Supporting Information for Publication

Alkane/water partition coefficient calculation based on modified AM1 method and Internal Hydrogen Bonding sampling using COSMO-RS

Panagiotis C. Petris*,†, Paul Becherer‡, Johannes G. E. M. Fraaije*†

* Leiden Institute of Chemistry, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands,

‡ Siemens Industry Software Netherlands B.V., Galileiweg 8, 2333 BD Leiden, The Netherlands.

*E-mail: panos.petris@siemens.com

S1. Supporting Information Files Description and Explanation

(i) AM1-BCC COSMO parameters file
Our model is based on a Bond Charge Correction (BCC) of COSMO charges applied to the AM1 semi-empirical method. We have included all details related to this correction in the am1-bcc.cff force field file. Therein, we include the atom types and the corresponding bond parameters. Depending on the type of the bond we suggest the bond charge correction that should be applied.

To be able to work with the published AM1-BCC atom types together with the molecules in the Culgi Molecules Database, we slightly modified the representation of the molecules. Instead of assigning aromatic bonds by a strict application of Hückel's 4n+2 rule, we used single/double bond representations for two cases: aromatic rings with one or more exocyclic double bonds to oxygen (such as uracil) and aromatic 5-rings connected to aromatic 6-rings. Other aromatic rings retain their representation with (only) aromatic (partial double) bonds. The SMARTS patterns in the parameters table are meant to be used with this modified representation.

The multi-atom SMARTS patterns in the AM1-BCC parameters table match a group of atoms, but they define the atom type of the first atom in the pattern. For example, the pattern \([OX1]=[C;r3,r4,r5,r6,r7,r8,r9][O,N]\) matches a group of three atoms typically
found in lactam or lactone compounds, but for our purposes it defines/detects the atom type of the sp2-hybridized oxygen in that group (the first atom in the pattern).

(ii) Kuhn’s table

As presented in the methodology section of the publication, apart from the BCC incorporation into the AM1 semi-empirical method, we have, also, developed an algorithm to identify Internal Hydrogen Bond (IHB) and generate conformations based on whether IHB’s are formed or not. To identify an IHB we have used Kuhn’s rings as presented in the Kuhntable.csv file. We have included the SMARTS pattern, the hydrogen bonding type, the number of atoms present in the ring, the Kuhn code given in the original work, as well as the Cambridge Structure Database fraction and example. As pointed in the manuscript we have to add 5 new rings to make sure that small molecules containing IHB’s would, also, be allowed to form IHB. Further information regarding the way Kuhn ended up in these very important structures can be found in Kuhn et al.(1)

(iii) BioByte dataset
Following up, we present the Biobyte.csv file, in which we have listed all molecules from BioByte dataset participating in our research. The molecule name, the SMILES string, the solvent name in which partitioning is derived, and ABC water/alkane partition coefficient is included. We have, also, added a column that indicates whether or not this particular molecule exhibits IHB’s.

(iv) SAMPL5 dataset

In the SAMPL5.csv file the drug-like molecules used in the SAMPL5 challenge are presented. As in the case of the Biobyte dataset we have kept the most important columns, as the id, the molecule name (as given in the challenge), the SMILES string, the experimental water/cyclohexane distribution coefficient alongside the uncertainty in the measurement, as well as the result of our model. The last column indicates whether or not this particular molecule exhibits IHB’s.

S2. AM1, ABC and DFT comparison for pyridine-3-sulfonamide molecule
We have separately investigated the performance of the AM1 semi-empirical method and the ABC model for the pyridine-3-sulfonamide molecule that contains the special functional group of sulfonamide. In Figure S1, we illustrate the COSMO surface in the form of a color map using AM1, ABC and DFT methods. Red surface areas correspond to a positive COSMO charge density, while the blue areas carry a negative COSMO charge density. Grey COSMO surface indicates neutral COSMO charge density. It is graphically evident that the ABC COSMO surface is closer to the DFT counterpart, while AM1 exhibits a somewhat different surface profile, with regions of extreme negative and positive COSMO charge densities. This semi-quantitative demonstration amplifies the already observed improvement that ABC introduces, when compared to the AM1 semi-empirical method, for practically all of the functional groups investigated in this work.
Figure S1: The COSMO surface, in color map, for the pyridine-3-sulfonamide molecule calculated using the AM1 semi-empirical method, the ABC model and DFT with the modified NWChem software package.\(^{(2)}\)

S3. Orbital-COSMO interaction

The total electron density \(\rho(x)\) is given by the sum of the squares of the molecular orbitals. This can be expressed in terms of the molecular orbital coefficients and the basis functions, ultimately giving this expression in terms of the basis functions \(\chi_\mu\) and the density matrix \(P\)

\[
\rho(r) = \sum_\mu \sum_\nu P_{\mu\nu} \chi_\mu(r) \chi_\nu(r). \quad (S.1)
\]

To get the corresponding electrostatic energy for the interaction between the orbitals and the (discretized) COSMO charges we first need the electrostatic potential at
the locations \( \mathbf{r}_i \) of the COSMO points. This requires integrating \( 1/|\mathbf{r} - \mathbf{r}_i| \) over the electron density. With the STO basis functions used in semiempirical methods this is not trivial.

The required integrals cannot be obtained in closed form and there is no fast scheme to evaluate them. This is in contrast with the case of GTO basis sets where such schemes do exist.

In the spirit of semi-empirical methods, we assume that the cross-terms vanish, and for the products of basis functions we use the MNDO\(^{(3), (4)}\) approximation that is also used for the 2-electron repulsion integrals. This will then give the effect of each atom as an expansion in monopoles, dipoles, and quadrupoles. Moreover, instead of considering the multipoles as having a finite spatial extent (‘physical’ multipoles) as in NDDO, we treat them as ‘pure’ multipoles to get straightforward analytical expressions for the electrostatic potential. This scheme does throw away a lot of detail, which may be a problem. But this is the approach that MOPAC \(^{(5), (6)}\) takes for COSMO and it is without doubt the simplest and fastest approach.
For definiteness, we will now give the expressions for the monopole, dipole, linear quadrupole, and square quadrupole potentials as derived from the NDDO multipoles. The monopole simply gives (in atomic units)

\[ V_1(\mathbf{r}) = \frac{e}{r} \]  \hspace{1cm} (S.2)

For the dipole we need to use the \( D_1 \) from the NDDO parameters and take the limit where the charge separation \( d \to 0 \) while keeping \( d \times q \) constant. In the NDDO case, the charge \( q = \pm e/2 \) at distance \( D_1 \). At that point, \( d \times q = \pm D_1 \times e/2 \). The limit then becomes

\[ d \to 0, q = \pm \frac{ed_1}{2d} \]  \hspace{1cm} (S.3)

and we get, for a dipole with charges \( \pm e/2 \) at the location \( (x,y,z) = (\pm d,0,0) \):

\[
V_2(\mathbf{r}) = \lim_{d \to 0} \left[ \frac{q}{\sqrt{(x+d)^2 + y^2 + z^2}} + \frac{-q}{\sqrt{(x-d)^2 + y^2 + z^2}} \right] \\
= \lim_{d \to 0} \frac{ed_1}{2d} \left[ \frac{1}{\sqrt{(x+d)^2 + y^2 + z^2}} - \frac{1}{\sqrt{(x-d)^2 + y^2 + z^2}} \right] \\
= D_1e\frac{r^2}{r^3} \\
= D_1e\frac{x}{r^3} \\
\]

This follows from a series expansion in \( d \) around \( d = 0 \). The quadrupoles can be treated similarly, though the limit is a bit different and we should set the quadrupole moment \( d^2 \).
× q constant as d→0. For the square quadrupole with charges

\[ [ + e/4, - e/4, - e/4, + e/4] \] at the location \((x, y, z) = \]

\[ [(−d, −d,0),(−d, +d,0),( +d, −d,0),( +d, +d,0)] \] we find

\[ V_{4}^{xy} = D_{2}e \frac{3xy}{r^{5}} \quad (S.5) \]

and for the linear quadrupole with \([ + e/4, −e/2, + e/4] \) at the location \((x, y, z) = \]

\[ [(−2d,0,0),(0,0,0),( +2d,0,0)] \] this is

\[ V_{4}^{xx} = D_{2}e \left[ \frac{1}{r^{3}} - \frac{3x^{2}}{r^{5}} \right] \quad (S.6) \]

In this way we can compute the electrostatic potential for a given molecular geometry as a function of the density matrix \(P\) by replacing the electron density from a single matrix element,

\[ \rho_{\mu\nu}(\mathbf{r}) = P_{\mu\nu}\chi_{\mu}(\mathbf{r})\chi_{\nu}(\mathbf{r}), \quad (S.7) \]

by the corresponding multipole (or multipole + monopole for the \(xx, yy\) and \(zz\) orbitals).

Again, note that the off-diagonal blocks of \(P\) are assumed to vanish.
For calculating the MO coefficients this works a bit differently as we do not (yet) have \( P \) in that case. The multipole expansion is then used in much the same way, but will figure directly in the Fock matrix, without reference to the density matrix.

**S4. COSMO-SAC parameterization scheme**

The optimization was loosely based on the COSMO-SAC re-optimization by a team from NIST. All relevant differences (be it pressure, partition coefficient or actual energies) were converted to energies, expressed in kJ/mol. For the pressure we used the same method as the NIST team, namely take the natural logarithm of the ratio of experimental and observed pressure, interpret that as an energy difference in \( k_B T \) and convert to kJ/mol; for partition coefficients, use the natural logarithm instead of the base 10 logarithm, interpret the difference in ‘ln P’ as an energy difference expressed in \( k_B T \), and convert to kJ/mol. In this way, the quadratic errors for all datapoints can be summed to give a single MSE (and thus a single RMSE value that can then be minimized).

However, observing that the typical RMSE for the partitioning data was then about 10 times larger than that for the VLE and \( G_{ex} \) data, we weighed the former with 0.1
compared to the latter (that is, we weighed with 0.01 when summing the quadratic errors to give the objective function to be minimized). We arrived at a typical RMSE for the partitioning of about 4.5 kJ/mol (which corresponds to about 0.8 log10 units). For the VLE and $G_{ex}$ data the typical RMSE is 0.3 kJ/mol. To compare this with the full DFT COSMO calculations and the raw AM1 calculations, see the Table S1.

**Table S1.** COSMO-SAC re-optimization results for AM1, AM1-BCC and DFT COSMO calculations. Deviation from experimental results (RMSE) in kJ/mol and native units.

| Method | VLE pressure $G_{ex}$ | log $P_{octanol/water}$ | log $P_{hexadecane/water}$ |
|--------|-----------------------|-------------------------|-----------------------------|
|        | kJ/mol % kJ/mol $k_BT$ | kJ/mol log$_{10}$       | kJ/mol log$_{10}$           |
| AM1    | 0.394 17% 0.363 0.146 | 4.93 0.864 | 5.15 0.902 |
| ABC    | 0.316 14% 0.286 0.115 | 4.27 0.748 | 4.71 0.825 |
| DFT    | 0.302 13% 0.264 0.106 | 3.65 0.639 | 4.22 0.739 |

These results put the ABC method in between DFT and AM1 for partition coefficients, and close to DFT for VLEs and related calculations.
AUTHOR INFORMATION

Corresponding Author

*E-mail: panos.petris@siemens.com

REFERENCES

(1) Kuhn, B.; Mohr, P.; Stahl, M., Intramolecular Hydrogen Bonding in Medicinal Chemistry. *J. Med. Chem.* **2010**, *53*(6), 2601-2611.

(2) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W., NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, **2010**, *181*(9), 1477-1489.

(3) Dewar, M. J. S.; Thiel, W., Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899-4907.

(4) Dewar, M. J. S.; Thiel, W., A semiempirical model for the two-center repulsion integrals in the NDDO approximation. *Theorem. Chim. Acta*, **1977**, *46*, 89-104.

(5) Stewart P., MOPAC 7, **1993**, [http://openmopac.net/Downloads/Downloads.html](http://openmopac.net/Downloads/Downloads.html)
(6) Stewart P., MOPAC 7.1, 2006, http://openmopac.net/Downloads/Downloads.html

(7) Paulechka, E.; Diky, V.; Kazakov, A.; Kroenlein, K.; Frenkel, M., Reparameterization of COSMO-SAC for phase equilibrium properties based on critically evaluated data. *J. Chem. Eng. Data*, **2015**, *60*, 3554-3561.