AIChE Journal
DOI 10.1002/aic.16194

An open and extensible sigma-profile database for COSMO-based models

F. Ferrarini, G. B. Flôres, A. R. Muniz, and R. de P. Soares*

Laboratório Virtual de Predição de Propriedades - LVPP, Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, Rua Ramiro Barcelos, 2777, Bairro Santana, Porto Alegre - RS, Brazil, ZIP 90035-007

E-mail: rafael.pelegrini@ufrgs.br

Abstract

COSMO-based activity coefficient models have become an interesting alternative to the prediction of the behavior of substances in mixture. These models depend on the pure substance information known as sigma-profile. The present work aims to create and freely distribute a sigma-profile database for a wide range of molecules using the GAMESS software. Different quantum chemistry theories for the calculation of the electronic structure and basis sets were tested, in order to define an accurate and efficient approach for computing the profiles. To check the accuracy of the profiles obtained, infinite dilution activity coefficients calculated by the COSMO-SAC model were compared with experimental data for binary mixtures. Among the studied alternatives, the Hartree-Fock method with the TZVP basis set led to results with the best agreement with the experimental data, within a reasonable computational cost. Using this approach, a database was created based solely free computational tools.

Introduction

The design of separation processes in industry requires the accurate prediction of the phase equilibrium behavior of the substances involved[1]. Activity coefficient models, or excess Gibbs energy models, are typically used to quantify the non-ideality of the liquid phase mixtures, attributed to

Received: Aug 29, 2017; Revised: Feb 19, 2018; Accepted: Apr 14, 2018

differences in size, shape, and interaction between the species. A variety of methods have been traditionally used, *e.g.*, NRTL[2], Wilson[3], and UNIFAC[4]. More recently the so called COSMO-based models[5,6] have emerged as an interesting alternative. These models have a stronger predictive power, when compared to the traditional models, since they rely on computational quantum chemistry calculations. They allow determining mixture properties without using experimental data[7].

One example of COSMO-based method is the COSMO-SAC[6] model, a variant of the original COSMO-RS model[5]. This model has many applications in complex systems, such as the prediction behavior of polymer solution and ionic liquids[8–10]. In the COSMO-SAC model, each molecule is described by an apparent surface charge density distribution, determined by the COSMO[11] method. These charge densities are usually represented by a single variable function denoted as *sigma-profile* or simply σ-profile, in order to simplify the mathematical treatment. The σ-profile establishes a relationship between the fraction of the molecular surface to a corresponding charge density. The calculation of the activity coefficient according COSMO-SAC requires then the σ-profile, the van der Waals surface area and volume of each molecule present in the mixture. This makes the model genuinely predictive[12], depending only on a small set of universal parameters.

The σ-profile of a given molecule is obtained from its electronic structure, computed by quantum chemistry methods. There is a variety of first principles methods which can be employed for this task, such as the well-known Hartree-Fock (HF) method and its variants that include effects of electronic correlation, as well as quantum Density Functional Theory (DFT) calculations. The objective of the HF-based methods is to find an approximate solution of the Schröedinger's equation for the interacting all-electron system[13], obtaining then the molecular electronic wave function ($\phi(r)$), from which any measurable information could be extracted, including the σ-profile[14,15]. In DFT calculations, the objective is to determine a 3-D function called electronic density distribution ($n(r)$), from which observable properties can be obtained[16]. Both $\phi(r)$ and $n(r)$ functions do not have a general exact analytical form, and are usually expressed through linear combinations of mathematical functions (Gaussians, plane waves, among others), called basis sets[15,16]. Each quantum mechanical (QM) method and variants has its own particularities and involves different kinds of approximations. The accuracy of the predictions brought by these methods and the associated computational cost depend strongly on the theory used and on the quality of basis sets employed for function expansion. Consequently, these settings are expected to affect the computed COSMO σ-profile and consequently the predictions for the activity coefficients and related properties.

Considering this significant variety in QM methodologies, some studies in the literature have evaluated the influence of the approach used to compute the electronic structure in the predictive capacity of COSMO methods. Franke and Hannebauer[17] have tested different QM approaches in conjunction with the COSMO-RS model to estimate infinite dilution activity coefficient (IDAC) values. The accuracy of the predictions was verified by comparison with experimental data for a set of 70 species, with a total of 375 IDAC data points. The equilibrium molecular geometries for the molecules were previously determined in vacuum (in the absence of COSMO solvent-induced effects). They concluded that the QM method indeed has influence on COSMO-RS predictions, and their most accurate parameterization is based on screening charge densities calculated by the MP2 method. Chen et al.[18] carried out a similar study using the COSMO-SAC model. The authors used a modified version of this model (called revised COSMO-SAC model, that was proposed by Hsieh et al.[19]). The authors present a different expression for the temperature dependence and for the hydrogen bonding interactions. New parameters were introduced to improve the description of this intermolecular interaction. Besides IDAC values, other properties such as vapor-liquid and liquid-liquid equilibrium (VLE and LLE) data were predicted. Again, the molecular geometry of each involved species was obtained in vacuum (no COSMO effects). In this study the authors also used σ-profiles from an existing database from Virginia Tech (VT)[7], and obtained the same information using different QM approaches as well. The authors found that the original COSMO-SAC is sensitive to the quantum method used, and observed that the revised version is not significantly sensitive to this, allowing using a lower-quality QM method without losing accuracy.

In the work of Wang et al.[20] the authors used COSMO-SAC model with σ-profiles obtained with the GAMESS Quantum Chemistry package[21], and then compared with those available in VT-2005 database[7] (generated with DMol^3)[22]. However, the equilibrium geometry was not obtained for each compound and the COSMO-SAC global parameters were not optimized. The authors concluded that the open-source GAMESS package is a good alternative for σ-profiles generation for uses in COSMO-SAC calculations.

More recently, in the study carried by Paulechka et al.[23], another sigma-profile database was created. This database contains 897 compounds and were used to fit two COSMO-SAC type models parameters. For this estimation, binary liquid mixture data, like equilibrium vapor pressure, activity coefficients, and excess enthalpy, were used. The authors tested different approaches of DFT theories, like B3LYP and BP functionals, combined with different basis sets, and they found the combinations of B3LYP with 6-311G(d,p) basis represented a good accuracy and computational cost.

Considering that obtaining the σ-profile is the most time consuming step in the application of COSMO-based methods to estimate thermodynamic mixture properties, it is desirable to have access to a database containing previously calculated profiles for different substances of interest. These databases are desired to be freely available and easily extensible by the community. There are a few existing databases such as VT-database[7] and Dortmund Data Bank[24]. However, the inclusion of new species requires using commercial software as $\left(\mathrm{DMol}^3\right)$[22] and Gaussian 03[25], which may not be available for some groups interested in collaborating on the extension of the database.

It is noteworthy that there are attempts to produce σ-profiles without directly using computational chemistry packages, for instance by means of group contribution[26,27]. Or even further, by suggesting empirically adjusted σ-profiles as assumed in the F-SAC[28,29] model. Computational effort can be reduced using these approaches, but usually introduce a more empirical character and reduce the prediction power.

Therefore, an open access database of σ-profiles that could be easily extended and improved by users with help of freely distributed tools would be of great value for the chemical thermodynamics community. The present work aims to provide high quality σ-profiles to a large number of distinct chemical species and store them in a freely available database that satisfies these requirements, employing the also freely available GAMESS quantum chemistry package[21].

As discussed in the previous paragraphs, the profiles should be generated using a consistent QM methodology to avoid the introduction of systematic errors. The influence of different QM methods and basis sets for the calculation of the electronic structure was then evaluated in this work, and modifications with respect to the previous studies were implemented, such as the determination of the equilibrium molecular geometries including solvation effects as predicted by COSMO and re-estimation of the universal parameters of the model. The accuracy and efficiency of the methodologies were evaluated by comparison of predicted IDAC values for 760 binary mixtures with experimental data and the associated computational times. Among these 760 binary mixtures, 659 contain only substances that do not present hydrogen bonding interactions. The remaining ones are aqueous mixtures, characterized by the presence of this kind of intermolecular interaction. Based on this analysis, we determined the most suitable methodology to be used in the creation of our database, which is freely available at https://github.com/lvpp/sigma/.

Computational methods and studied substances

Before the construction of the database containing σ-profiles for a large number of molecules, some preliminary tests were performed. A smaller set containing substances belonging to different chemical classes was selected. This set includes hydrocarbons (linear, cyclic, aromatic and aliphatic), aldehydes, ethers, esters, among others. Table 1 shows the substances selected for this preliminary study. The creation of this set (smaller one) of compounds was necessary due to a high computational cost to obtain the equilibrium geometries and the profiles for all compounds (in this work, the optimization of geometry was conducted considering COSMO effects).

[Table 1 about here.]

In the calculation of σ-profiles, the computational chemistry package GAMESS[21] was used, which is free for academic use (including source code) and can be easily modified to generate the

σ-profiles required in COSMO-based calculations. The use of GAMESS package for QM calculations is recommended for COSMO-based applications, as reported by Wang et al.[20] in their study. The initial structure of each compound was drawn and pre-optimized using the AVOGADRO software[30]. For this pre-optimization, the default option available was considered (UFF force field with Steepest Descent algorithm). With this, the initial conformation of each compound was not so far from that one that has been optimized with GAMESS package. Four different QM methods were used in the calculation of the electronic structure of the molecules, namely the Hartree-Fock (HF)[14,15] method, the Møller-Plesset perturbation theory (MP2)[31], and the DFT method using two well-known hybrid exchange-correlation functionals, the BP86[32,33] and B3LYP[32,34]. These methods have been used in similar previous studies[17,35]. Different basis sets were tested: the triple-zeta basis sets TZV and TZVP[36], and the Pople basis sets 3-21G[37–39], 3-21++G(d,p), 6-311G[40], 6-311G(d,p) and 6-311G++(d,p), in order to check the influence of the number of functions on the set and the presence of extra functions that include diffuse and polarization effects. For each combination of method and basis set, the geometry of the molecules was optimized until the largest component of the gradient is less than 1×10^{-4} Ha/Bohr. The optimization was carried out considering that the molecule is immersed in a solvent with a probing radius of 1.4 Å. After the geometry optimization, the sigma profile was obtained by a minor modification in the GAMESS code version *Dec 5, 2014*, as described in https://github.com/lvpp/sigma.

In the present study, some global parameters of COSMO-SAC were adjusted for each combination of QM method/basis set, namely the standard and average radius (r_{eff} and r_{avg}, respectively), the polarization factor (f_{pol}), the cutoff value for hydrogen-bonding interactions (σ_{HB}), and the constant for the hydrogen-bonding interaction (C_{HB}).

For that, the original COSMO-SAC model, presented by Lin and Sandler[6] was used in the present study. The only modification adopted in this work, was the inclusion of r_{avg}, introduced in the work of Mullins et al.[7]. If the so called COSMO-SAC (2010) version[19] were considered, multiple hydrogen bonding constants should be adjusted. Thus, we decided to use the COSMO-SAC (2002) version because of its smaller set of parameters. The accuracy of the infinity dilution activity coefficients (IDAC) predicted by the COSMO-SAC model using σ-profiles generated for each tested combination (basis set and QM method) was evaluated by direct comparison with the experimental data collected in the works of Soares et al.[28] and Soares e Gerber[41].

To evaluate the quality of the predictions brought by COSMO-SAC using the σ-profiles for each combination of basis sets and theory, we plotted the logarithm of the experimental IDAC values versus those predicted by the model. For comparison purposes, values of IDAC were calculated using COSMO-SAC with two other sets of σ-profiles already available: one from the VT database[7] and another generated using the semi empirical QM method POA1[42] (obtained with the quantum chemistry package MOPAC[43]). Besides the COSMO-SAC model, we also used the UNIFAC(Do)[4] model to estimate IDAC values for the sake of comparison. This comparison is important, because the latter model is commonly used in phase equilibrium calculations and presents good accuracy, however it requires extensive sets of experimental data for the estimation of its parameters[4,44]. The computational time required to generate the σ-profiles (structural optimization + σ-profile computation) for the full set of substances described in Table 1 is also reported (using a desktop equipped with a Intel® Core™ i7-3770S @ 3.10 GHz processor and 8 GB of RAM), in order to evaluate the CPU cost for each combination of QM method and basis set.

For the estimation of the COSMO-SAC universal parameters ($\mathrm{r_{eff}}$, $\mathrm{f_{pol}}$, σ_{HB}, and $\mathrm{C_{HB}}$), the objective function was defined as the average absolute deviation (AAD) between the natural logarithm of calculated IDAC values and the experimental ones:

$$\mathrm{AAD} = \frac{1}{\mathrm{NP}} \sum_{\mathrm{i}}^{\mathrm{NP}} \left| \ln \gamma^{\infty}_{\mathrm{i,mod}} - \ln \gamma^{\infty}_{\mathrm{i,exp}} \right| \tag{1}$$

where NP is the number of experimental points, and $\ln \gamma^{\infty}_{i,\mathrm{mod}}$ and $\ln \gamma^{\infty}_{i,\mathrm{exp}}$ are, respectively, the natural logarithm of IDAC calculated by the model and experimental values. To minimize the objective function, the Nelder-Mead[45] method was used. Values presented by Wang and Sandler[46] for the parameters $\mathrm{r_{eff}}$, $\mathrm{r_{avg}}$, $\mathrm{f_{pol}}$, $\mathrm{C_{HB}}$, and σ_{HB} were chosen as the initial estimates. For the first round of parameter, $\mathrm{r_{avg}}$ was assumed as 1.0 Å, only to determine which QM method presents the best accuracy in IDAC values prediction. Once the best method was chosen, the value of $\mathrm{r_{avg}}$ was also optimized.

In addition to IDAC values, some vapor-liquid equilibrium (VLE) and liquid-liquid equilibrium (LLE) diagrams were computed and compared to experimental data to assess the quality of the predictions of each approach. The Modified Raoult's law and the Gamma-Gamma Method were applied to VLE and LLE calculations, respectively. These diagrams allow to evaluate the accuracy of σ-profiles used in COSMO-SAC to predict the behavior of substances in mixtures at finite dilution. The deviations from Raoult's law (ideal mixtures) and formation of azeotropes can be easily identified in these charts.

Results and Discussion

Determination of best QM method

As discussed in the Introduction, the choice of QM method and basis sets has an influence on the computed σ-profiles. This effect can be seen in Figure 1-(a), which depicts the 3-D apparent surface charges around an acetone molecule as well as the σ-profiles obtained by the HF method employing different basis sets. Similarly, Figure 1-(b) presents σ-profiles obtained by different QM methods employing the 6-311G(d,p) basis set. The apparent surface charges image shown in Figure 1 is characterized by induced negative charge density regions near the hydrogen atoms (blue) and positive around the oxygen (red). Neutral regions (green) are also present near the carbon atoms. The one-dimensional representation of this surface, i.e., the σ-profile, has distinguished peaks associated to each one of these regions (one with positive σ, one negative and one neutral). Even though the profiles predicted by the different QM methods and basis sets look qualitatively similar, there are significant quantitative differences among some of them. Clearly, the direct use of the σ-profiles for evaluation purposes is not practical. Therefore, a most suitable approach would be to predict the properties of non-ideal mixtures containing a given molecule using a COSMO-based method and compare them to experimental data, as discussed in the previous section.

[Figure 1 about here.]

A summary of the results obtained by employing the various basis sets and QM methods to generate the σ-profiles for the COSMO-SAC model is listed in Table 2. This table provides the values of fitted parameters for each set of calculations, as well as some performance parameters, namely the corresponding correlation coefficient (R^2) of the experimental versus predicted IDAC parity plots, the average absolute deviation (AAD), and the total CPU time required to generate the profiles for the whole list of substances listed in Table 1. According to the correlation coefficient values reported in the table, the sigma profiles that lead to IDAC values closer to those obtained experimentally are the ones obtained by using the HF method with the TZVP basis set. The same combination of QM method and basis sets is the one that gives the lowest AAD value. Interestingly, other combinations (HF with 6-311G(d, p) or 6-311++G(d, p), for example) also lead to similar values of R^2 and AAD. Reasons for differences and similarities in the results summarized in Table 2 will be discussed in the next paragraphs.

[Table 2 about here.]

Figure 2 illustrates the average absolute deviation and total CPU time to construct the whole set of σ-profiles for each combination of QM theories and basis sets. The results in Figure 2-(a) show the effect of basis sets on the accuracy of predictions. The use of larger basis sets (increasing the number of primitive Gaussians to describe valence and core orbitals, from 3-2xx to 6-3xx[37,40]) and inclusion of diffuse or polarization functions (as denoted by ++, (d,p) and P notations[15,36,47]) clearly leads, for most combinations, to an improvement on the results. However, the computational cost increases proportionally to the size of the basis set, as illustrated on Figure 2-(b).

Figure 2 also allows analyzing the dependence of the accuracy of results and CPU cost on the QM method employed. The AAD values (Figure 2-(a) and Table 2) show that the choice of the QM method does not have a significant impact on the quality of the predictions; similar AAD values are obtained for all methods if the same basis set is used. However, the associated computational time varies considerably, being HF the cheapest and MP2 the most costly (the latter requires up to ~ 10 times more CPU time when compared to the former).

Same observations can be analogously made by analyzing the IDAC parity plots depicted in Figures 3 and 4. The closer the points are to the diagonal, the better the accuracy of predicted IDAC values. Visual inspection of Figure 3 shows that the use of larger basis sets (i.e., increasing the number of functions and including diffuse and polarization effects) leads to a better agreement of predicted IDAC with experimental data. Similarly, IDAC parity plots created by using the same basis set (TZVP) and different QM theories to generate the σ-profiles (Figure 4) shows that the accuracy of results are quite similar, demonstrating that the choice of QM method does not have a significant influence on the accuracy of results.

[Figure 2 about here.]

[Figure 3 about here.]

These results show that the choice of the basis set used to generate the σ-profiles presented a more significant impact on the final calculation of the IDAC values compared to the QM method employed. The inclusion of electronic correlation effects in the QM method (comparing HF versus MP2) does not seem to have a significant influence on the σ-profiles and the COSMO-predicted properties, but requires considerably longer CPU times. The presented analysis suggests that the use of a larger basis set, including polarization effects (such as 6-311G(d.p), 6-311++ G(d,p) and TZVP) and a cheaper first-principles QM method (such as HF) makes a suitable strategy to generate a large database of accurate σ-profiles within a reasonable computational time.

According to the results presented in this section, the smaller deviations between predicted and experimental IDAC values were obtained by the use of HF method along with the TZVP basis set. In the study of Paulechka et al.[23] the best combination of QM theory and basis set was found to be B3LYP with 6-311G(d,p). We have tested this combination to calculate the σ-profiles, but the results did not presented as good accuracy as observed with HF/TZVP. In addition to that, the CPU time consumed to obtain the profiles (using B3LYP/6-311G(d,p)) was larger than when using HF/TZVP method (in our study).

Thus the HF/TZVP approach could be pointed out as the most appropriate choice to generate the σ-profiles database, with a reasonable CPU time required. Other similar combinations of theory and basis sets could also be used to perform the calculations. We then decided to use the HF/TZVP combination in the creation of our database, and we strongly suggest using the same approach for additional calculations to be included in the database, for consistency purpose.

After deciding to use the σ-profiles computed by the HF/TZVP combination, the r_{avg} parameter was also optimized along with the other universal parameters. The estimation was achieved indirectly, by varying r_{avg} from 0.5 to 1.8 Å while reestimating all other parameters. The AAD results for this procedure are shown in Figure 5. The minimum AAD value was observed with r_{avg}=1.4 Å, and the corresponding values of the new parameters for the calibration of COSMO-SAC model with σ-profiles obtained with HF/TZVP are: f_{pol} = 0.8041; r_{eff} = 1.1565 Å; C_{HB} = 58250.49 $kcal/mol$ Å$^4/e^2$; σ_{HB} = 0.01008 e/Å^2, resulting in the performance parameters AAD = 0.2106 and R^2 = 0.9908.

[Figure 4 about here.]

[Figure 5 about here.]

Further comparisons can be made to evaluate the accuracy of the thermodynamic mixture properties predicted by COSMO-SAC using the σ-profiles from the database generated in this work. IDAC predictions were compared with those obtained by COSMO-SAC using σ-profiles available from other databases, as well as with predictions of the UNIFAC(Do)[4] model. These results are available in Table 2 and Figure 6.

In Figure 6-(b), results obtained when using σ-profiles comqing from the quantum semi-empirical POA1 method[42] (using the MOPAC package) are shown. The semi-empirical methods are considerable cheaper but not as accurate than the other approaches used in this work for electronic structure calculations, as can be seen in Figure 6-(a). Interestingly, POA1 method presented results better than several HF/DFT based methods with different basis sets when only the correlation coefficient is analyzed. However, the physical meaning of each parameter must be considered too. For example, the polarization factor (f_{pol}) parameter. In the paper published by Klamt[5], he mentioned that the misfit energy constant (α) must be corrected ($\alpha' = f_{pol}\alpha$), due to polarizations effects. When in contact, two molecules will adjust their wave functions and, then, the misfit energy should be reduced due to the polarizability of each molecule. In this paper, Klamt calculated and presented 0.60 as aproximation to f_{pol}. Although, he found 0.64 as an optimal value for it. Clearly, values higher than 1.0 (or much lower than 0.6) are not physically sound. When POA1 method was used and the COSMO-SAC global parameters were fitted, the value of f_{pol} was larger than 1.6, a number that does not have a physical sense. For the other QM methods, on the other hand, the f_{pol} parameter keeps between 0.5 and 1.0. In the work of Chen et al.[18], another semi-empirical method (PM6[48]) was used. The authors concluded the performance of COSMO-SAC model when PM6 is used to predict VLE, LLE, IDAC, and Kow is much worse than that based on other DFT-based calculations.

Results obtained when using COSMO-SAC with σ-profiles from the VT database[7] are shown in Figure 6-(c). Finally, results for the UNIFAC(Do) model[4], are shown in Figure 6-(d). The empirical UNIFAC(Do) model can predict very well IDAC values for the region below 5 logarithmic units, since most of these data where probably included in the correlation of its parameter matrices. For larger IDAC values (higher than 5 logarithm units), corresponding to aqueous solutions, the UNIFAC(Do) results are not as precise. COSMO-SAC has the interesting advantage of being a predictive model that uses only pure substance information. UNIFAC requires estimating new interaction parameters for every functional group to be included in its current parameters package. The parity plot obtained with σ-profiles from the VT database[7] present a very good correlation, with R^2 and AAD values quite close to the obtained by using our approach (HF theory and TZVP basis set). These comparisons indicate that our database of sigma profiles leads to reliable results when used with the COSMO-SAC method, being able to be used to predict thermodynamic properties for mixtures involving the available species. The inclusion of new species in a consistent manner (employing the aforementioned combination of QM method and basis set) will expand the spectra of mixtures which can be analyzed.

[Figure 6 about here.]

VLE and LLE data Prediction

Finally, we have evaluated the accuracy of the predictions of vapor-liquid (VLE) and liquid-liquid (LLE) equilibrium diagrams for binary mixtures by using COSMO-SAC with the σ-profiles from our database. Figures 7, 8, and 10 depict diagrams for a series of binary mixtures consisting of species available in our database.

[Figure 7 about here.]

[Figure 8 about here.]

The results of Figures 7, 8, and 10 show that the COSMO-SAC model is able to predict properly the equilibrium diagrams using the σ-profiles from our database (generated with the HF/TZVP combination). For sake of comparison, same diagrams were created using COSMO-SAC with σ-profiles from VT-database[7] and the UNIFAC(Do)[4] model. In general, the quality of the predictions of COSMO-SAC is similar, with a performance slightly superior when using the profiles from our database (except for the case depicted in Figure 7-(b)). As expected, UNIFAC(Do) presented the best results in general.

Figure 8 shows VLE results of aqueous systems and other ones containing alcohols. In general, the diagrams show that COSMO-SAC combined with σ-profiles from our database presents better results, when compared with profiles from VT[7]. It is noteworthy that these are genuine predictions, since neither methanol nor ethanol were included in the parameter estimation procedure and the hydrogen bond energy was optimized only for water. For all four mixtures, UNIFAC(Do) predicted very well these equilibrium data.

As discussed previously, the HF/TZVP combination showed a good accuracy in IDAC calculations. As expected, this combination also leads better results than the much faster semi-empirical method POA1. The same trend was also observed in the prediction of VLE data, as depicted in Figure 9 for some representative systems. Analyzing the average absolute relative deviation (AARD) of pressure data of acetone/n-hexane at 338K, HF/TZVP presented lower deviation (AARD = 0.020) while the deviation for POA1 was 0.228. Other example is equilibrium data of chloroform and n-heptane, when HF/TZVP leads to lower deviations (AARD = 0.022), when compared with POA1 (AARD = 0.078).

[Figure 9 about here.]

The LLE diagrams in Figure 10-(a) and -(b) are well predicted by COSMO-SAC model using the proposed σ-profile database. This diagrams consist in water-hydrocarbon mixtures, demonstrating again the suitability of the sigma profiles from our database to be used in COSMO-SAC predictions of thermodynamic properties of mixtures. However, for alcohol-water mixtures, shown in Figure 10-(c) and Figure 10-(d), some deviations can be seen when compared with the experimental data. These results could be improved by including a specific hydrogen bond energy parameter for alcohol molecules, as in the work of Chen et al.[18].

[Figure 10 about here.]

Extended σ-profile Database

The comprehensive set of tests carried out and discussed in the previous sections demonstrates that the use of COSMO-SAC along with σ-profiles generated with the Hartree-Fock (HF) method and the TZVP basis set leads to accurate predictions of IDAC values and VLE/LLE diagrams. This combination showed a good balance between computational cost and accuracy. Thus, an extended database containing the σ-profiles for 1625 substances was created using the proposed approach, which is now freely available in the GitHub platform. Besides the σ-profiles (*.gout* files), all the atomic configuration files (*.mol*) and scripts required to run the calculations in GAMESS are available in the database. All the files can be found on the following WWW address: `https://github.com/lvpp/sigma`. The material contained there is free for any purpose.

Using *internet distributed development* jargon, the academic community can expand the database by *forking* the current database. Later, extensions from *forked* databases can be incorporated in the primary one by *merging*. Potential collaborators must keep in mind that the same method and quantum package suggested in this work should be used for consistency. The atomic configuration (*.mol* files) and script files available in the database for the already existing substances can be used as templates for additional calculations.

The accuracy of the extended σ-profile database available on GitHub can also be verified by constructing an IDAC parity plot. With the extended database a larger number of species and mixtures (in addition to the ones used to calibrate and validate the approach, described in Table 1) is possible. The results shown in Figure 11 correspond to experimental data of 3013 binary mixtures. The good accuracy of results remains within the precision of the COSMO-SAC model, for the extended database with a correlation coefficient and AAD of 0.9655 and 0.3769, respectively. Most of data exhibits deviations within the limits of one logarithmic unit, either positive or negative.

[Figure 11 about here.]

Finally, it must be stressed that the construction of the σ-profiles database proposed in this work was accomplished using only freely available computational tools. This should enable a larger number of research groups to collaborate with the inclusion of new species of interest, helping to make this database larger and more complete.

Conclusion

The major goal of this work was to create a freely available and extensible σ-profiles database for use in conjunction with COSMO-based models. In order to establish an accurate, consistent, and efficient procedure to compute the profiles, different basis sets and theories for the calculation of the electronic structure were tested using the GAMESS package. This package is freely available to academic purposes, including source code, and any user can contribute with new calculations to the database.

The accuracy and effectiveness of each combination of method and basis set was evaluated by predicting values of infinite dilution activity coefficients (IDAC) for binary mixtures using the

COSMO-SAC model. The results showed that the COSMO-SAC model presents a good performance in the prediction of IDAC values, and the Hartree-Fock (HF) method combined with the TZVP basis set was considered as the best approach for the generation of σ-profiles, among the studied methods.

Based on the results of the preliminary analysis employing 35 substances, an extended σ-profile database was generated containing 1625 substances, and made freely available at `https://github.com/lvpp/sigma`. This extended database also gives good predictions, with an absolute average deviation in IDAC logarithm of 0.3769 for a set of 3013 binary mixtures. Since the proposed database was constructed using only freely available computational tools, other authors of the academic community can easily generate σ-profiles for other substances and (optionally) deposit the information in the original database.

Acknowledgments

This work was partially supported by CNPq-Brazil under grants no. 454557/2014-0 and 304046/2016-7.

Literature Cited

1 Leonhard, K.; Veverka, J.; Lucas, K. A comparison of mixing rules for the combination of COSMO-RS and the Peng-Robinson equation of state. *Fluid Phase Equilibria* **2009**, *275*, 105–115.

2 Renon, H.; Prausnitz, J. M. Local compositions in thermodynamic excess functions for liquid mixtures. *AIChE Journal* **1968**, *14*, 135–144.

3 Wilson, G. M. Vapor-Liquid Equilibrium. XI. A New Expression for the Excess Free Energy of Mixing. *Journal of the American Chemical Society* **1964**, *86*, 127–130.

4 Jakob, A.; Grensemann, H.; Lohmann, J.; Gmehling, J. Further Development of Modified UNIFAC (Dortmund): Revision and Extension 5. *Industrial & Engineering Chemistry Research* **2006**, *45*, 7924–7933.

5 Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry* **1995**, *99*, 2224–2235.

6 Lin, S.-T.; Sandler, S. I. A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model. *Industrial & Engineering Chemistry Research* **2002**, *41*, 899–913.

7 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*, 4389–4415.

8 Yang, L.; Sandler, S. I.; Peng, C.; Liu, H.; Hu, Y. Prediction of the Phase Behavior of Ionic Liquid Solutions. *Industrial & Engineering Chemistry Research* **2010**, *49*, 12596–12604.

9 Yang, L.; Xu, X.; Peng, C.; Liu, H.; Hu, Y. Prediction of vapor-liquid equilibrium for polymer solutions based on the COSMO-SAC model. *AIChE Journal* **2010**, *56*, 2687–2698.

10 Yang, L.; Chang, C.-W.; Lin, S.-T. A novel multiscale approach for rapid prediction of phase behaviors with consideration of molecular conformations. *AIChE Journal* **2016**, *62*, 4047–4054.

11 Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.

12 Bouillot, B.; Teychené, S.; Biscans, B. An evaluation of COSMO-SAC model and its evolutions for the prediction of drug-like molecule solubility: Part 1. *Industrial and Engineering Chemistry Research* **2013**, *52*, 9276–9284.

13 Schrödinger, E. An undulatory theory of the mechanics of atoms and molecules. *Physical Review* **1926**, *28*, 1049–1070.

14 Ira N. Levine, Quantum Chemistry. 2000.

15 Szabo, A.; Ostlund, N. S. Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory. 1996.

16 Koch, W.; Holthausen, M. C. *Neural Networks*; 2001; Vol. 3; p 294.

17 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 : PCCP* **2011**, *13*, 21344–50.

18 Chen, W. L.; Hsieh, C. M.; Yang, L.; Hsu, C. C.; Lin, S. T. A Critical Evaluation on the Performance of COSMO-SAC Models for Vapor-Liquid and Liquid-Liquid Equilibrium Predictions Based on Different Quantum Chemical Calculations. *Industrial and Engineering Chemistry Research* **2016**, *55*, 9312–9322.

19 Hsieh, C.-M.; Sandler, S. I.; Lin, S.-T. Improvements of COSMO-SAC for vaporliquid and liquidliquid equilibrium predictions. *Fluid Phase Equilibria* **2010**, *297*, 90–97.

20 Wang, S.; Lin, S.-T.; Watanasiri, S.; Chen, C.-C. Use of GAMESS/COSMO program in support of COSMO-SAC model applications in phase equilibrium prediction calculations. *Fluid Phase Equilibria* **2009**, *276*, 37–45.

21 Schmidt, M. W.; Baldridge, K. K.; Boatz, J.; Elbert, S.; 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*, 1347–1363.

22 Delley, B. From molecules to solids with the DMol3 approach. *Journal of Chemical Physics* **2000**, *113*, 7756–7764.

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*, 3554–3561.

24 Dortmund Data Bank. 2017; `www.ddbst.com`.

25 Frisch, M. J. et al. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

26 Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density profiles for COSMO-RS(01). *AIChE Journal* **2007**, *53*, 3231–3240.

27 Mu, T.; Rarey, J.; Gmehling, J. Group contribution prediction of surface charge density distribution of molecules for COSMO-SAC. *AIChE Journal* **2009**, *55*, 3298–3300.

28 Soares, R. d. P.; Gerber, R. P. Functional-Segment Activity Coefficient Model. 1. Model Formulation. *Ind. Eng. Chem. Res.* **2013**, 11159–11171.

29 Soares, R. D. P.; Gerber, R. P.; Possani, L. F. K.; Staudt, P. B. Functional-Segment Activity Coefficient Model. 2. Associating Mixtures. *Ind. Eng. Chem. Res.* **2013**, *52*, 11172–11181.

30 Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics* **2012**, *4*, 17.

31 Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Physical Review* **1934**, *46*, 618–622.

32 Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38*, 3098–3100.

33 Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B* **1986**, *33*, 8822–8824.

34 Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37*, 785–789.

35 Mu, T.; Rarey, J.; Gmehling, J. Performance of COSMO-RS with Sigma Profiles from Different Model Chemistries. *Industrial & Engineering Chemistry Research* **2007**, *46*, 6612–6629.

36 Schafer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *The Journal of Chemical Physics* **1994**, *100*, 5829.

37 Binkley, J. S.; Pople, J. a.; Hehre, W. J. Self-Consistent Molecular-Orbital Methods. 21. Small Split-Valence Basis-Sets for 1st-Row Elements. *J. Am. Chem. Soc.* **1980**, *102*, 939–947.

38 Gordon, M. S.; Binkley, J. S.; Pople, J. A.; Pietro, W. J.; Hehre, W. J. Self-consistent molecular-orbital methods. 22. Small split-valence basis sets for second-row elements. *Journal of the American Chemical Society* **1982**, *104*, 2797–2803.

39 Pietro, W. J.; Francl, M. M.; Hehre, W. J.; DeFrees, D. J.; Pople, J. A.; Binkley, J. S. Self-Consistent Molecular-Orbital Methods.24. Supplemented Small Split-Valence Basis-Sets for 2nd-Row Elements. *J. Am. Chem. Soc.* **1982**, *104*, 5039–5048.

40 Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

41 Soares, R. d. P.; Gerber, R. P. Functional-Segment Activity Coefficient Model. 1. Model Formulation. *Industrial & Engineering Chemistry Research* **2013**, *52*, 11159–11171.

42 Gerber, R. P.; Soares, R. P. Assessing the reliability of predictive activity coefficient models for molecules consisting of several functional groups. *Brazilian Journal of Chemical Engineering* **2013**, *30*, 1–11.

43 J. J. P. Stewart, *MOPAC2016*; Stewart Computational Chemistry: Colorado Springs, CO, USA, 2016.

44 Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal* **1975**, *21*, 1086–1099.

45 Nelder, J. A.; Mead, R. A Simplex Method for Function Minimization. *The Computer Journal* **1965**, *7*, 308–313.

46 Wang, S.; Sandler, S. I.; Chen, C. C. Refinement of COSMO-SAC and the applications. *Industrial and Engineering Chemistry Research* **2007**, *46*, 7275–7288.

47 Cramer, C. J. *Essentials of Computational Chemistry*; 2004; Vol. 42; pp 334–342.

48 Stewart, J. J. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007**, *13*, 1173–1213.

List of Tables

Table 1: Selected substances for the calculation of σ-profiles in the accuracy tests.

Chemical Funcion	**Substances**
Hydrocarbons	Saturated: 2,2,3-trimethylbutane ; 2,2,4-trimethylpentane; 2,3,4-trimethylpentane; 2,4-dimethylpentane; 2-methylpentane; cyclohexane; methylcycloexane; n-decane; n-heptane; n-hexane; n-nonane; n-octane; n-pentane.
	Unsaturated: 1-butene; 1-heptene;1-hexene.
	Aromatics: benzene; toluene
Ketones	acetone (propanone); methyl-ethyl-ketone.
Organic Halides	1,2-dichloroethane; chloroform; carbon tetrachloride.
Aldehydes	isobutyraldhyde; n-butyraldhyde.
Esters	ethyl acetate; methyl acetate.
Ethers	diethyl ether; dimethyl ether; methyl-n-butyl ether.
Nitriles	isobutyronitrile; n-butyronitrile.
Others	triethylamine; tetrahydrofuran; water.

Table 2: Summary of results of the tests employing different QM methods and basis sets for the calculation of the σ-profile, with r_{avg} = 1.0 Å.

			Fitted COSMO-SAC Parameters				Fit Performance	
		CPU Time (h)	f_{pol}	r_{eff} (Å)	C_{HB} ($kcal/mol$ Å$^4/e^2$)	σ_{HB} ($e/$Å)	AAD	R^2
HF	3-21G	2.31	0.6271	1.1728	59 183.75	0.0110	0.3787	0.9740
	3-21++G(d,p)	7.32	0.6483	1.1013	72 413.24	0.0127	0.2924	0.9805
	6-311G	3.94	0.5449	1.1727	60 969.94	0.0127	0.2610	0.9887
	6-311G(d,p)	9.83	0.8145	1.0955	130 174.25	0.0120	0.2597	0.9885
	6-311++G(d,p)	24.96	0.7479	1.1333	58 976.68	0.0113	0.2437	0.9895
	TZV	8.18	0.5495	1.1846	63 798.19	0.0133	0.2433	0.9901
	TZVP	16.70	0.8095	1.0864	252 310.10	0.0136	0.2315	0.9905
MP2	3-21G	6.63	0.5771	1.2672	55 502.77	0.0123	0.5270	0.9420
	3-21++G(d,p)	47.13	0.6001	1.2053	205 821.19	0.0162	0.3624	0.9754
	6-311G	21.61	0.4642	1.2836	74 515.51	0.0160	0.4743	0.9583
	6-311G(d,p)	85.50	0.6396	1.3038	73 147.00	0.0135	0.4683	0.9611
	6-311++G(d,p)	197.11	0.6034	1.2708	62 789.81	0.0148	0.4459	0.9650
	TZV	49.64	0.4936	1.2433	64 024.06	0.0174	0.4322	0.9643
	TZVP	150.24	0.5820	1.2854	71 540.21	0.0147	0.4500	0.9642
BP86	3-21G	9.65	0.6108	1.2700	51 584.49	0.0117	0.4771	0.9529
	3-21++G(d,p)	26.43	0.5863	1.2023	82 111.21	0.0149	0.3099	0.9831
	6-311G	15.53	0.5057	1.2732	84 416.93	0.0160	0.3773	0.9717
	6-311G(d,p)	30.95	0.6739	1.2990	58 190.95	0.0151	0.4190	0.9681
	6-311++G(d,p)	57.24	0.6027	1.2605	85 589.71	0.0144	0.3728	0.9751
	TZV	21.33	0.4926	1.2481	45 234.07	0.0160	0.3672	0.9741
	TZVP	38.49	0.6024	1.2725	70 098.87	0.0150	0.3829	0.9732
B3LYP	3-21G	11.08	0.6318	1.2604	79 647.45	0.0125	0.5252	0.9438
	3-21++G(d,p)	30.16	0.6293	1.1877	219 570.01	0.0162	0.3455	0.9776
	6-311G	17.16	0.5254	1.2644	86 173.40	0.0154	0.4016	0.9678
	6-311G(d,p)	32.35	0.6704	1.2764	69 717.64	0.0126	0.4219	0.9671
	6-311++G(d,p)	61.52	0.6372	1.2405	65 971.56	0.0137	0.3870	0.9731
	TZV	25.42	0.4981	1.2417	129 892.60	0.0167	0.3991	0.9688
	TZVP	43.34	0.6243	1.2579	49 420.45	0.0138	0.4044	0.9705
VT	DMOL^3	—	0.6469	1.4586	37 842.27	0.0126	0.3788	0.9773
POA1	MOPAC	—	1.6388	0.9911	388 300.17	0.0099	0.2594	0.9872
UNIFAC(Do)	—	—	—	—	—	—	0.3124	0.8511

List of Figures

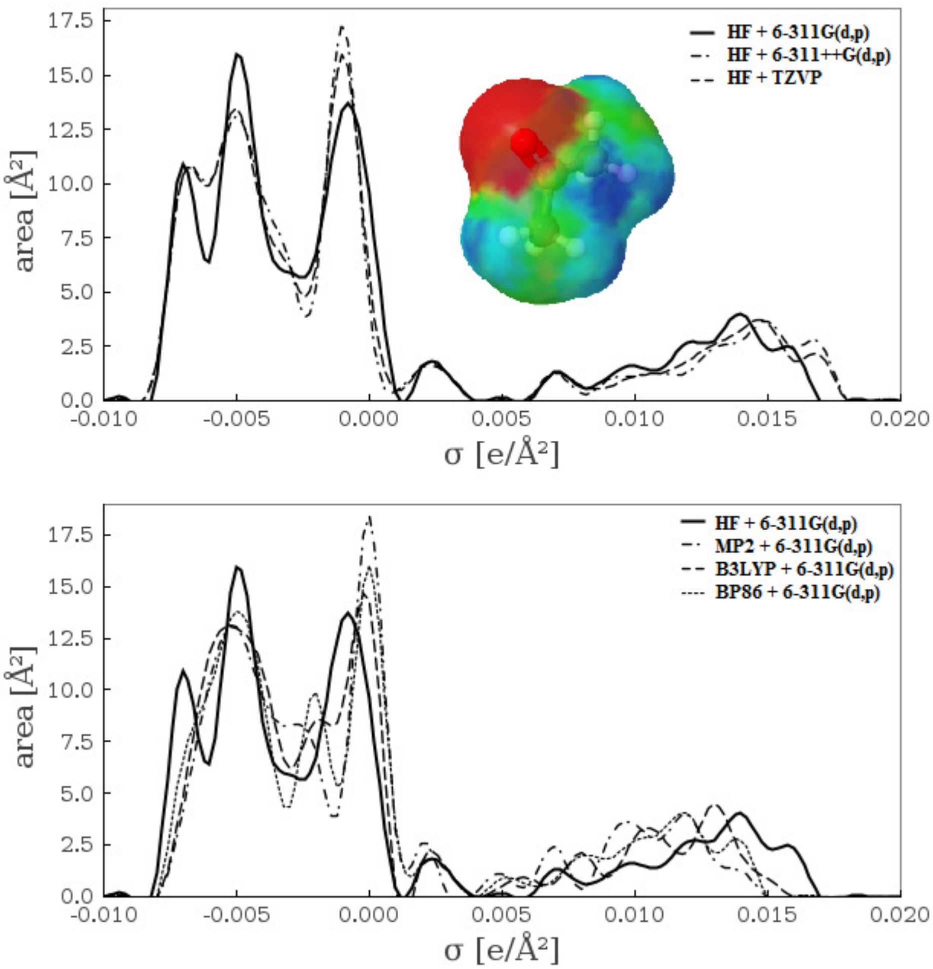
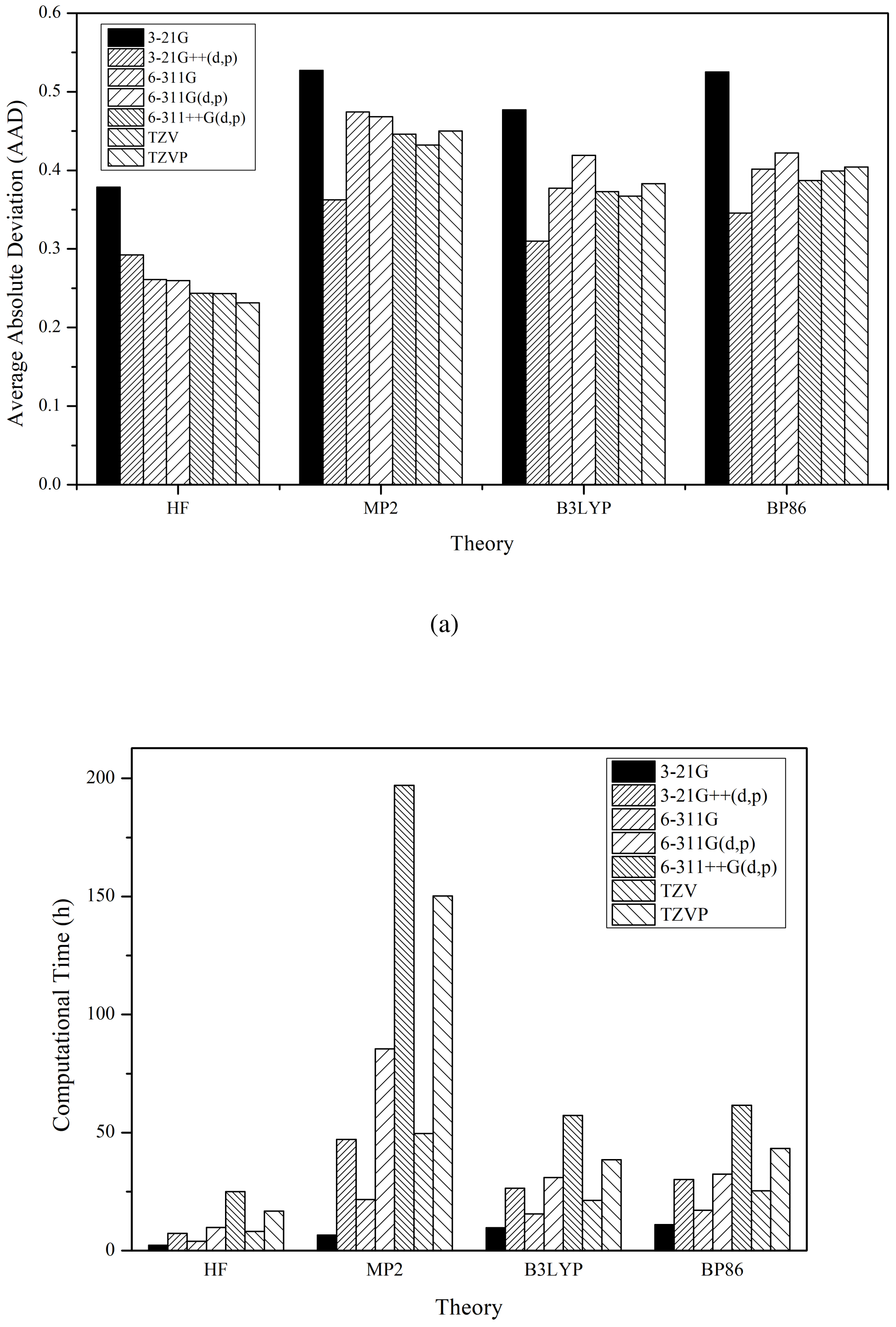


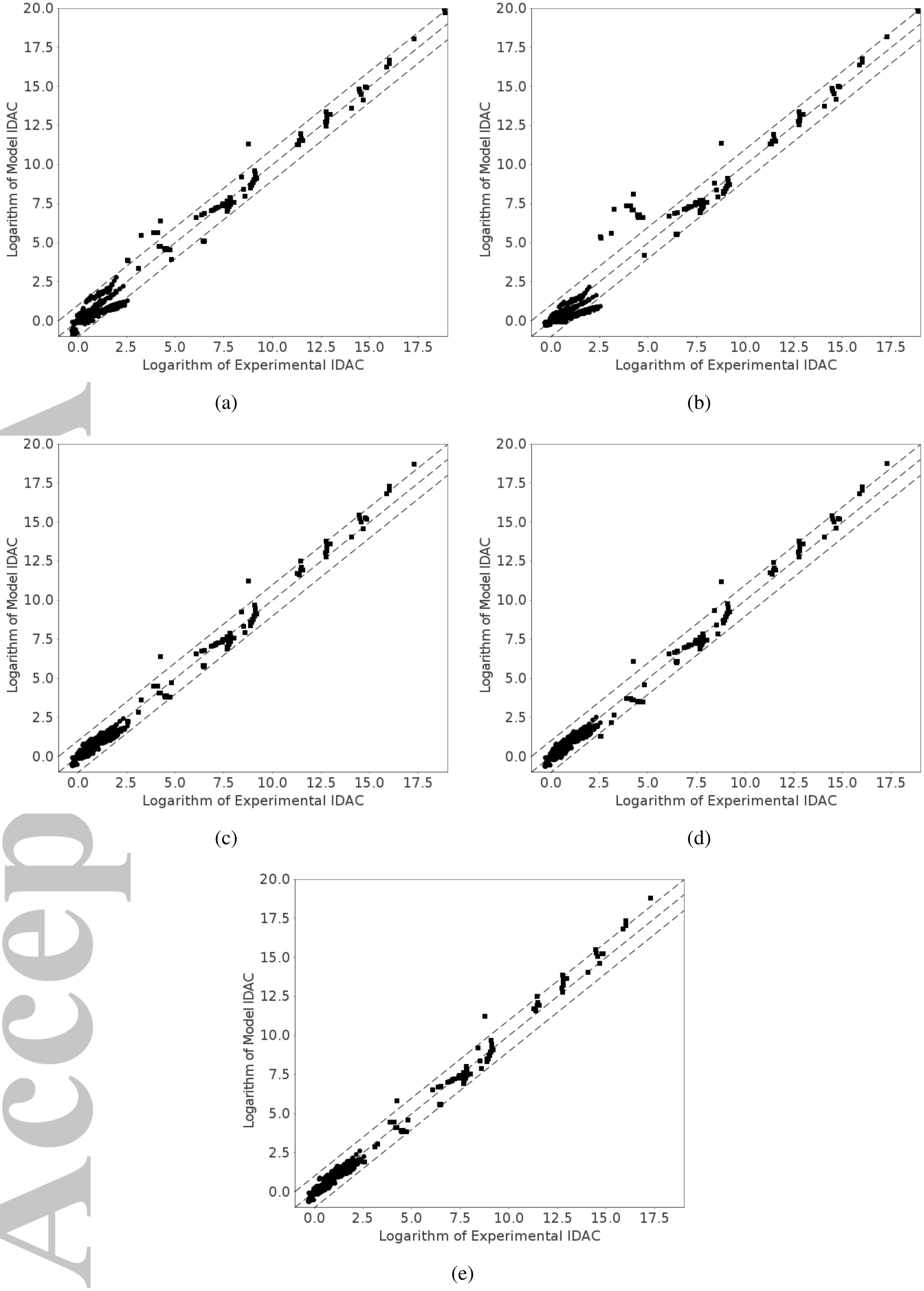
Figure 1: σ-profiles for acetone obtained with (a) HF method with different basis sets, and (b) different QM theories with the 6-311G(d,p) basis set. The tridimensional image was created with HF method and 6-311G(d,p) basis set.



(a)

(b)

Figure 2: Deviation to experimental IDAC data (a) and computational time for calculation of electronic structure of the 35 molecules studied (b) for the different quantum chemistry methods that were tested.



(a)

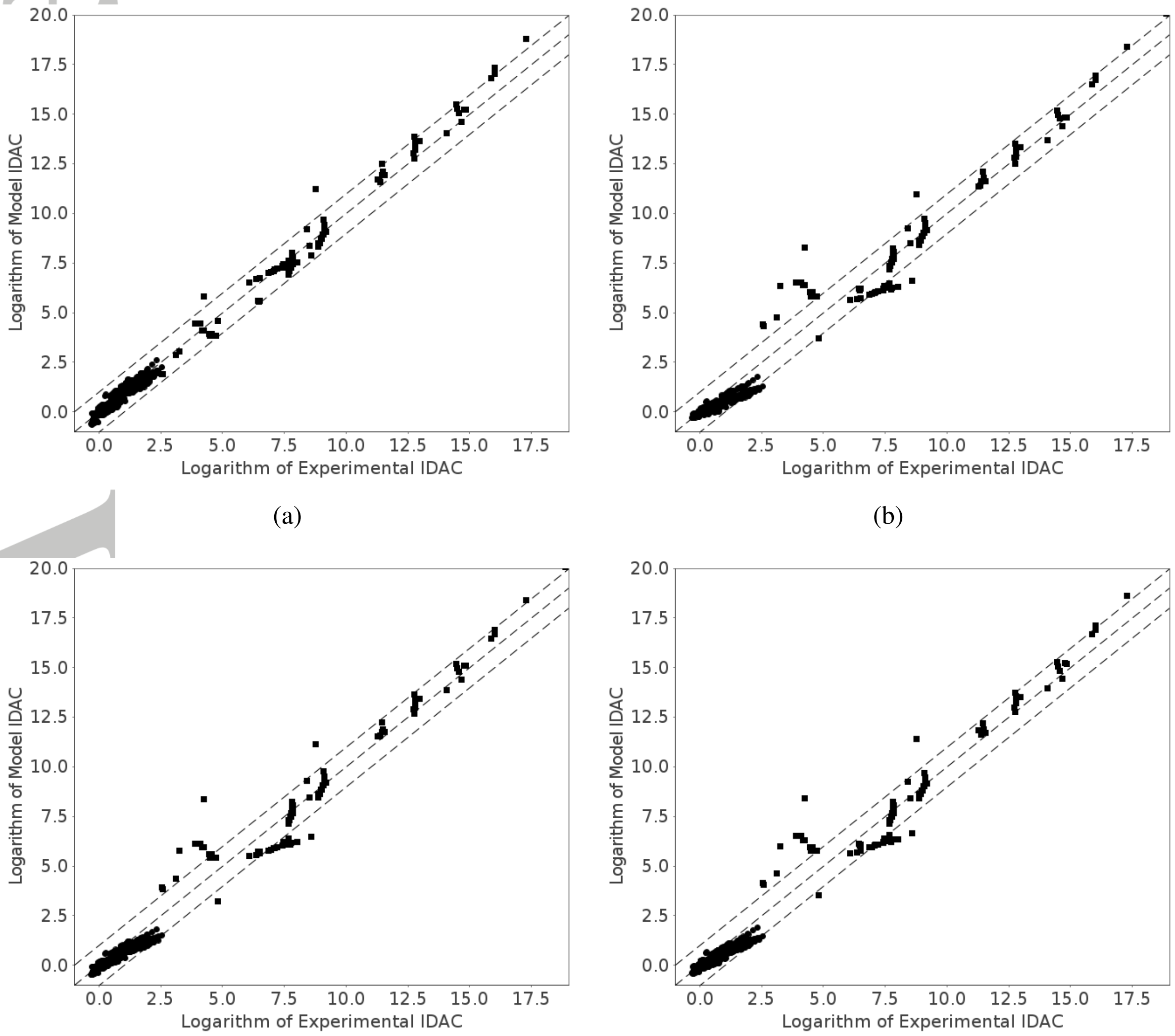
(b)

(c)

(d)

(e)

Figure 3: IDAC parity plots obtained using HF method with different basis sets: (a) 3-21G; (b) 6-311G; (c) 6-311G(d,p); (d) 6-311++G(d,p); and (e) TZVP.



(a)

(b)

(c)

(d)

Figure 4: Parity plots obtained using TZVP as base with different theories: (a) HF; (b) MP2; (c) B3LYP, and (d) BP86.

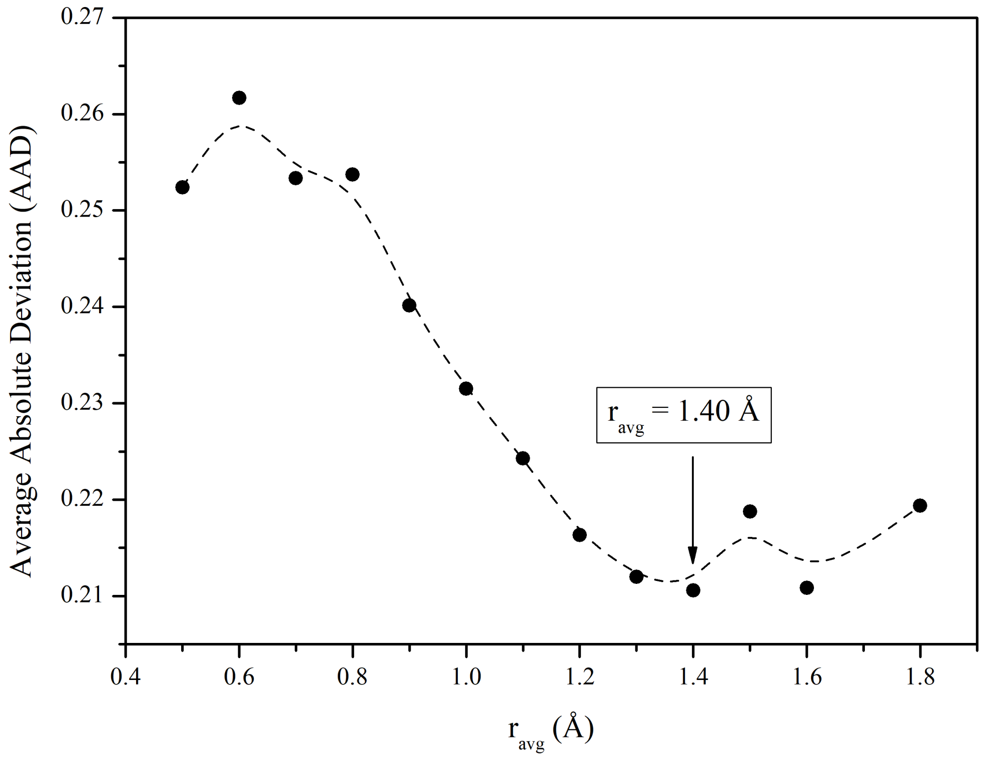
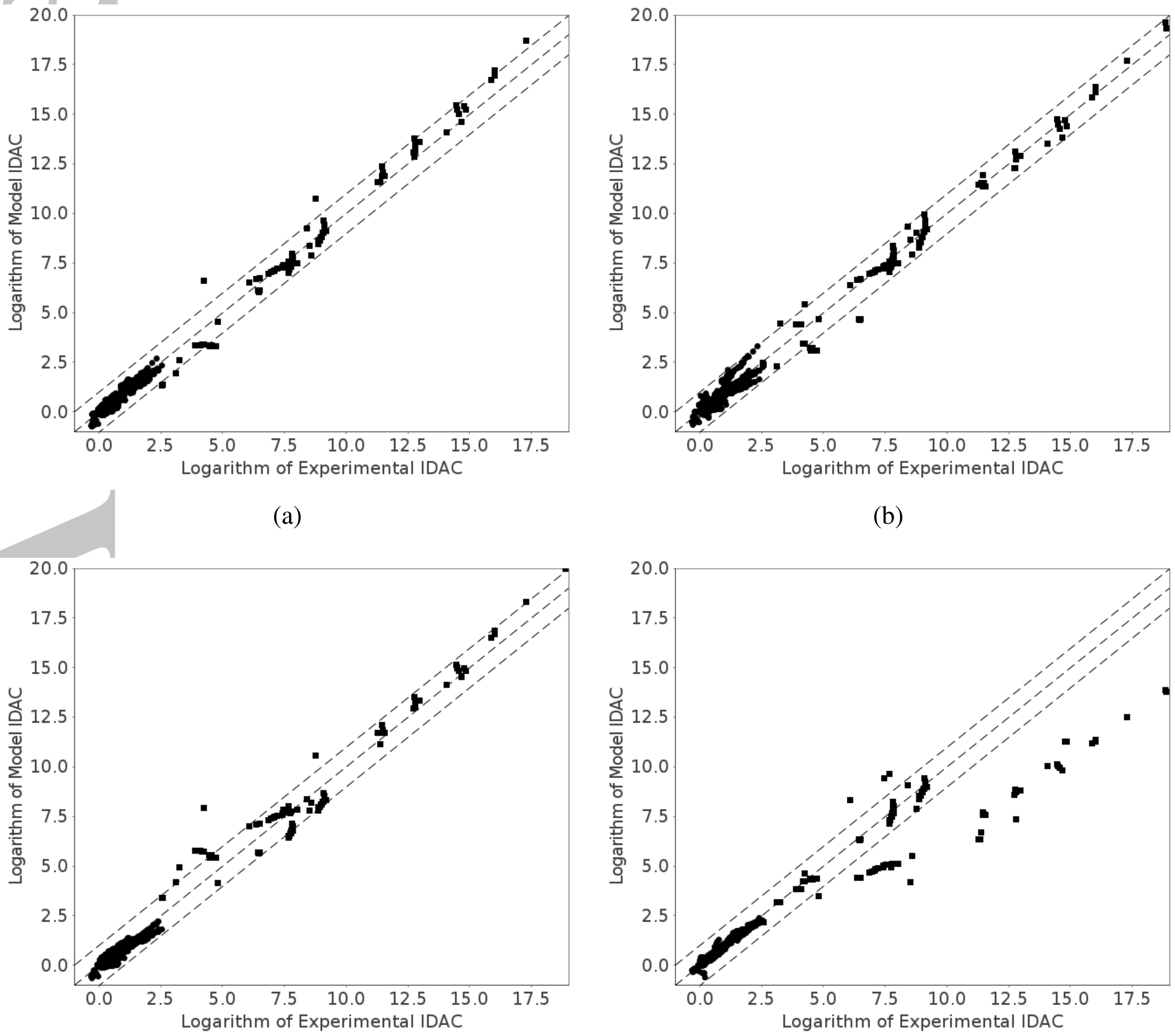


Figure 5: Plot of average absolute deviation (AAD) behavior as funcion of r_{avg} using HF theory and TZVP basis set. The curve through the data serves as a guide to the eyes for the reader.



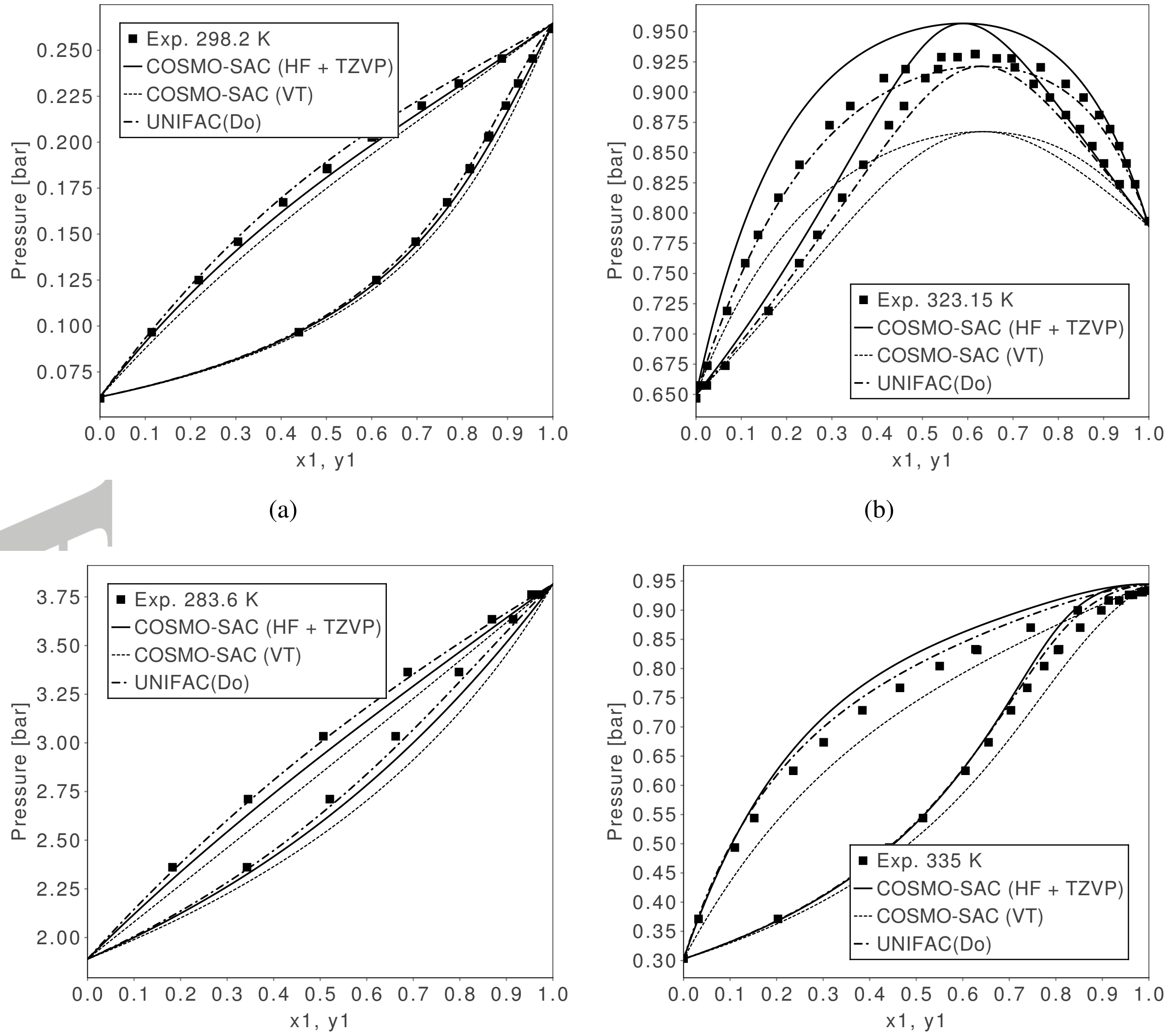
(a)

(b)

(c)

(d)

Figure 6: Parity plots that were obtained with (a) HF + TZVP, (b) POA1 method, (c) VT database, all with COSMO-SAC, and (d) UNIFAC(Do) model.



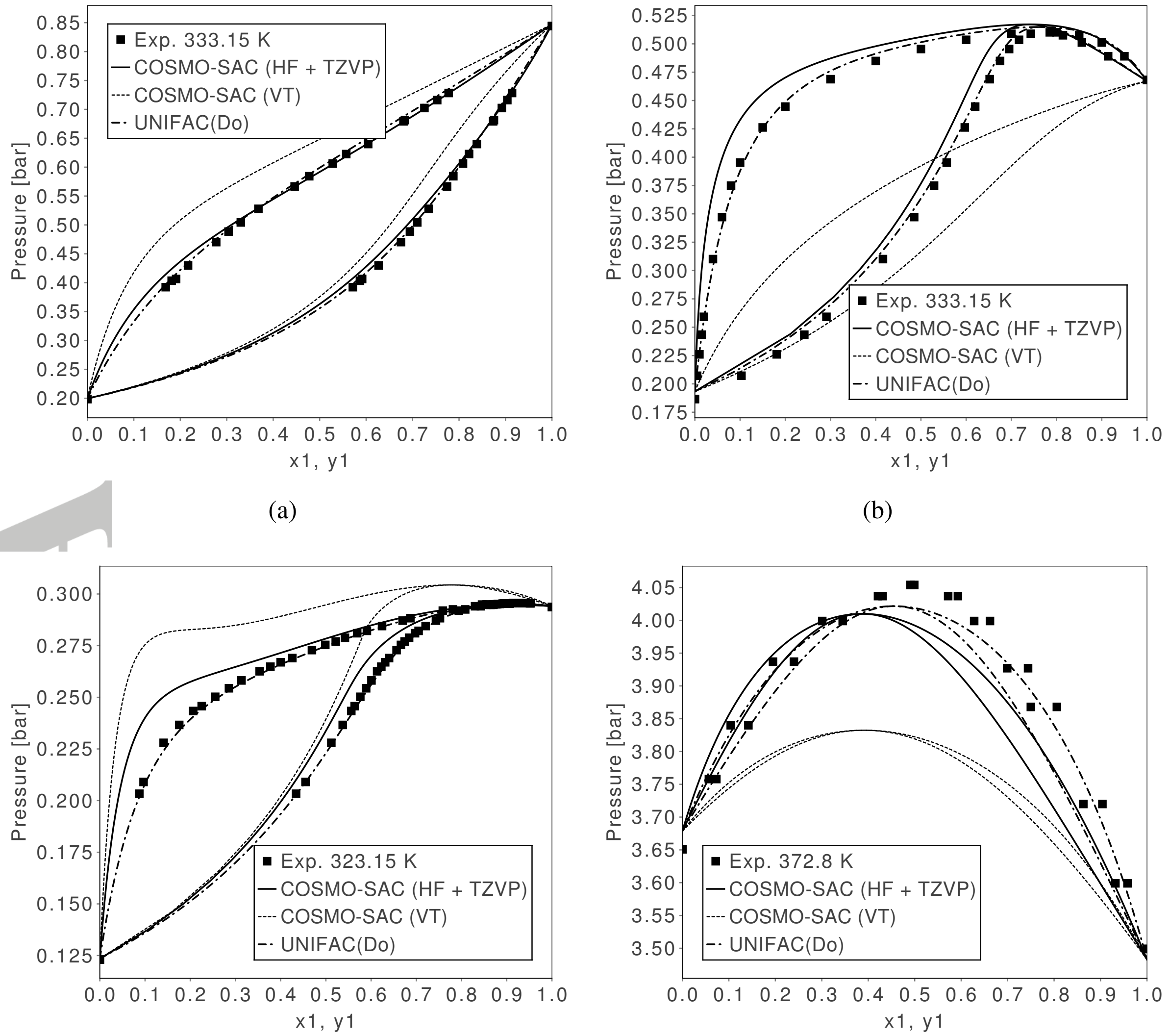
(a)

(b)

(c)

(d)

Figure 7: VLE diagrams for binary mixtures predicted by COSMO-SAC using sigma profiles from our generated database, VT database, and UNIFAC(Do): (a) (1) chloroform and (2) n-heptane at 298.2 K; (b) (1) methyl acetate and (2) 1-hexene at 323.15 K; (c) (1) dimethyl ether and (2) 1-butene at 283.6 K; and (d) (1) isobutyraldehyde and (2) n-heptane at 335 K.



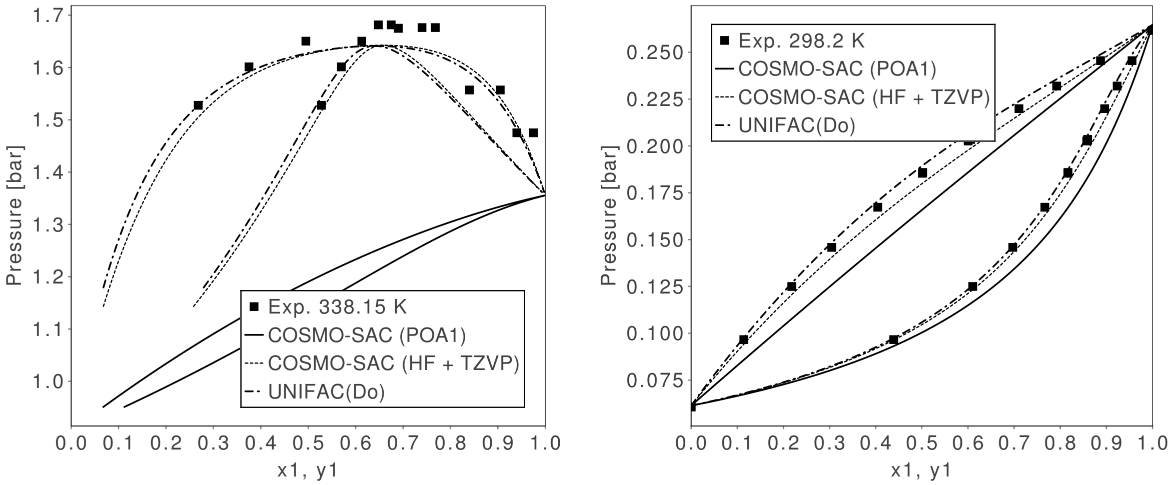
(a)

(b)

(c)

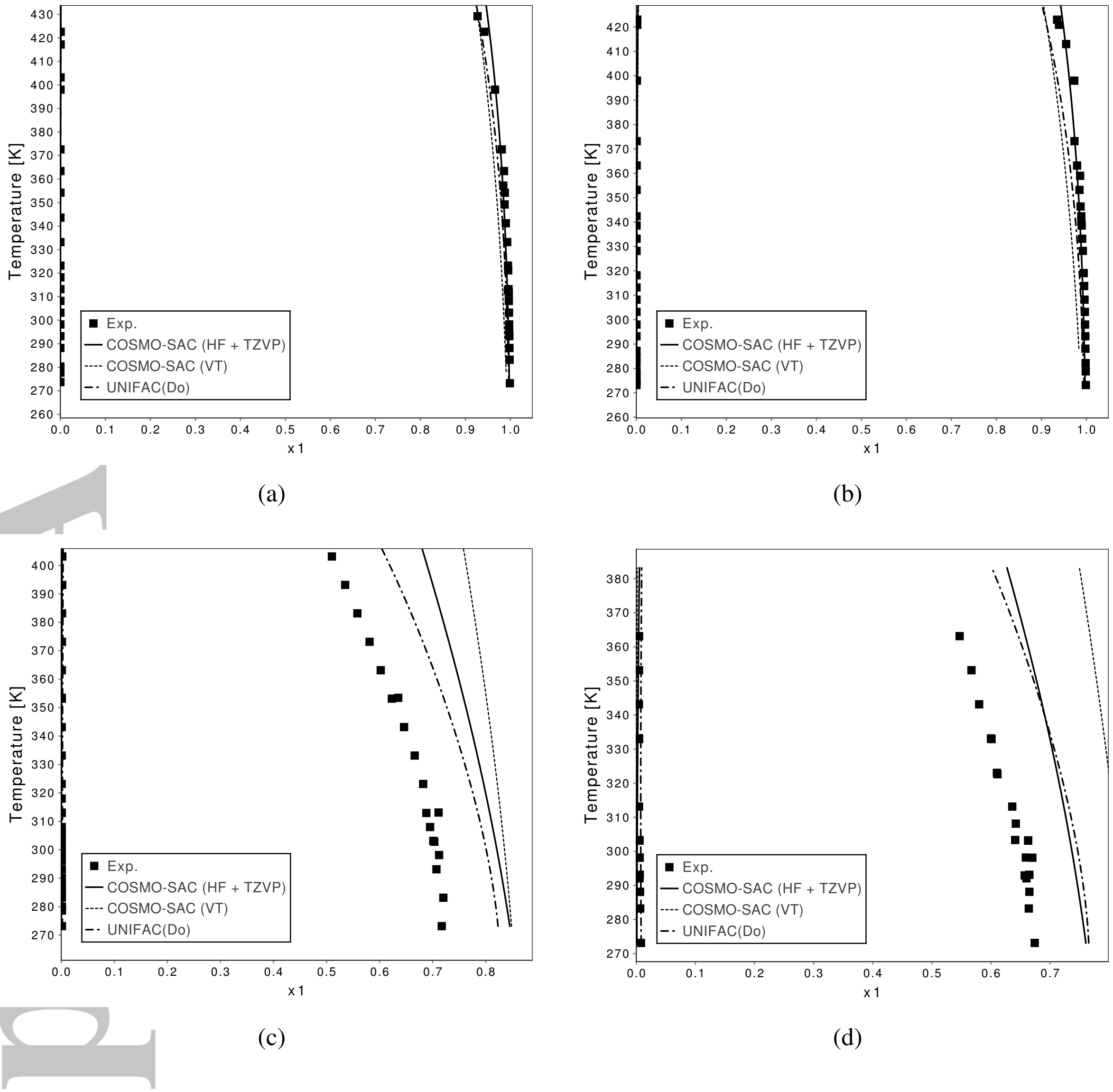
(d)

Figure 8: VLE diagrams for binary mixtures predicted by COSMO-SAC using sigma profiles from our generated database and VT database, and UNIFAC(Do): (a) (1) methanol and (2) water at 333.15.2 K; (b) (1) ethanol and (2) toluene at 333.15 K; (c) (1) ethanol and (2) water at 323.15 K; and (d) (1) methanol and (2) acetone at 372.8 K



(a) (b)

Figure 9: VLE diagrams for binary mixtures predicted by COSMO-SAC using sigma profiles generated with HF/TZVP and POA1 methods, and UNIFAC(Do): (a) (1) acetone and (2) n-hexane at 338.15 K; (b) (1) chloroform and (2) n-heptane at 298.2 K



(a)

(b)

(c)

(d)

Figure 10: LLE diagrams for binary mixtures predicted by COSMO-SAC using sigma profiles from our generated database and VT database, and UNIFAC(Do): (a) (1) toluene and (2) water; (b) (1) benzene and (2) water; (c) (1) 1-hexanol and (2) water; and (d) (1) 3-methyl-1-butanol and (2) water.

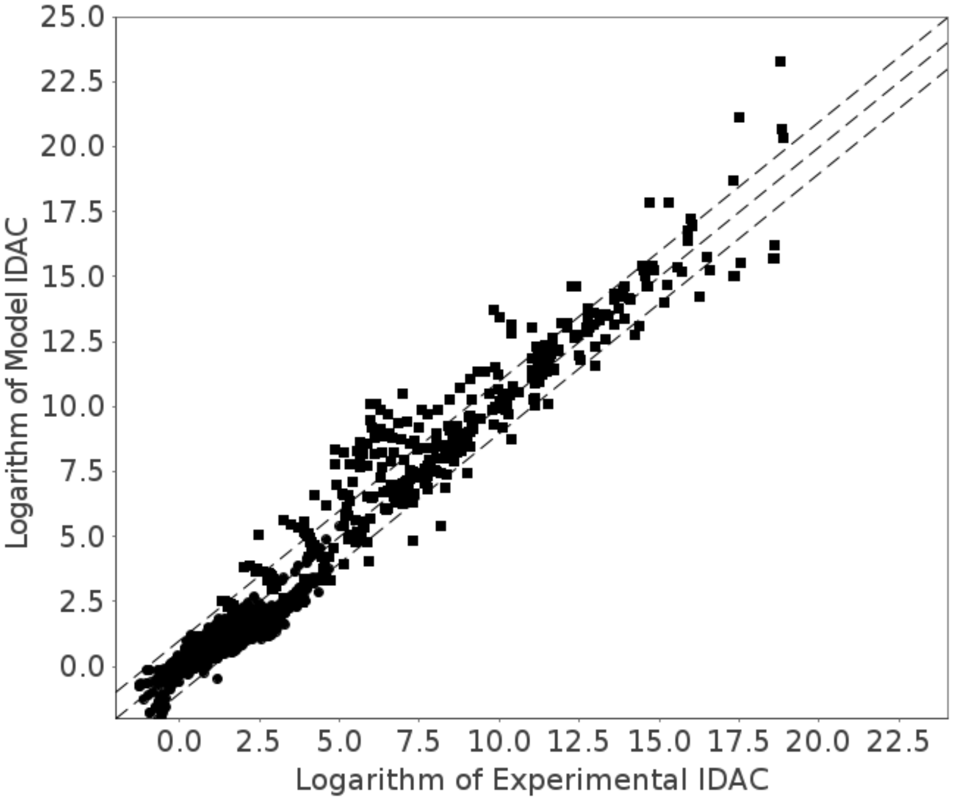


Figure 11: IDAC parity plot obtained using HF method with TZVP basis set for the extended database, comprising 3013 experimental points.