] RMS = $\left[\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{ln\gamma_i^{calc}} - \boldsymbol{ln\gamma_i^{expt}}\right)^2\right]^{1/2}$

[b] the computable data in mod-UNIFAC is 52 and 2276 in the two types of systems respectively (2328 in total)

Table 7. Difference in the prediction accuracy between two versions of the COSMOSAC model (positive values indicate the revised model being more accurate)

Method \ Difference[a]	VLE		LLE		IDAC[b]	IDAC[c]	K_{ow}
	AARD-*P* (%)	AAD-*y* (%)	RMS	N_{sys}[d]	RMS-lnγ_i	RMS-lnγ_i	RMS-logK_{ow}
VWN-BP/DNP (Original)	1.80	0.68	0.0326	21	-0.073	0.269	0.415
VWN-BP/DNP (Re-optimized)	1.36	0.35	-0.0010	11	-2.923	0.110	-0.063
b3lyp/6-31G(d,p)	1.28	0.34	0.0019	10	-3.545	0.237	0.182
b3lyp/6-31G(2d,p)	1.73	0.50	0.0041	17	-4.402	0.510	0.261
b3lyp/6-31+G(d,p)	0.50	0.18	0.0176	1	-0.661	-0.052	-0.154
b3lyp/6-311G(d,p)	1.25	0.31	-0.0046	12	-3.191	0.181	0.100
b3lyp/6-31G(d,p)-cosmo	1.30	0.35	-0.0013	15	-3.361	0.198	0.098
PM6	3.28	1.42	0.0313	48	0.590	0.322	-0.003

[a] positive values indicate the revised model being more accurate, negative values indicate the original model being more accurate

[b] the IDAC systems when water or alcohol is infinite dilute in alkane (cf. Table 6)

[c] the IDAC systems which exclude the system when water or alcohol is infinite dilute in alkane (cf. Table 6)

[d] the additional number of computable system in LLE using the revised COSMOSAC model compared to that using the original COSMOSAC model

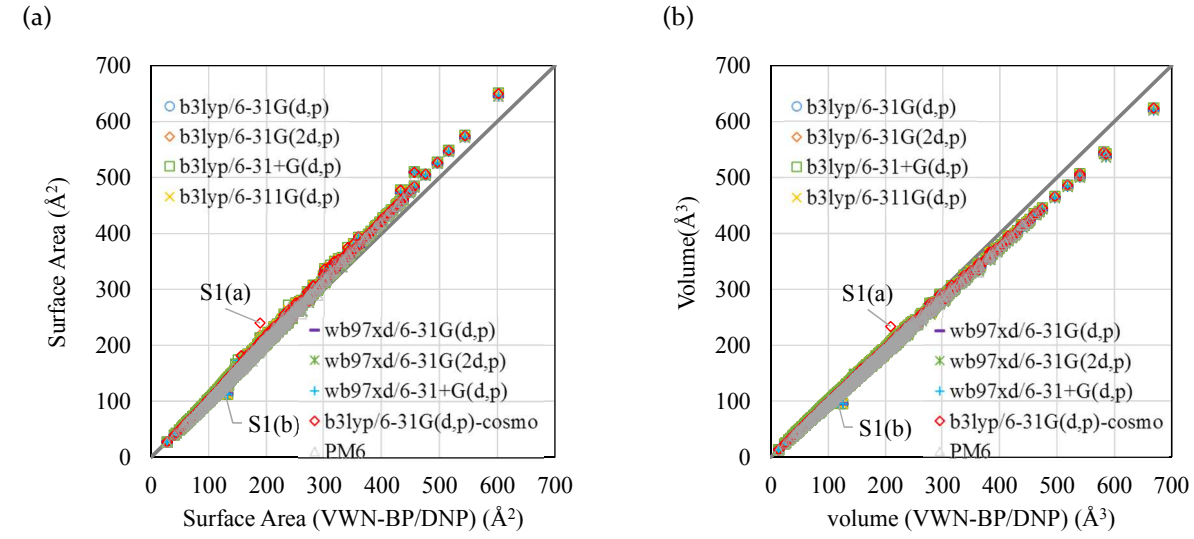

Fig. 1. Comparison of the surface area (a) and molecular volume (b) of 1405 molecules from 9 different QM/COSMO methods to those from VWN-BP/DNP.

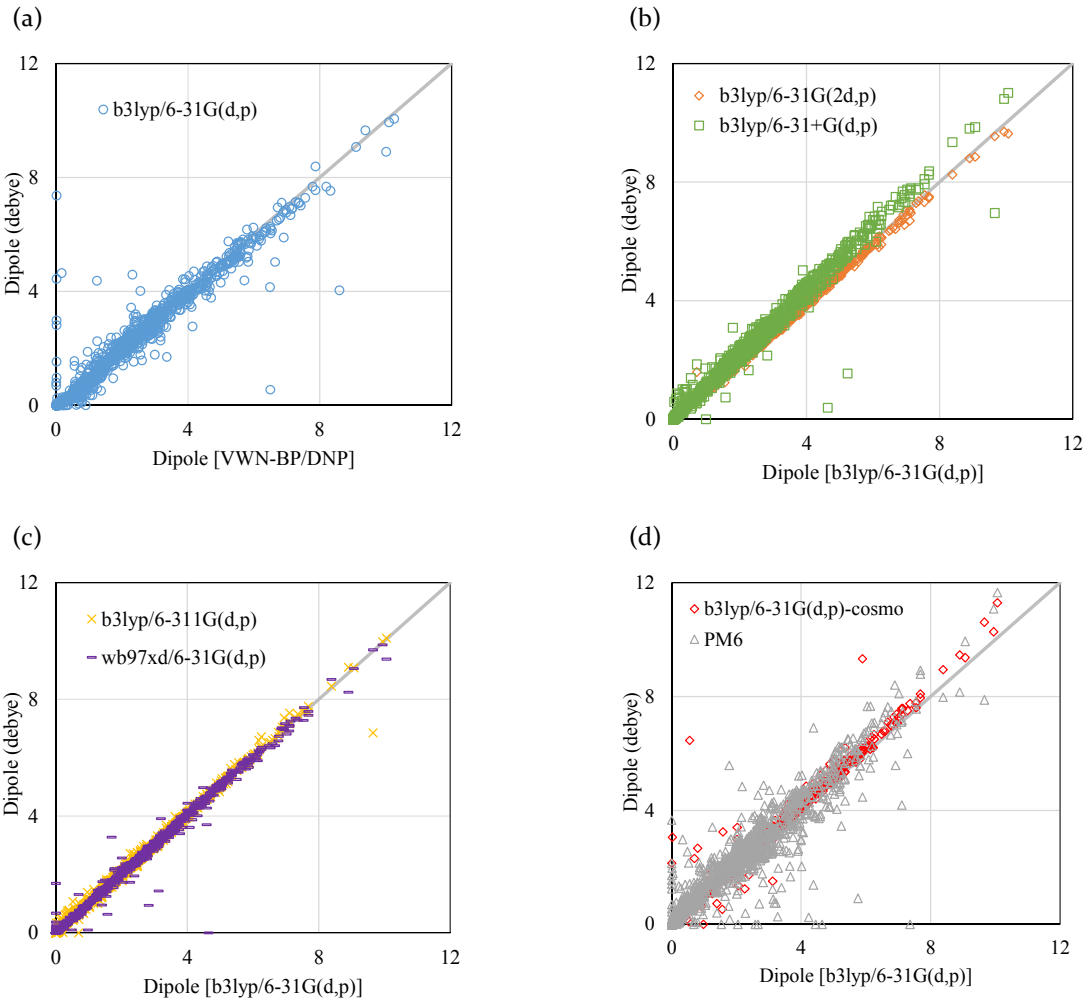

Fig. 2. Comparison of the dipole moment (eq. 16) of 1405 molecules determined from different QM/COSMO methods: (a) b3lyp/6-31G(d,p) vs. VWN-BP/DNP; (b) b3lyp/6-31G(2d,p) and b3lyp/6-31G+(d,p) vs. b3lyp/6-31G(d,p); (c) b3lyp/6-311G(d,p) and wb97xd/6-31G(d,p) vs. b3lyp/6-31G(d,p); (d) b3lyp/6-31G(d,p)-cosmo and PM6 vs. b3lyp/6-31G(d,p).

(a) ethane (b) chloroform

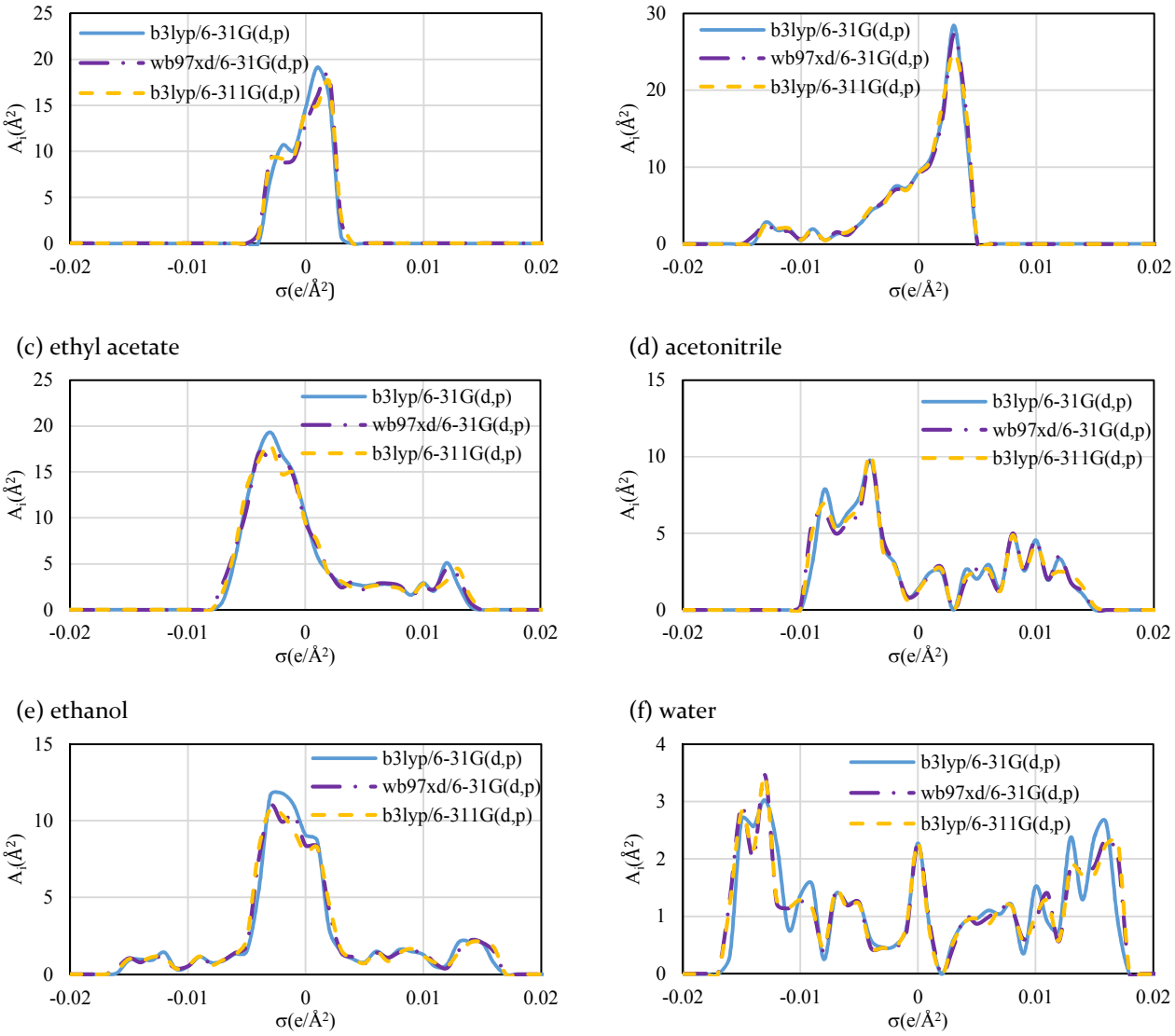

Fig. 3. Comparison of σ-profile of selected compounds ((a) ethane, (b) chloroform, (c) ethyl acetate, (d) acetonitrile, (e) ethanol, (f) water) from different QM/COSMO methods: b3lyp/6-31G(d,p), wb97xd/6-31G(d,p) and b3lyp/6-311G(d,p).

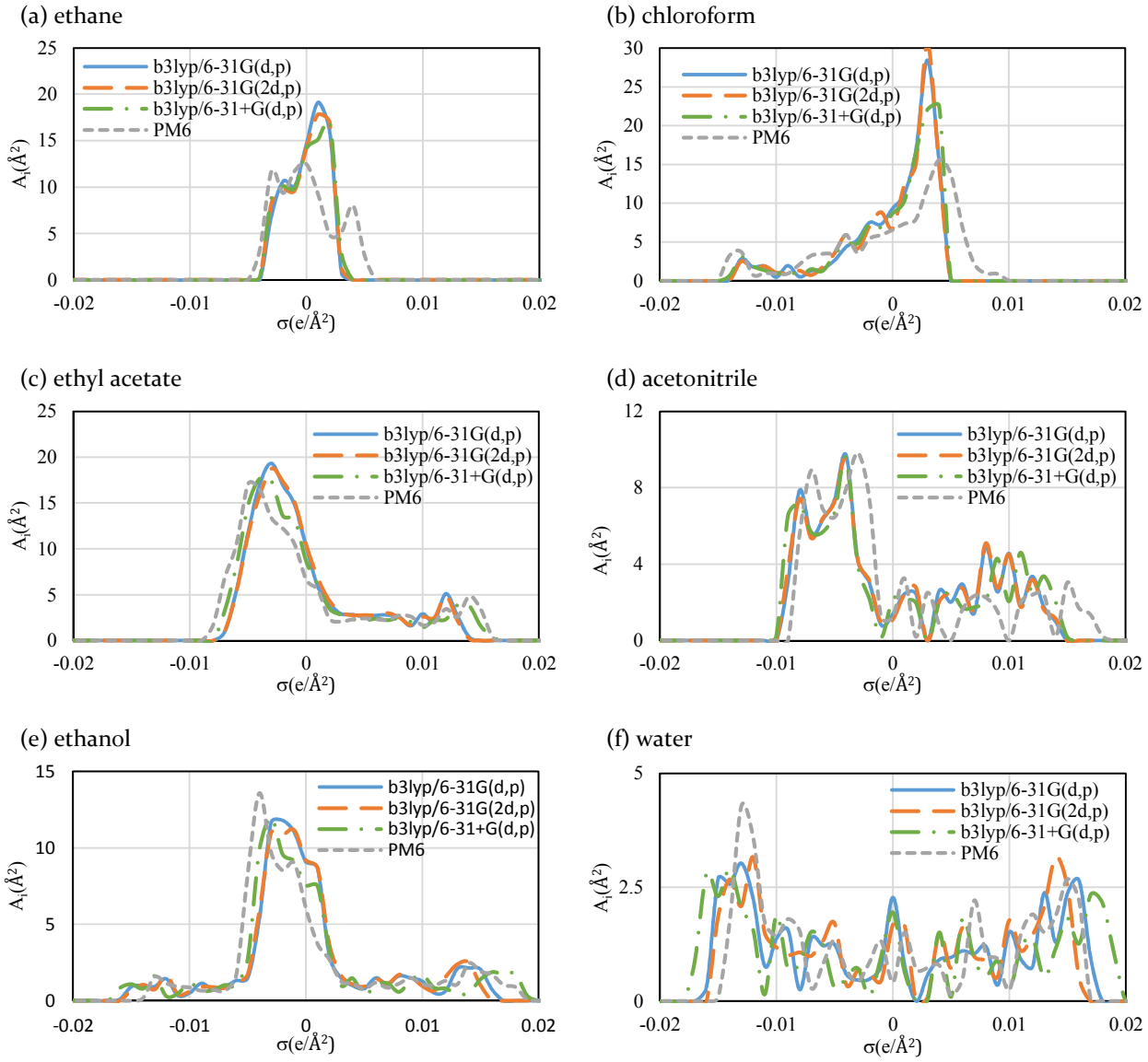

Fig. 4. Comparison of σ-profile of selected compounds ((a) ethane, (b) chloroform, (c) ethyl acetate, (d) acetonitrile, (e) ethanol, (f) water) from different QM/COSMO methods: b3lyp/6-31G(d,p), b3lyp/6-31G(2d,p), b3lyp/6-31+G(d,p) and PM6.

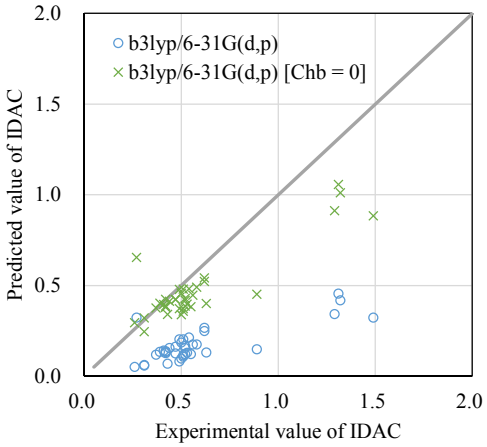

Fig. 5. Comparison of the IDAC systems, chloroform and molecules containing hydrogen-bond acceptor, between the original COSMO-SAC model (open circles) and the original COSMO-SAC model without any hydrogen bond interaction (crosses, by setting C_{hb} = 0)

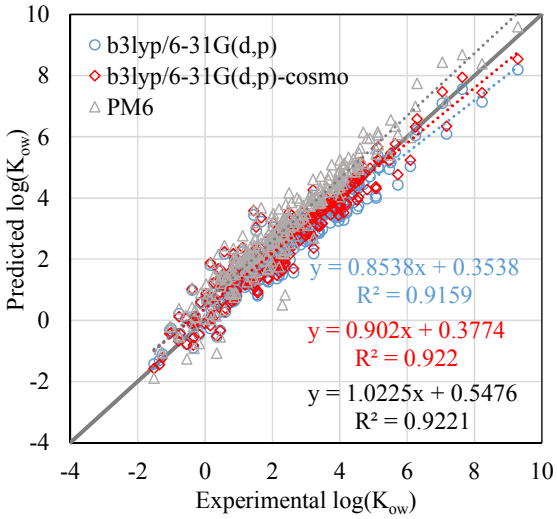

Fig. 6. The comparison of logK_{ow} predicted from the revised COSMO-SAC model with experiments. The calculations are based on b3lyp/6-31G(d,p) (open circles), b3lyp/6-31G(d,p)-cosmo (open diamonds), and PM6 (open triangles).

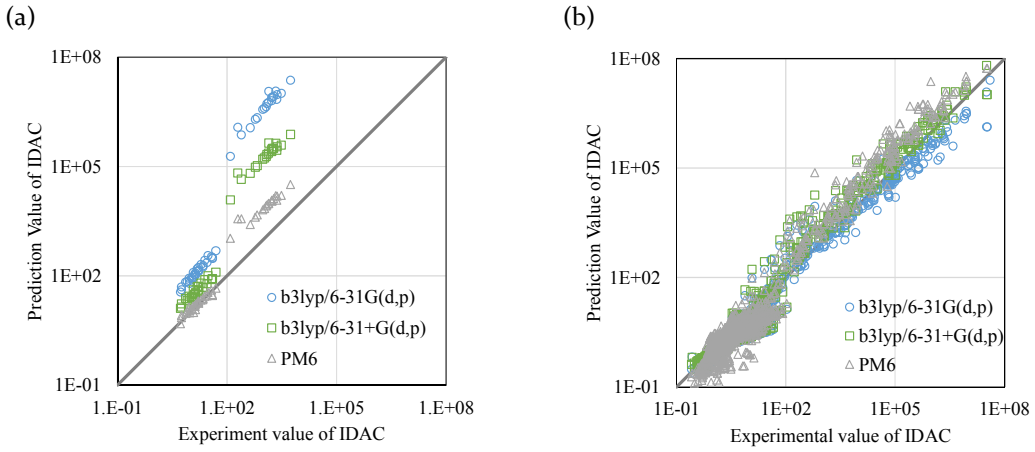

Fig. 7. Comparison of the infinite dilution activity coefficient prediction from revised COSMO-SAC based on b3lyp/6-31G(d,p) (open circles), b3lyp/6-31+G(d,p) (open squares), and PM6 (open triangles) with experiment for (a) water or alcohol being infinite dilute in alkane and (b) all other systems.

Authors are required to submit a graphic entry for the Table of Contents (TOC) that, in conjunction with the manuscript title, should give the reader a representative idea of one of the following: A key structure, reaction, equation, concept, or theorem, etc., that is discussed in the manuscript. Consult the journal's Instructions for Authors for TOC graphic specifications.

Insert Table of Contents artwork here