Supplementary Information

An Improvement to COSMO-SAC for Predicting Thermodynamic Properties

Ruichang Xiong, Stanley I. Sandler* and Russell I. Burnett

Department of Chemical and Biomolecular Engineering,

University of Delaware, Newark, Delaware, USA

*Author to whom correspondence should be addressed (sandler@udel.edu)

1. Dispersion Derivation

$$E_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}\, m_\alpha m_\beta$$

$$A_{mix}^{disp} = -\frac{\sum_i \sum_j \sum_{\alpha\in i} \sum_{\beta\in j} x_i x_j E_{\alpha\beta}}{\underline{V}_{mix}}$$

$$A_{i/S}^{disp} = \frac{\partial\left(N_{tot} A_{mix}^{disp}\right)}{\partial N_i}\Bigg|_{T,V,N_{1,j\neq i}} = \frac{\partial}{\partial N_i}\left[-\frac{N_{tot}\sum_j \sum_k \sum_{\alpha\in j}\sum_{\beta\in k} x_j x_k E_{\alpha\beta}}{\underline{V}_{mix}}\right]$$

$$= -\frac{\partial}{\partial N_i}\left[\frac{\left(\sum_j N_j\right)}{\left(\dfrac{\sum_j N_j \underline{V}_j}{\sum_j N_j}\right)}\frac{\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} N_j N_k E_{\alpha\beta}}{\left(\sum_j N_j\right)^2}\right] = -\frac{\partial}{\partial N_i}\left[\frac{\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} N_j N_k E_{\alpha\beta}}{\sum_j N_j \underline{V}_j}\right]$$

$$= -\frac{\left(\sum_j N_j \underline{V}_j\right)\left[\dfrac{\partial}{\partial N_i}\left(\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} N_j N_k E_{\alpha\beta}\right)\right] - \left(\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} N_j N_k E_{\alpha\beta}\right)\left[\dfrac{\partial}{\partial N_i}\left(\sum_j N_j \underline{V}_j\right)\right]}{\left(\sum_j N_j \underline{V}_j\right)^2}$$

$$= -\frac{2V_{mix}\left[\sum_i\sum_{j\neq i}\sum_{\alpha\in i}\sum_{\beta\in j} N_j E_{\alpha\beta} + \sum_i\sum_i\sum_{\alpha\in i}\sum_{\beta\in i} N_i E_{\alpha\beta}\right] - \left(\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} N_j N_k E_{\alpha\beta}\right)\underline{V}_i}{V_{mix}^{\ 2}}$$

$$= \frac{\underline{V}_i}{V_{mix}^{\ 2}}\left(\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} N_j N_k E_{\alpha\beta}\right) - \frac{2}{V_{mix}}\left[\sum_i\sum_{j\neq i}\sum_{\alpha\in i}\sum_{\beta\in j} N_j E_{\alpha\beta} + \sum_i\sum_i\sum_{\alpha\in i}\sum_{\beta\in i} N_i E_{\alpha\beta}\right]$$

$$= \frac{\underline{V}_i}{\underline{V}_{mix}^{\ 2}}\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} x_j x_k E_{\alpha\beta} - \frac{2}{\underline{V}_{mix}}\left[\sum_i\sum_{j\neq i}\sum_{\alpha\in i}\sum_{\beta\in j} x_j E_{\alpha\beta} + \sum_i\sum_i\sum_{\alpha\in i}\sum_{\beta\in i} x_i E_{\alpha\beta}\right]$$

$$= \frac{\underline{V}_i}{V_{mix}^{\ 2}}\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} x_j x_k \sqrt{\varepsilon_\alpha \varepsilon_\beta}\, m_\alpha^j m_\beta^k - \frac{2}{\underline{V}_{mix}}\left[\sum_{j\neq i}\sum_{\alpha\in i}\sum_{\beta\in j} x_j \sqrt{\varepsilon_\alpha \varepsilon_\beta}\, m_\alpha^i m_\beta^j + \sum_{\alpha\in i}\sum_{\beta\in i} x_i \sqrt{\varepsilon_\alpha \varepsilon_\beta}\, m_\alpha^i m_\beta^i\right]$$

$$= \frac{\underline{V}_i}{V_{mix}^{\ 2}}\sum_j\sum_k\sum_{\alpha\in j}\sum_{\beta\in k} x_j x_k \sqrt{\varepsilon_\alpha \varepsilon_\beta}\, m_\alpha^j m_\beta^k - \frac{2}{\underline{V}_{mix}}\sum_j\sum_{\alpha\in i}\sum_{\beta\in j} x_j \sqrt{\varepsilon_\alpha \varepsilon_\beta}\, m_\alpha^i m_\beta^j$$

2. Quantum Mechanics (QM) COSMO Calculation Settings

Table S1 shows the settings used in the COSMO calculations at density functional theory (DFT) in the DMol^3 software package.

Table S1. DFT/COSMO calculation settings used in the DMol^3 software package.

Functional	vwn-bp	Cosmo_Solvent_Radius	1.30
Pseudopotential	none	Cosmo_A-Matrix_Cutoff	7.00
Basis_Version	v4.0.0	Cosmo_Radius_Incr	0.000
Basis	dnp	Cosmo_A-Constant	1.882190
Atom_Rcut	5.50 Angstrom	Cosmo_B-Constant	0.010140
Charge	0.000	Cosmo_RadCorr_Incr	0.150000
Spin_Polarization	unrestricted	Mulliken_Analysis	charge
Occupation	Thermal	Hirshfeld_Analysis	charge
Symmetry	on	SCF_Density_Convergence	0.000001
Integration_Grid	fine	SCF_Charge_Mixing	0.200 10.000
Aux_Density	octupole	SCF_DIIS	6
Cosmo	on	SCF_Iterations	50
Cosmo_Dielectric	999.999	SCF_Number_Bad_Steps	13
Cosmo_Grid_Size	1082	SCF_Direct	on
Cosmo_Segments	92	SCF_Restart	off

Table S2 shows the settings used in the COSMO calculations at density functional theory (DFT) in the ADF software package.

Table S2. DFT/COSMO calculation settings used in the ADF software package.

XC potential in SCF	GGA:BP	solvation method	COSMO
XC energy after SCF	Default	solvent	CRS
Relativity (ZORA)	Scalar	Default radii	Klamt
Basis set	TZP	Surface	Delley
Frozen core	Small	Ndiv	4
Integration	Becke Good	Charge determination method	CONJ
Density used in XC-potential	Exact	Convergence	1.00E-06
Total charge	0.0	Max. iterations	300
Spin polarization	0.0	Correct for outlying charge	Yes
Coordinates used for optimization	Delocalized	Calculate Coulomb interaction	EXACT
Optimization method	New	Disc scaling	0.01
Number of geometry iterations	100	Disc Legendre order	4
Maximum number of SCF cycles	50	Disc tolerance	0.1
Convergence method	DIIS	Handle charges	VAR
Mixing	0.2	Charges in which SCF cycles	ALL

3. Liquid Density Fitting

There are 1456 compounds used for the fitting. The correlation between COSMO volume and molecular volume from experiment (expt. volume) at normal boiling point temperatures (NBTs) are shown in **Fig. S1**. The COSMO volume is from COSMO calculations in DMol^3 software package (**Fig. S1**(a)) and ADF software package (**Fig. S1**(b)). The molecular volume from experiment is converted from the liquid density from DIPPR. There is strong polynomial correlation between COSMO volume and expt. volume at NBTs.
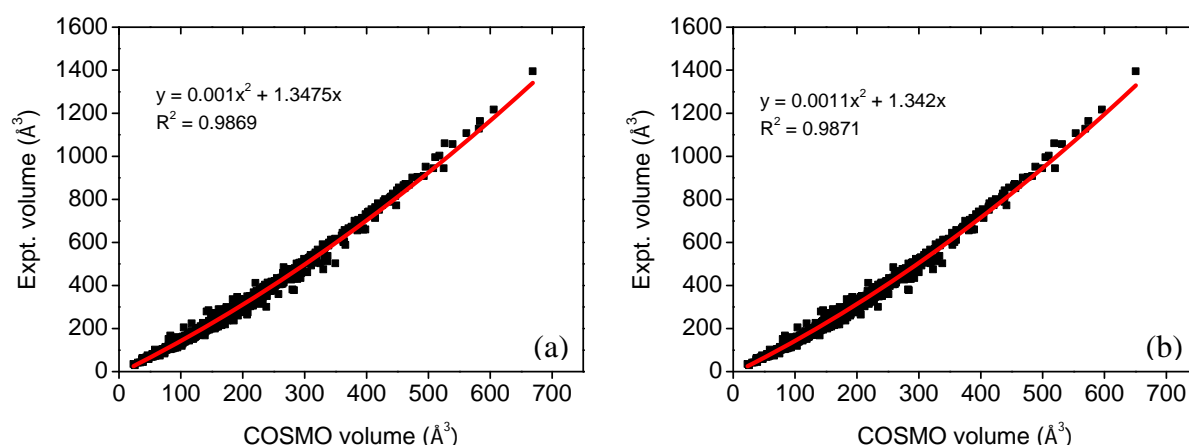

Fig. S1. Correlation between COSMO volume and molecular volume from experiment at normal boiling point temperatures (NBTs): (a) DMol^3 and (b) ADF.

The correlation between COSMO volume and expt. volume at room temperature (RT, 298.15 K) are shown in **Fig. S2**. The COSMO volume from DMol^3 is shown in **Fig. S2**(a), and that from ADF is shown in **Fig. S2**(b). Unlike at NBTs, there is strong linear correlation at RT.
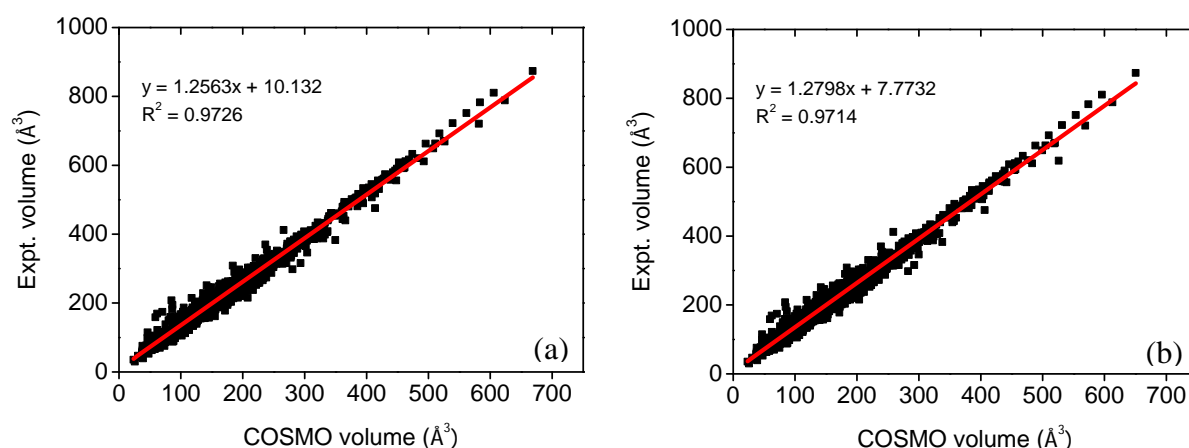

Fig. S2. Correlation between COSMO volume and molecular volume from experiment at room temperature (RT): (a) DMol^3 and (b) ADF.

The correlation between COSMO volume and expt. volume at melting point temperatures (MTs) are shown in **Fig. S3**. The COSMO volume from DMol^3 is shown in **Fig. S3**(a), and that from ADF is shown in **Fig. S3**(b). Again, there is strong linear correlation at MTs.
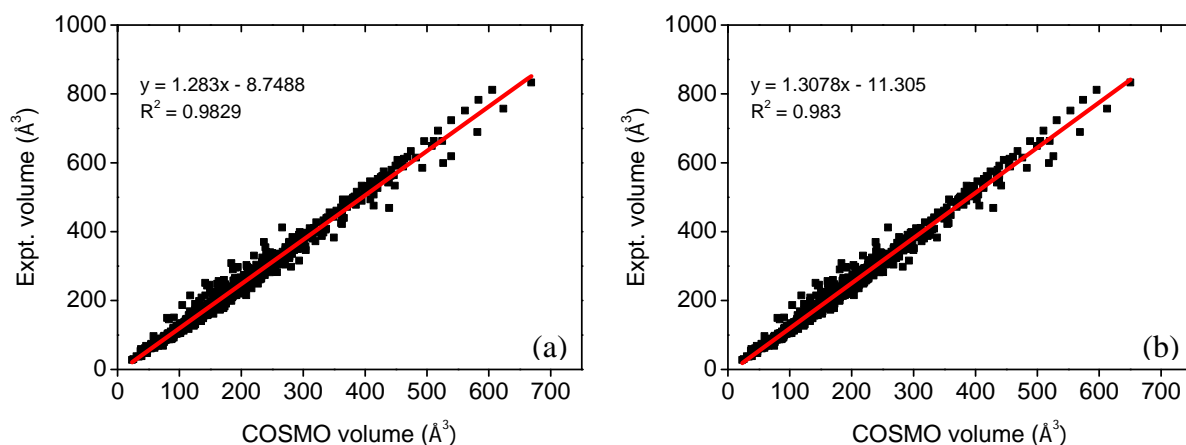

Fig. S3. Correlation between COSMO volume and molecular volume from experiment at melting temperatures (MTs): (a) DMol^3 and (b) ADF.

By combining these data points at the different temperatures, we fit a temperature-dependent relationship between COSMO volume and expt. volume by scaling the temperature to the NBTs since NBT can be estimated from group contribution methods when the experimental data are not available. The temperature-dependent relationship for predicting the molecular volume in DMol^3 platform is

$$V = 0.000971 \times T_r^2 \times V_{cosmo}^2 + \left(0.233 \times T_r^2 + 1.14\right) \times V_{cosmo} - 4.42 \tag{S1}$$

where $T_r = T / T_b$ is the reduced temperature based on the NBT (T_b) and V_{cosmo} is the COSMO volume from the quantum software. Similarly, the temperature-dependent relationship for predicting the molecular volume in ADF platform is

$$V = 0.001040 \times T_r^2 \times V_{cosmo}^2 + \left(0.224 \times T_r^2 + 1.16\right) \times V_{cosmo} - 6.16 \tag{S2}$$

The predicted molecular volume (related to liquid density) is shown in **Fig. S4**. The data in **Fig. S4**(a) are calculated using eqn. S1 from DMol^3 platform, and those in **Fig. S4**(b) are calculated using eqn. S2 from ADF platform.
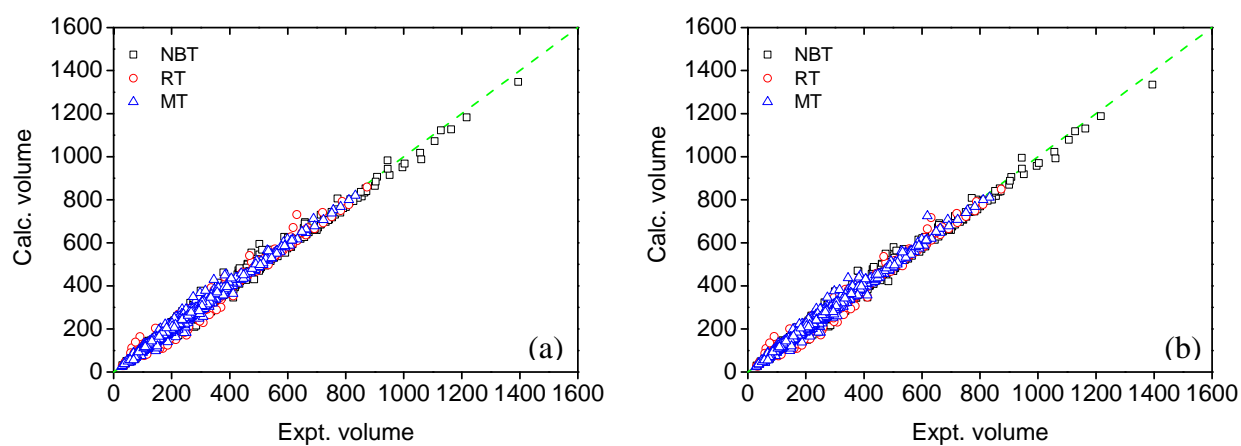
Fig. S4. Predictions of liquid density (molecular volume) from (a) DMol^3 and (b) ADF platform compared to experimental data

4. Turning on Mixture Dispersion in COSMO-SAC (2007) Model

It is found in the manuscript that the dispersion interaction increases prediction precision significantly in our COSMO-SAC (2013) model. It was assumed in the 2007 model that the dispersion effect is negligible and so the dispersion is disabled in the calculation of the mixture thermodynamic properties. It would be curious how the 2007 model works by simply turning on the dispersion interaction using the same parameters provided in the 2007 model for the calculation of the pure fluid properties such as vapor pressures. The results from 2007 model with dispersion are shown in **Fig. S5**. By comparing to the original 2007 model, the 2007 model with dispersion turning on does not improve the accuracy but become worse. This should be because the dispersion parameters were not parameterized for the mixture calculation, not due to that the dispersion does not work. The conclusion here is that by simply turning on the dispersion for 2007 model, the precision is not improved.
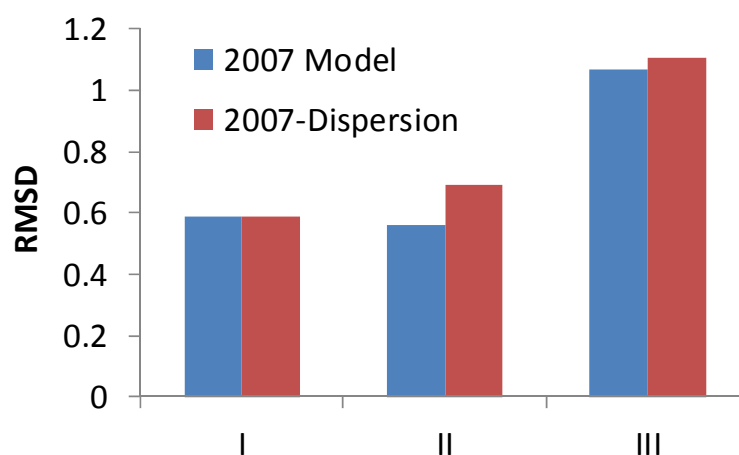

Fig. S5. Comparisons between the 2007 model and the 2007 model turning on dispersion for three different systems: (I) activity coefficient predicitions for 2353 binary systems; (II) octanol-water partition coefficient predicitons for 405 smaller compounds; (III) octanol-water partition coefficient predicitions for 89 larger pharmaceutical compounds.

5. Additional Parameters for the Prediction of Pure Compound Thermodynamic Properties

Additional sets of COSMO-SAC (2013) model parameters for the prediction of pure compound thermodynamic properties are shown in Table S3. These sets of parameters were biased for fitting the pure compound experimental data in the optimization.

Table S3. COSMO-SAC (2013) model parameters biased for the prediction of pure compound thermodynamic properties. In the manuscript, they are referred to as 2013-DMol[3]-pure and 2013-ADF-pure, respectively, for use with DMol[3] and ADF.

		2013-Dmol^3-pure	2013-ADF-pure
Universal Parameters			
Parameter	Unit	Value	Value
a_{eff}	Å^2	0.85	0.75
r_{avg}	Å	7.15	6.95
c_{es}	kcal/mol•Å^4/e^2	8932	9020
$c_{OH\text{-}OH}$	kcal/mol•Å^4/e^2	3572	3324
$c_{OH\text{-}OT}$	kcal/mol•Å^4/e^2	2894	2516
$c_{OT\text{-}OT}$	kcal/mol•Å^4/e^2	1677	1223
q_s	-	0.57	0.60
σ_0	e/Å^2	0.007	0.007
q_0	Å^2	79.53	79.53
Atom Bonding Specific Parameters			
Atom type	R_i^{el} (Å)	ε_i/R (K•Å^3)	ε_i/R (K•Å^3)
H	1.30	0	0
C(sp^3)	2.00	39551	42479
C(sp^2)	2.00	34703	37793
C(sp)	2.00	22333	23743
N(sp^3)	1.83	24851	36047
N(sp^2)	1.83	20995	27820
N(sp)	1.83	3662	5161
O(sp^3-H)	1.72	11033	13186
O(sp^3)	1.72	7726	8258
O(sp^2)	1.72	6932	7716
O(sp^2-N)	1.72	12375	11880
F	1.72	7836	8157
P	2.12	53944	56643
S	2.16	63711	67683
Cl	2.05	40236	42223
Br	2.16	61391	64170
I	2.32	101311	106489

Table S4. Comparison of HB and non-HB surface areas from different models for four selected compounds.

Compound	Model	Surface area (Å^2)		
		non-HB	HB	Total
Ethanol	2007 & 2010	72.2	16.2	88.4
	2013-DMol^3	69.9	18.5	88.4
	2013-ADF	81.1	8.9	90.0
Aniline	2007 & 2010	121.5	14.7	136.2
	2013-DMol^3	119.0	17.2	136.2
	2013-ADF	129.6	7.1	136.7
2-butanone	2007 & 2010	107.5	13.3	120.8
	2013-DMol^3	105.3	15.5	120.8
	2013-ADF	117.0	6.8	123.8
Water	2007 & 2010	16.7	26.6	43.3
	2013-DMol^3	13.3	30.0	43.3
	2013-ADF	28.7	14.3	43.0

Table S5. RMSD for the octanol-water partition coefficients. The compounds are classified based on elements and functional groups. N_c is number of data points in the sample set. The last four columns show the RMSD of 2013 models relative to that of 2010 and 2007 models. The negative sign means the decreased RMSD for the new model. "Mixed" means that the compound contains N and F, Cl, Br, I, P or S. "Mutiple" means that the compound contains two or more strong functional groups such as -OH, -C=O, -COOH, etc.

|	Sample set	Nc	RMSD				Relative RMSD difference to 2010 model		Relative RMSD difference to 2007 model	
			2013-DMol^3	2013-ADF	2010	2007	2013-DMol^3	2013-ADF	2013-DMol^3	2013-ADF
N/F/Cl/Br/I/P/S-containing compounds	N-containing	224	0.90	0.85	0.93	0.92	-3%	-9%	-2%	-8%
	F-containing	19	0.63	0.88	0.64	0.69	-2%	36%	-9%	26%
	Cl-containing	167	0.49	0.51	0.67	0.64	-27%	-24%	-23%	-19%
	Br-containing	16	0.21	0.23	0.25	0.30	-15%	-11%	-28%	-25%
	I-containing	7	0.33	0.27	0.46	0.53	-29%	-41%	-38%	-49%
	P-containing	7	0.64	0.51	0.96	0.71	-33%	-47%	-10%	-28%
	S-containing	23	0.40	0.39	0.43	0.46	-7%	-9%	-12%	-15%
	Mixed	58	0.94	0.88	1.16	0.96	-19%	-24%	-2%	-9%
Others	hydrocarbon	156	0.35	0.36	0.48	0.49	-28%	-25%	-29%	-27%
	acid	60	0.47	0.36	0.46	0.46	2%	-22%	1%	-23%
	alcohol	77	0.45	0.44	0.47	0.51	-4%	-6%	-12%	-14%
	aldehyde	7	0.51	0.37	0.35	0.51	47%	5%	0%	-29%
	ester	60	0.31	0.36	0.41	0.35	-24%	-11%	-12%	4%
	ether	26	0.50	0.62	0.70	0.64	-29%	-11%	-22%	-2%
	furan	3	0.27	0.19	0.36	0.36	-25%	-46%	-24%	-46%
	ketone	27	0.41	0.46	0.46	0.46	-11%	0%	-12%	-1%
	multiple	55	0.85	0.94	0.91	0.84	-6%	4%	2%	13%
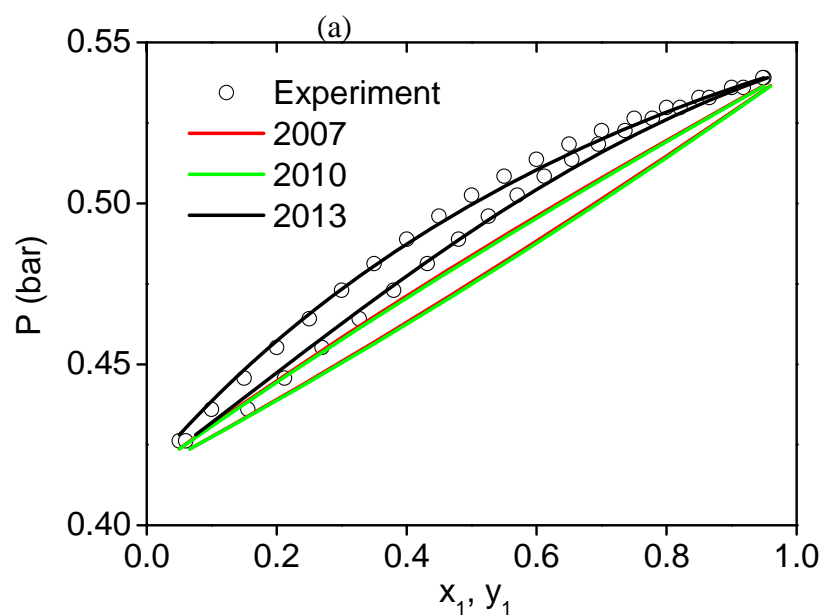
Fig. S6. Comparison of the vapor-liquid equilibrium predictions with experimental data for hexane + tetrachloromethane system at 323.15 K.