Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field

Ian M Kenney, Oliver Beckstein, Bogdan Iorga

Journal of Computer-Aided Molecular Design manuscript No.
(will be inserted by the editor)

Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field

Ian M. Kenney · Oliver Beckstein · Bogdan I. Iorga

Received: date / Accepted: date

Abstract All-atom molecular dynamics (MD) simulations were used to predict water-cyclohexane distribution coefficients D_{cw} of a range of small molecules as part of the SAMPL5 blind prediction challenge. Molecules were parameterized with the transferable all-atom OPLS-AA force field, which required the derivation of new parameters for sulfamides and heterocycles and validation of cyclohexane parameters as a solvent. The distribution coefficient was calculated from the solvation free energies of the compound in water and cyclohexane. Absolute solvation free energies were computed by an established protocol using windowed alchemical free energy perturbation with thermodynamic integration. This protocol resulted in an overall root mean square error (RMSE) in $\log D_{cw}$ of almost 4 log units and an overall signed error of -3 compared to experimental data. There was no substantial overall difference in accuracy between simulating in NVT and NPT ensembles. The signed error suggests a systematic error but the experimental D_{cw} data on their own are insufficient to

OB was supported in part by grant ACI-1443054 from the National Science Foundation. BII was supported in part by grants ANR-10-LABX-33 (LabEx LERMIT) and ANR-14-JAMR-0002-03 (JPIAMR) from the French National Research Agency (ANR).

I. M. Kenney
Department of Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA

O. Beckstein
Department of Physics and Center for Biological Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA
Tel.: +1 480 727 9765
Fax: +1 480 965-4669
E-mail: oliver.beckstein@asu.edu

B. I. Iorga
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Saclay, Labex LERMIT, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France
Tel.: +33 1 69 82 30 94
Fax: +33 1 69 07 72 47
E-mail: bogdan.iorga@cnrs.fr

uncover the source of this error. Preliminary work suggests that the major source of error lies in the hydration free energy calculations.

Keywords molecular dynamics · solvation free energy · OPLS-AA force field · ligand parameterization · free energy perturbation · thermodynamic integration · cyclohexane-water distribution coefficients

1 Introduction

The distribution coefficient D_{AB} of a small molecule quantifies the partitioning of a molecule between two immiscible phases A and B. Of particular importance in drug discovery are distribution coefficients between the aqueous phase and hydrophobic solvents, which mimic to some degree biological hydrophobic environments such as the lipid bilayer of the cell membrane. Distribution coefficients can be used to describe and model the distribution of molecules in chemical and biological systems. In the drug discovery process, they are key quantities for the design of drugs that can diffuse across cell membranes (blood brain barrier, epithelial lining of the gut) and thus reach their site of action inside the body or a cell itself [1].

In principle, distribution coefficients should also be good benchmark systems for the evaluation of the predictive power of quantitative computational methods [2], similar to the hydration free energy calculations of previous SAMPL challenges [3–7].

For the SAMPL5 challenge we employed classical all-atom molecular dynamics (MD) simulations in explicit solvent with additive and transferable force fields to predict distribution coefficients. We are also interested in the question if distribution coefficients might be useful target observables in the process of parameterizing small molecules and drug-like compounds.

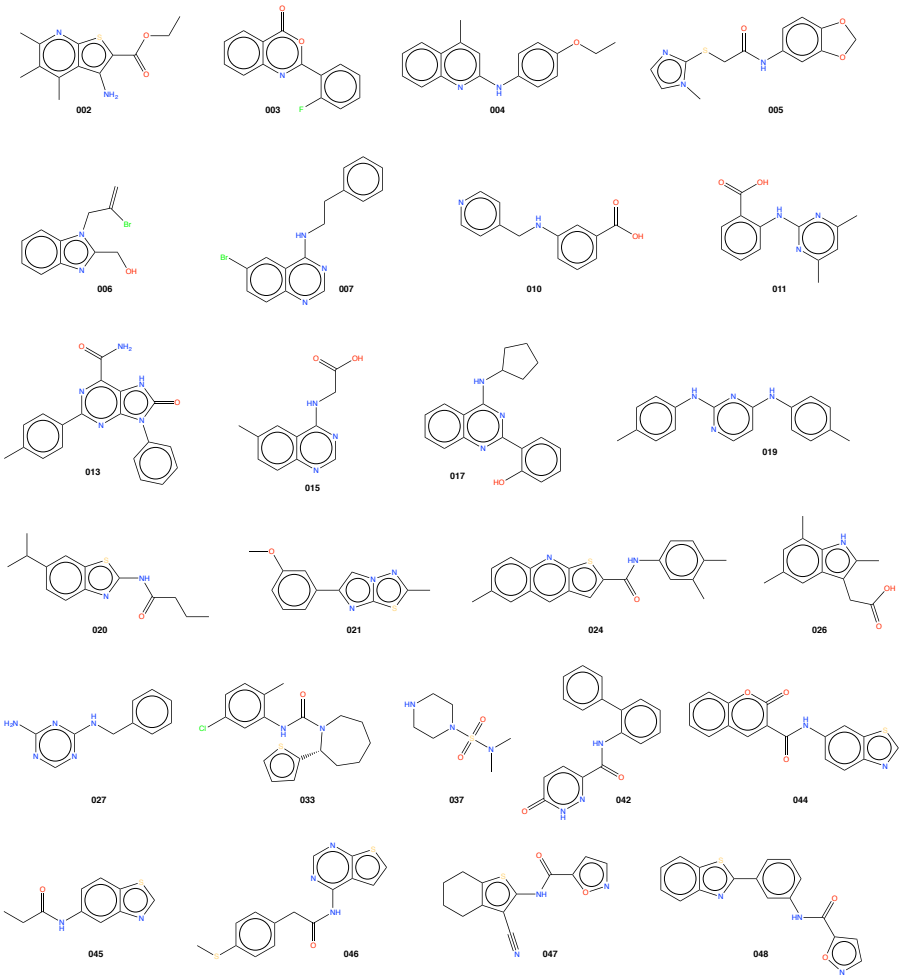
The data set provided for the SAMPL5 challenge consisted of 53 small, drug-like molecules (Figures 1a and 1b) for which water-cyclohexane distribution coefficients [8] were measured at Genentech using a mass spectrometry-based assay [9]. Their distribution coefficients had not been published but were known to the SAMPL5 organizers.

Here we employed explicit solvent MD simulations in conjunction with a windowed alchemical free energy perturbation approach to compute absolute solvation free energies of the compounds in water (ΔG_w) and cyclohexane (ΔG_c). The cyclohexane-water partition coefficient D_{cw} (or rather, its base-10 logarithm, indicated by log) is then

$$\log D_{cw} = (\Delta G_w - \Delta G_c)(kT)^{-1} \log e, \tag{1}$$

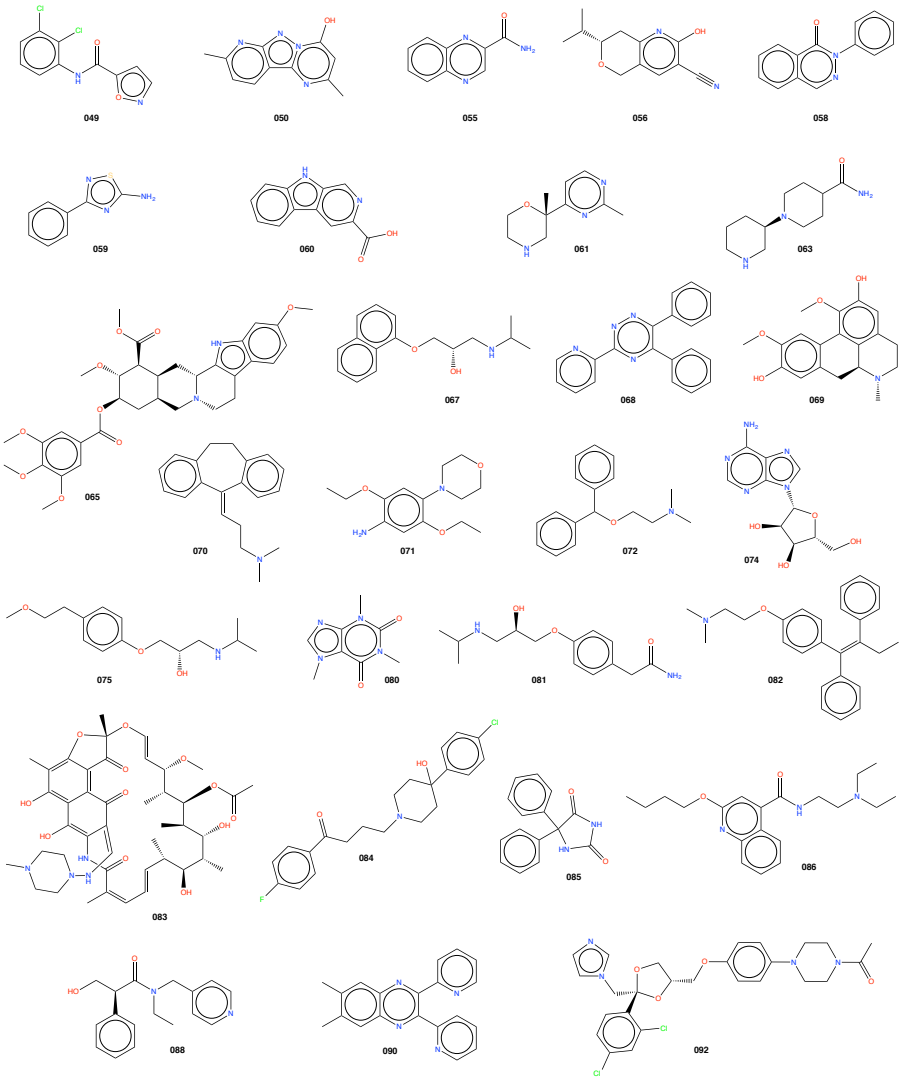
where $k = 1.987207 \times 10^{-3}\,\mathrm{kcal \cdot mol^{-1} \cdot K^{-1}}$ is Boltzmann's constant, T is the temperature, and e Euler's number. If $\log D_{cw}$ is zero then the solute is equally likely to be found in the water and the cyclohexane phase. A $\log D_{cw} < 0$ indicates that the polar water phase is favored whereas for $\log D_{cw} > 0$ the hydrophobic cyclohexane is preferred.

The interactions between all atoms in the system were parameterized on the basis of the classical OPLS-AA force field [10]. OPLS-AA is a transferable force field



1a Chemical structures of the 53 compounds included in the SAMPL5 data set.

with partial atomic charges determined from calculations on small model compounds; unlike other classical force fields, these partial charges are considered fixed and part of the atom type in the same way as the Lennard-Jones potential parameters. This leads to a rich set of atom types that can be directly applied to an atom in another molecule that experiences the same chemical environment as the atom in the model compound. Thus, in principle OPLS-AA is a good force field for the parameterization of small molecules based on chemical rules without the requirement of molecule-specific adaptations such as additional partial charge calculations.



1b Chemical structures of the 53 compounds included in the SAMPL5 data set (continued).

2 Methods

Calculations were performed with protocols similar to our previous work in the SAMPL3 [11] and SAMPL4 [12] challenges but for completeness we describe the essential details below.

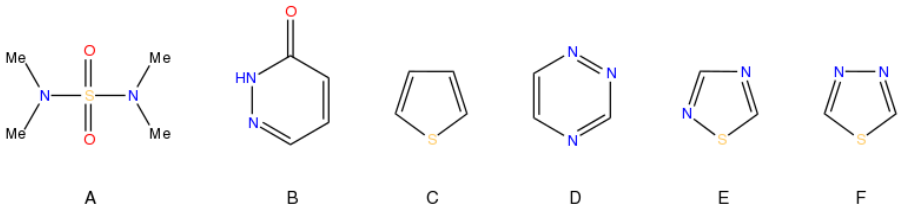


Fig. 2: Chemical structures of the motifs for which new OPLS-AA parameters had to be generated; see Table 1 for the parameters.

2.1 Force field parameters

The SAMPL5 data set contained 53 compounds (see Figures 1a and 1b together with their SAMPL5 ID numbers), which were parameterized in the protonation and tautomeric forms as provided by the organizers, with the exception of compound **042** for which tautomeric forms were considered (see below). The SMILES string of each compound was converted to PDB format with CORINA version 3.60 (`http://www.molecular-networks.com`). Molecules were parameterized with the OPLS-AA all-atom force field [13–19]. The OPLS-AA force field files that were bundled with Gromacs 4.6.5 [20, 21] were used as a starting point and extended with parameters found in the literature [11, 12, 22–28]. Topologies were generated using the MOL2FF algorithm (O. Beckstein and B. I. Iorga, unpublished), which automatically assigns OPLS-AA atom types based on the chemical function as determined by the CACTVS Chemoinformatics Toolkit (`http://www.xemistry.com/`). All force field parameters for the SAMPL5 compounds were deposited in the Ligandbook repository `https://ligandbook.org` [29] (see Table 3 for the individual LigandbookIDs).

The OPLS-AA force field has been parameterized together with the TIP4P water model [13–19] and in order to remain consistent with this standard choice, we also employed the standard TIP4P parameters [30]. For simulations with cyclohexane we generated parameters with MOL2FF and tested them with simulations of bulk cyclohexane (see Results).

2.2 Parameterization of missing OPLS-AA parameters

Force field parameters were not available in the published OPLS-AA force field for a number of chemical groups (*N*-substituted sulfamide, 2*H*-pyridazin-3-one, thiophene, 1,2,4-triazine, 1,2,4-thiadiazole and 1,3,4-thiadiazole, see Figure 2) that were present in the SAMPL5 data set. CM5-derived charges were calculated for these chemical groups using Gaussian09 (revision D.01) [31] and following the protocol recently described by Jorgensen and colleagues [32, 33], with a scaling factor of 1.20. These newly generated parameters are presented in Table 1. When necessary, other missing atom types, bond, angle and dihedral bonding parameters were adapted from the existing ones using the original OPLS-AA philosophy or obtained in a manner similar to the published OPLS-AA protocol [14].

Table 1: New OPLS-AA parameters for *N*-substituted sulfamide, 2*H*-pyridazin-3-one, thiophene, 1,2,4-triazine, 1,2,4-thiadiazole and 1,3,4-thiadiazole chemical groups (CM5 charges).

name[a]	type[b]	Z[c]	m (u)[d]	q (e)[e]	σ (nm)[f]	ε (kJ·mol^{-1})[g]	
opls_474A	SY	16	32.0600	0.852	0.355	1.046000	; **S** in sulfamide
opls_475A	OY	8	15.9994	−0.444	0.296	0.711280	; **O** in sulfamide
opls_480A	N	7	14.0067	−0.464	0.325	0.711280	; **N** in sulfamide
opls_483A	HC	1	1.0080	0.132	0.250	0.125520	; C**H**$_x$–N in sulfamide
opls_482A	CT	6	12.0110	−0.155	0.350	0.276144	; **C**H$_3$–N in sulfamide
opls_484A	CT	6	12.0110	−0.023	0.350	0.276144	; R**C**H$_2$–N in sulfamide
opls_484B	CT	6	12.0110	0.109	0.350	0.276144	; R$_2$**C**H–N in sulfamide
opls_484C	CT	6	12.0110	0.241	0.350	0.276144	; R$_3$**C**–N in sulfamide
opls_750A	NZ	7	14.0067	−0.282	0.325	0.711280	; N1 in 2*H*-pyridazin-3-one
opls_238A	N	7	14.0067	−0.352	0.325	0.711280	; N2 in 2*H*-pyridazin-3-one
opls_235A	C	6	12.0110	−0.312	0.375	0.439320	; C3 in 2*H*-pyridazin-3-one
opls_142A	CM	6	12.0110	−0.088	0.355	0.317984	; C4 in 2*H*-pyridazin-3-one
opls_142B	CM	6	12.0110	−0.084	0.355	0.317984	; C5 in 2*H*-pyridazin-3-one
opls_277A	C_2	6	12.0110	0.060	0.375	0.439320	; C6 in 2*H*-pyridazin-3-one
opls_236A	O	8	15.9994	−0.456	0.296	0.878640	; O3 in 2*H*-pyridazin-3-one
opls_241A	H	1	1.0080	0.434	0.000	0.000000	; H2 in 2*H*-pyridazin-3-one
opls_144A	HC	1	1.0080	0.154	0.242	0.125520	; H4 in 2*H*-pyridazin-3-one
opls_144B	HC	1	1.0080	0.146	0.242	0.125520	; H5 in 2*H*-pyridazin-3-one
opls_279A	HC	1	1.0080	0.156	0.242	0.062760	; H6 in 2*H*-pyridazin-3-one
opls_633A	S	16	32.0600	0.104	0.355	1.046000	; S in thiophene
opls_567A	CW	6	12.0110	−0.172	0.355	0.292880	; C2 in thiophene
opls_568A	CS	6	12.0110	−0.152	0.355	0.317984	; C3 in thiophene
opls_569A	HA	1	1.0080	0.140	0.242	0.125520	; H2 in thiophene
opls_570A	HA	1	1.0080	0.132	0.242	0.125520	; H3 in thiophene
opls_641A	N	7	14.0067	−0.244	0.325	0.711280	; N1 in 1,2,4-triazine
opls_641B	N	7	14.0067	−0.264	0.325	0.711280	; N2 in 1,2,4-triazine
opls_642A	CQ	6	12.0110	−0.248	0.355	0.292880	; C3 in 1,2,4-triazine
opls_641C	N	7	14.0067	−0.422	0.325	0.711280	; N4 in 1,2,4-triazine
opls_145A	CA	6	12.0110	0.106	0.355	0.292880	; C5 in 1,2,4-triazine
opls_145B	CA	6	12.0110	0.077	0.355	0.292880	; C6 in 1,2,4-triazine
opls_643A	HA	1	1.0080	0.174	0.242	0.125520	; H3 in 1,2,4-triazine
opls_643B	HA	1	1.0080	0.163	0.242	0.125520	; H5 in 1,2,4-triazine
opls_643C	HA	1	1.0080	0.162	0.242	0.125520	; H6 in 1,2,4-triazine
opls_633B	S	16	32.0600	0.284	0.355	1.046000	; S in 1,2,4-thiadiazole
opls_635A	NB	7	14.0067	−0.443	0.325	0.711280	; N2 in 1,2,4-thiadiazole
opls_634A	CR	6	12.0110	0.211	0.355	0.292880	; C3 in 1,2,4-thiadiazole
opls_635B	NB	7	14.0067	−0.461	0.325	0.711280	; N4 in 1,2,4-thiadiazole
opls_634B	CR	6	12.0110	0.058	0.355	0.292880	; C5 in 1,2,4-thiadiazole
opls_638A	HA	1	1.0080	0.178	0.242	0.125520	; H3 in 1,2,4-thiadiazole
opls_638B	HA	1	1.0080	0.173	0.242	0.125520	; H5 in 1,2,4-thiadiazole
opls_633C	S	16	32.0600	0.118	0.355	1.046000	; S in 1,3,4-thiadiazole
opls_634C	CR	6	12.0110	0.067	0.355	0.292880	; C2 in 1,3,4-thiadiazole
opls_635C	NB	7	14.0067	−0.301	0.325	0.711280	; N3 in 1,3,4-thiadiazole
opls_638C	HA	1	1.0080	0.175	0.242	0.125520	; H2 in 1,3,4-thiadiazole

[a] proposed OPLS-AA atom type name [b] bonded type [c] atomic number
[d] atomic mass in atomic mass units $m_u = 1.660538921 \times 10^{-27}$ kg [e] partial charge in elementary charges $e = 1.602176565 \times 10^{-19}$ C
[f] length parameter of the OPLS-AA Lennard-Jones potential $V_{LJ}(r) = 4\varepsilon[(\sigma/r)^{12} - (\sigma/r)^6)]$ [10]
[g] energy well depth of the OPLS-AA Lennard-Jones potential $V_{LJ}(r)$

In this SAMPL5 challenge we tentatively evaluated the parameterization of fused rings (for which no parameters are available in the OPLS-AA force field) using the parameters of the individual rings. This approach is not fully validated yet , but if it proves to be useful it will greatly facilitate the parameterization of new heterocyclic systems and—following the OPLS-AA philosophy—extend the modularity and the transferability of the parameters.

Compound **042** also provided a good opportunity to test the parameterization of two alternative tautomeric forms of a heterocycle, i.e., aromatic (3-hydroxy-pyridazine) and non-aromatic (2*H*-pyridazin-3-one, which represents the structure provided in the SAMPL5 data set).

From the parameterization point of view, the compounds from the SAMPL5 data set can be classified in four groups: *group 1*, with compounds containing chemical moieties available in the OPLS-AA force field: **004**, **010**, **011**, **019**, **026**, **027**, **049**, **056**, **061**, **063**, **065**, **067**, **069**, **070**, **071**, **072**, **074**, **075**, **081**, **082**, **084**, **086**, **088**; *group 2*, with compounds containing chemical moieties absent from the OPLS-AA force field, that were parameterized during this work: **005**, **033**, **042**, **047**, **059**, **068**, **092**; *group 3*, with compounds containing fused rings for which parameterization used the parameters of individual rings: **007**, **015**, **017**, **020**, **044**, **045**, **048**, **050**, **055**, **060**, **090**; *group 4*, with compounds presenting a combination of the issues mentioned above, chemical moieties difficult to parameterize, and high conformational complexity: **002**, **003**, **006**, **013**, **021**, **024**, **037**, **046**, **058**, **080**, **083**, **085**.

During the SAMPL5 challenge preliminary calculations of $\log D_{cw}$ for a few simple compounds with known solvation free energies in water and cyclohexane (ΔG_w and ΔG_c) showed that the distribution coefficient could be predicted with an error of about 0.9 logD units (data not shown). In our submission of the SAMPL5 predictions, this value was used as estimated uncertainty of the method for groups 1 and 2, and was increased arbitrarily to 1.1 and 1.3 for groups 3 and 4 to account for the more difficult parameterization.

2.3 Conformational flexibility

Considering the size and the macrocyclic structure of compound **083**, two different conformations (the one provided in the SAMPL5 data set and a second one, generated using CORINA) were considered as input structures for our protocol. By using different starting structures for the simulations we wanted to evaluate the sensitivity of the results to the initial conditions.

2.4 Hydration free energy and distribution coefficient calculation

Solvation free energies were calculated via alchemical free energy perturbation (FEP) MD simulations of each molecule in a water box. All simulations were performed with the MDPOW Python package (`https://github.com/Becksteinlab/MDPOW`) with Gromacs 4.6.5 [21, 34] as its MD engine. A periodic rhombic dodecahedral simulation cell was employed. In simulations with water, the minimum distance between

a solute and a box face was 1 nm whereas this distance was increased to 1.5 nm for cyclohexane as solvent.

The simulations were run as Langevin dynamics (integration time step 2 fs) for temperature control, with the friction coefficient for each particle computed as mass/0.1 ps [35]. For simulations in the *NPT* ensemble, the average pressure was maintained near the target value 1 bar with an isotropic Parrinello-Rahman barostat [36] with relaxation time constant $\tau_p = 1$ ps and compressibility $\kappa_T = 4.6 \times 10^{-5}$ bar^{-1}. The grid-based neighbor list was updated every five time steps. Lennard-Jones interactions were calculated up to a cutoff of 1 nm and a dispersion correction was applied to energy and pressure to account for van der Waals interactions beyond the cutoff in a mean field manner [37]. Coulomb interactions were evaluated with the SPME method [38] with a short range cutoff of 1 nm, 0.12 nm Fourier grid spacing, sixth order spline interpolation, and a relative tolerance of 10^{-6}. All bonds containing hydrogen atoms were constrained with the P-LINCS algorithm [39] using a twelfth order expansion with a single iteration.

Solvated systems were energy minimized and carefully relaxed with an MD simulation with a time step of 0.1 fs and duration of 5 ps. An initial equilibrium simulation at constant temperature and pressure ($T = 300$ K, $P = 1$ bar) was carried out for 15 ns. The last frame of the equilibrium simulation served as the starting configuration for the windowed FEP calculations. The FEP calculations were carried out (1) in the *NVT* ensemble and (2) in the *NPT* ensemble; in previous work we had exclusively used the *NVT* ensemble for the FEP calculations [11, 12] so we wanted to evaluate if there was a measurable difference between results from the two ensembles. Coulomb interactions (partial charges) were linearly switched off over five windows (coupling parameter $\lambda_{\text{Coul}} \in \{0, 0.25, 0.5, 0.75, 1\}$) while the van der Waals (Lennard-Jones) interactions were maintained (i.e. $\lambda_{\text{vdW}} = 0$); sixteen windows were used to switch off the Lennard-Jones term for the uncharged solute ($\lambda_{\text{Coul}} = 1$ and $\lambda_{\text{vdW}} \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1\}$). Each window was simulated for 5 ns. The van der Waals calculations used soft core potentials with the values suggested by Mobley and colleagues [35] ($\alpha = 0.5$, power 1, and $\sigma = 0.3$ nm). The calculations made use of the "`couple-intramol = no`" feature in Gromacs [20, 21, 34], which maintains intramolecular interactions while decoupling all intermolecular ones. Solvation free energies and statistical errors for the discharging and decoupling process were calculated with thermodynamic integration

$$\Delta G = \int_0^1 \left\langle \frac{\partial \mathscr{H}}{\partial \lambda} \right\rangle d\lambda, \tag{2}$$

where the derivative of the Hamiltonian $\mathscr{H}$ with respect to the coupling parameter λ, $\partial \mathscr{H} / \partial \lambda$, was saved for every time step. Eq. 2 was integrated numerically with the composite Simpson's rule [40] as implemented in SciPy (`http://www.scipy.org`) [41]. The error on ΔG was calculated by propagating the errors of the individual $\langle \partial \mathscr{H} / \partial \lambda \rangle$ FEP windows through Simpson's rule as described previously [11].

The total solvation free energy (transfer from gas phase to aqueous phase at the 1M/1M Ben-Naim standard state)

$$\Delta G_{\text{solv}} = -(\Delta G_{\text{Coul}} + \Delta G_{\text{vdW}}) \tag{3}$$

is the sum of the Coulomb and van der Waals contributions, with the minus sign originating from the convention in Gromacs that $\lambda = 0$ corresponds to the fully coupled (solvated) state while $\lambda = 1$ describes a fully decoupled (gas-phase) solute.

In principle the distribution coefficient contains an average over all protonation and tautomeric states of the compound, which we do not take into account in our calculations. Instead we are calculating solvation free energies (Eq. 3) for one fixed state of the compound. The corresponding partition coefficient

$$\log P_{cw} = (\Delta G_w - \Delta G_c)(kT)^{-1} \log e \tag{4}$$

is only valid for the specific state of the molecule but we nevertheless make the approximation

$$\log D_{cw} \approx \log P_{cw}. \tag{5}$$

2.5 Error analysis

The error ε on $\log D_{cw}$ was calculated by error propagation from the errors of the individual free energies as

$$\varepsilon = \sqrt{\varepsilon^2_{\Delta G_c} + \varepsilon^2_{\Delta G_w}}(kT)^{-1} \log_{10} e. \tag{6}$$

The difference between experimental and computed water-cyclohexane distribution coefficients ("signed error") for each compound, labeled with its identification code 'id', was calculated as

$$\Delta_{\mathrm{id}} = \log D^{\mathrm{exp}}_{cw,\mathrm{id}} - \log D_{cw,\mathrm{id}}, \tag{7a}$$

$$\varepsilon_{\Delta,\mathrm{id}} = \sqrt{(\varepsilon^{\mathrm{exp}}_{\mathrm{id}})^2 + \varepsilon^2_{\mathrm{id}}}, \tag{7b}$$

with the uncertainty ε_Δ of Δ determined as the standard error from propagating the experimental and simulation errors (Eq. 6) through Eq. 7a.

The root mean square error (RMSE) was determined from the individual errors Δ as

$$\mathrm{RMSE} = \sqrt{\langle \Delta^2 \rangle} = \sqrt{N^{-1} \sum_{\mathrm{id}}^{N} \Delta^2_{\mathrm{id}}}, \tag{8}$$

the absolute unsigned error (AUE) as

$$\mathrm{AUE} = \langle |\Delta| \rangle = N^{-1} \sum_{\mathrm{id}}^{N} |\Delta_{\mathrm{id}}|, \tag{9}$$

and the signed mean error (ME, also called the "mean signed error", MSE) as

$$\mathrm{ME} = \langle \Delta \rangle = N^{-1} \sum_{\mathrm{id}}^{N} \Delta_{\mathrm{id}}. \tag{10}$$

The standard errors of the RMSE, AUE, and ME were estimated via error propagation of the individual uncertainties Eq. 7b through Eqs. 8–10 as

$$\varepsilon_{\mathrm{RMSE}} = \frac{1}{N\,\mathrm{RMSE}}\sqrt{\sum_{\mathrm{id}}^{N}\Delta_{\mathrm{id}}^{2}\varepsilon_{\Delta,\mathrm{id}}^{2}} = \frac{1}{\sqrt{N}}\sqrt{\frac{\langle(\Delta\varepsilon_{\Delta})^{2}\rangle}{\langle\Delta^{2}\rangle}}, \tag{11a}$$

$$\varepsilon_{\mathrm{ME}} = \varepsilon_{\mathrm{AUE}} = \frac{1}{\sqrt{N}}\sqrt{\langle\varepsilon_{\Delta}^{2}\rangle}. \tag{11b}$$

Eq. 11a followed the derivation of the root mean square error of prediction in Ref. [42] but remains more conservative by omitting a correction factor of $1/\sqrt{2}$.

3 Results and Discussion

The SAMPL5 set consisted of challenging compounds that required the introduction of a number of new OPLS-AA atom types. The computed distribution coefficients generally differed systematically from the experimental values, without any clear, discernible pattern based on the chemical character of the compounds. We discuss potential sources for the observed systematic error.

3.1 Validation of cyclohexane parameters

Cyclohexane was parameterized with the standard OPLS-AA alkane parameters, following the original work [14]. The parameterization was validated by (1) computing the density as a function of temperature, (2) calculation of the chemical potential and (3) calculation of the hydration free energy and comparison to experimental values.

The bulk density of cyclohexane was calculated from simulations with 140 cyclohexane molecules (cubic simulation cell with length 3 nm) of 100 ns length at temperatures from 273 K to 353 K and $P = 1$ bar (Table 2). Experimental data from 228 experiments in the temperature range 273 K to 353 K were obtained from the Reaxys database and compared to the computed values (Figure 3). A few experimental data points and one computed value are below the melting point (279.47 K [43]) and represent supercooled liquid; all reported values are below the boiling point (353.7 K [43]). Over the whole liquid range, the simulations slightly underestimate the density between -1% at low temperatures and -3.5% near the boiling point. At $T = 300$ K, the error is -2% but the computed density $0.7595 \pm 0.0001\,\mathrm{g\cdot cm^{-3}}$ (standard error of the mean) is close to the density $0.755 \pm 0.001\,\mathrm{g\cdot cm^{-3}}$ at 298 K that was reported for the original OPLS-AA parameterization [14]. Overall, the parameterization reproduces the density of cyclohexane satisfactorily.

The chemical potential of cyclohexane $\mu^{\mathrm{cyclohexane}}$ is the transfer free energy of a cyclohexane molecule from vacuum to the pure cyclohexane solvent, $\Delta G_c^{\mathrm{cyclohexane}}$. We calculated $\Delta G_c^{\mathrm{cyclohexane}}$ with the FEP protocol described above. The computed value is -4.00 ± 0.06 kcal/mol, which matches the experimental value -4.43 kcal/mol [44] rather well. The hydration free energy of cyclohexane $\Delta G_w^{\mathrm{cyclohexane}}$ was calculated in a similar way and the calculated value of 2.03 ± 0.05 kcal/mol agrees fairly well (< 1 kcal/mol error) with the experimental value 1.23 ± 0.60 kcal/mol [44].

Table 2: Density of cyclohexane.

T (K)	ρ_{exp} (g·cm^{-3}) [a]	ρ_{sim} (g·cm^{-3}) [b]	rel.error [c]
273	0.7967	0.7898(2)	−0.9%
300	0.7737(1)	0.7595(1)	−1.9%
310	0.7636(1)	0.7480(1)	−2.2%
350	0.7241	0.7001(1)	−3.5%

[a] Experimental densities were not available at the simulated temperature T so we estimated the density as an average over experimental values within $T \pm 2$ K; the error indicates the spread over this range as half of the difference between larges and smalles value. When no error is given, only a single value was found in the range. [b] Errors for simulated densities were estimated from a block average over five blocks of 20 ns length each and represent the standard error of the mean. [c] The relative error to experiment was calculated as $\rho_{sim}/\rho_{exp} - 1$.

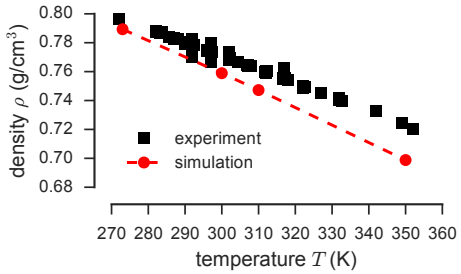


Fig. 3: Dependence of the density of cyclohexane on the temperature. Black squares are experimental data; red circles were computed from 100-ns MD simulations. The red dashed line was drawn to guide the eye.

Based on the good agreement of the calculated density and the free energies of solvation of cyclohexane in water and in cyclohexane with experimental values, we consider the cyclohexane parameters validated.

3.2 Parameterization of new OPLS-AA atom types

A number of compounds from the SAMPL5 data set required atom-types for several chemical groups that were absent from the OPLS-AA force field: *N*-substituted sulfamide (**037**), 2*H*-pyridazin-3-one (**042**), thiophene (**002**, **024**, **033**, **046** and **047**), 1,2,4-triazine (**068**), 1,2,4-thiadiazole (**059**) and 1,3,4-thiadiazole (**021**) (Figure 2). We have generated these missing atom types using CM5 charges [45], following the recent work of Jorgensen and colleagues [32, 33] (Table 1). The availability of CM5 charges, derived from Hirshfeld charges, in the GAUSSIAN program [31] (starting

with version 09 revision D.01) has considerably simplified the applicability of this new parametrization protocol.

We could only validate these parameters for thiophene, with simulations carried out on thiophene and 2-methyl-thiophene for which experimental values of hydration free energies are available (-1.42 and -1.38 kcal/mol, respectively) [46]. The computed values are provided in Table S2 (Electronic supplementary material), showing an error in the prediction of -1.18 kcal/mol (NVT) and -1.56 kcal/mol (NPT) for thiophene, and -1.00 kcal/mol (NVT) and -1.39 kcal/mol (NPT) for 2-methyl-thiophene. These relatively high values might be related to a systematic error that we suspect to be present in our predictions (see discussion below).

This protocol seems to give relatively good results, at least when the subset parameterized with CM5 charges is compared to the whole data set. As will be discussed in more detail below, the root mean squared error (RMSE) for the whole data set (in NPT) is 3.95 ± 0.05 (standard error of the RMSE) and the absolute unsigned error (AUE) is 3.49 ± 0.05 (Table 3). RMSE and AUE of $\log D_{cw}$ predicted for the CM5 subset were better than the whole data set with 3.33 ± 0.10 and 2.87 ± 0.10 (NPT conditions), respectively. Although a definitive conclusion cannot be drawn from this small subset of ten SAMPL5 compounds out of 53, the results are encouraging and CM5 charges appear to be a promising approach.

The SAMPL5 data set contained several compounds presenting heterocyclic systems with fused rings, for which no parameters were available in the OPLS-AA force field: thieno[2,3-*b*]pyridine (**002**), 1*H*-benzo[*d*]imidazole (**006**), benzo[*d*]thiazole (**020**, **044**, **045**, **048**), imidazo[2,1-*b*][1,3,4]thiadiazole (**021**), thieno[2,3-*b*]quinoline (**024**), thieno[2,3-*d*]pyrimidine (**046**), pyrido[2',3':3,4]pyrazolo[1,5-*a*]pyrimidine (**050**), 9*H*-pyrido[3,4-*b*]indole (**060**). For all these heterocyclic systems we tentatively evaluated an original approach involving the use of force field parameters of the individual rings composing these systems. The charge of the 'bridgehead' atoms is obtained by summing the charges of the corresponding atoms in the individual rings, and those of the hydrogen atoms connected to them. The overall RMSE and AUE of $\log D_{cw}$ values predicted for these 11 compounds with non-parameterized fused-ring heterocyclic systems were 3.59 ± 0.11 and 3.09 ± 0.10 (NPT conditions), which are similar to those obtained for the ensemble of 53 compounds from the SAMPL5 data set (Table 3).

Finally, an analysis of the results obtained for compound **042** with two alternative aromatic and non-aromatic tautomeric forms shows that the non-aromatic 2*H*-pyridazin-3-one form of the heterocycle gives better results than the aromatic 3-hydroxy-pyridazine one (error in the $\log D_{cw}$ prediction of -2.85 ± 0.32 and -5.22 ± 0.32, respectively). These results are in agreement with the fact that the non-aromatic 2*H*-pyridazin-3-one is the major form (5.68 kcal/mol more stable than the aromatic 3-hydroxy-pyridazine, according to DFT calculations carried out at the M06-2X/6-311+G(2df,2p)//B3LYP/6-31G(d) level using GAUSSIAN09 [31]), highlighting the importance of considering all representative forms of the compound of interest in the prediction of solvation free energies or distribution coefficients.

For the parameterization of compounds **005** and **092** we used the parameters of *N*-methyl-imidazole that we generated in our previous work [12]. For these compounds we obtained errors in the $\log D_{cw}$ prediction of 0.17 and -5.12 (NPT), re-

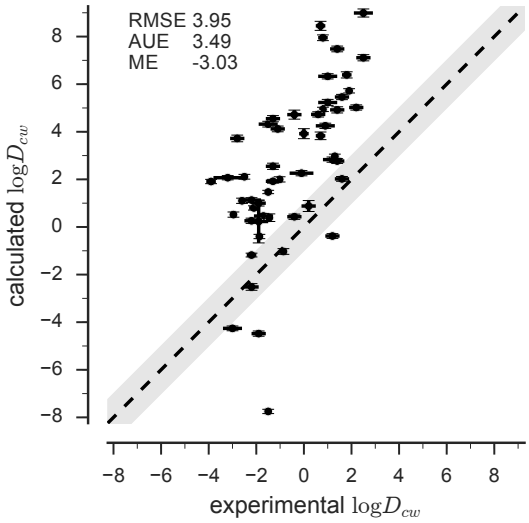


Fig. 4: Correlation between experimental and computed water-cyclohexane distribution coefficients $\log D_{cw}$ for simulations performed in the *NPT* ensemble. The gray band indicates ± 1 log-units from ideal correlation, shown by the dashed line. The root mean square error (RMSE), the absolute unsigned error (AUE), and the (signed) mean error (ME) are indicated. Error bars represent the error in the experiments or the error on the mean, derived from the simulations.

spectively. However, with only two compounds and such a spread in error, validation of the methyl-imidazole parameters by comparison to the $\log D_{cw}$ data is not feasible.

3.3 Predicted distribution coefficients

For each molecule, absolute solvation free energy calculations were carried out using topologies generated with standard OPLS-AA atom types and if necessary, the new parameters from Table 1. Calculations were performed in the *NVT* and *NPT* ensemble, with the values shown in Supplementary Table S1. A single compound from the SAMPL5 data set (**080**) has an experimental value of hydration free energy reported in literature (-12.82 ± 0.15 kcal/mol) [5] and a comparison with the computed values shows an error in the prediction of 0.71 kcal/mol and 0.55 kcal/mol in *NVT* and *NPT*, respectively.

From the solvation free energies, $\log D_{cw}$ were computed according to Eqs. 4 and 5. The distribution coefficients are tabulated in Table 3. The accuracy of the computed distribution coefficients was quantified by computing the root mean square error (RMSE), the mean absolute unsigned error (AUE), and the mean error (ME) from the experimental values.

Table 3: Computed ($\log D_{cw}$) and experimental ($\log D_{cw}^{\mathrm{exp}}$) water-cyclohexane distribution coefficients with error estimate for all SAMPL5 compounds. The difference Δ (Eq. 7a) between experimental and computed water-cyclohexane distribution coefficients is shown for each compound. The standard error of the mean in the last significant digits is given in parentheses (Eq. 7b). The root mean square error (RMSE), the absolute unsigned error (AUE), and the signed mean error (ME) were calculated according to Eqs. 8–10.

id	Exp.	Protocol *NVT*[a]		Protocol *NPT*		Ligandbook	Parameterization
	$\log D_{cw}^{\mathrm{exp}}$	$\log D_{cw}$	Δ	$\log D_{cw}$	Δ	ID[b]	
002	1.40(30)	3.14(11)	−1.74(32)	2.77(10)	−1.37(32)	2825	thiophene with CM5
003	1.90(10)	5.71(09)	−3.81(13)	5.72(9)	−3.82(13)	2826	
004	2.20(30)	4.19(11)	−1.99(32)	5.02(11)	−2.82(32)	2827	
005	−0.86(9)	−1.41(13)	0.55(16)	−1.03(12)	0.17(15)	2828	
006	−1.02(9)	0.80(11)	−1.82(14)	2.00(11)	−3.02(14)	2829	
007	1.40(30)	3.85(11)	−2.45(32)	4.92(13)	−3.52(33)	2830	
010	−1.70(40)	−0.18(10)	−1.52(41)	0.46(11)	−2.16(41)	2831	
011	−2.96(8)	0.48(10)	−3.44(13)	0.52(11)	−3.48(14)	2832	
013	−1.50(40)	3.84(13)	−5.34(42)	4.32(12)	−5.82(42)	2833	
015	−2.20(30)	−0.17(12)	−2.03(32)	0.26(11)	−2.46(32)	2834	
017	2.50(30)	5.89(11)	−3.39(32)	7.11(13)	−4.61(33)	2835	
019	1.20(40)	2.09(11)	−0.89(41)	2.83(11)	−1.63(41)	2836	
020	1.60(30)	1.22(10)	0.38(32)	2.02(10)	−0.42(32)	2837	
021	1.20(30)	−0.35(09)	1.55(31)	−0.38(10)	1.58(32)	2838	1,3,4-thiadiazole with CM5
024	1.00(40)	5.05(13)	−4.05(42)	5.23(13)	−4.23(42)	2839	thiophene with CM5
026	−2.60(10)	0.26(10)	−2.86(14)	1.10(11)	−3.70(15)	2840	
027	−1.87(7)	−0.43(09)	−1.44(11)	−0.40(9)	−1.47(11)	2841	
033	1.80(20)	5.90(12)	−4.10(23)	6.39(13)	−4.59(24)	2842	thiophene with CM5
037	−1.50(10)	−7.58(11)	6.08(15)	−7.75(9)	6.25(13)	2843	sulfamide with adapted sulfonamide
042	−1.10(30)	4.38(12)	−5.48(32)	4.12(12)	−5.22(32)	2845	aromatic tautomer form of 2*H*-pyridazin-3-one, with pyridazine parameters
044	1.00(40)	5.75(11)	−4.75(41)	6.33(11)	−5.33(41)	2847	
045	−2.10(20)	0.00(9)	−2.10(22)	0.80(9)	−2.90(22)	2848	
046	0.20(30)	0.52(18)	−0.32(35)	0.88(23)	−0.68(38)	2849	thiophene with CM5
047	−0.40(30)	−0.16(14)	−0.24(33)	0.43(11)	−0.83(32)	2850	thiophene with CM5
048	0.90(40)	3.55(11)	−2.65(41)	4.25(11)	−3.35(41)	2851	
049	1.30(10)	2.55(10)	−1.25(14)	2.97(9)	−1.67(13)	2852	
050	−3.20(60)	2.57(9)	−5.77(61)	2.07(9)	−5.27(61)	2853	
055	−1.50(10)	1.42(8)	−2.92(13)	1.47(7)	−2.97(12)	2854	
056	−2.50(10)	0.90(10)	−3.40(14)	2.11(10)	−4.61(14)	2855	
058	0.80(10)	4.66(9)	−3.86(13)	4.94(9)	−4.14(13)	2856	
059	−1.30(30)	1.47(8)	−2.77(31)	1.92(8)	−3.22(31)	2857	1,2,4-thiadiazole with CM5
060	−3.90(20)	1.64(9)	−5.54(22)	1.91(9)	−5.81(22)	2858	
061	−1.45(9)	−0.25(22)	−1.20(24)	0.39(16)	−1.84(18)	2859	
063	−3.00(40)	−5.04(10)	2.04(41)	−4.26(10)	1.26(41)	2860	
065	0.70(20)	8.17(18)	−7.47(27)	8.45(19)	−7.75(28)	2861	
067	−1.30(30)	3.27(13)	−4.57(33)	4.55(13)	−5.85(33)	2862	
068	1.40(30)	6.51(11)	−5.11(32)	7.48(12)	−6.08(32)	2863	1,2,4-triazine with CM5
069	−1.30(30)	2.12(12)	−3.42(32)	2.55(13)	−3.85(33)	2864	
070	1.60(30)	5.27(11)	−3.67(32)	5.46(11)	−3.86(32)	2865	
071	−0.10(50)	1.99(10)	−2.09(51)	2.26(11)	−2.36(51)	2866	
072	0.60(30)	4.44(12)	−3.84(32)	4.73(11)	−4.13(32)	2867	
074	−1.90(30)	−4.21(11)	2.31(32)	−4.48(11)	2.58(32)	2868	
075	−2.80(30)	2.65(14)	−5.45(33)	3.72(14)	−6.52(33)	2869	
080	−2.20(20)	−1.43(8)	−0.77(22)	−1.18(8)	−1.02(22)	2870	adapted purine and amide parameters
081	−2.20(30)	−2.50(11)	0.30(32)	−2.52(14)	0.32(33)	2871	
082	2.50(40)	8.52(14)	−6.02(42)	8.99(17)	−6.49(43)	2872	
083	−1.90(40)	−4.69(68)	2.79(79)	0.23(90)	−2.13(98)	2873	(1.5 ns equilibration MD)
084	0.00(20)	3.10(15)	−3.10(25)	3.92(21)	−3.92(29)	2874	
085	−2.20(40)	0.33(10)	−2.53(41)	1.13(10)	−3.33(41)	2875	hydantoin with adapted urea and amide parameters
086	0.70(20)	2.89(19)	−2.19(28)	3.83(16)	−3.13(26)	2877	
088	−1.90(30)	0.15(12)	−2.05(32)	1.00(12)	−2.90(32)	2878	
090	0.80(20)	8.04(12)	−7.24(23)	7.95(12)	−7.15(23)	2879	
092	−0.40(30)	3.80(17	−4.20(34)	4.72(19)	−5.12(36)	2880	
RMS Error (RMSE)			3.56(5)		3.95(5)		
Absolute Unsigned Error (AUE)			3.07(5)		3.49(5)		
Mean Error (ME)			−2.47(5)		−3.03(5)		

[a] These results represent our first submission to SAMPL5 (*#68*). [b] Parameters for Gromacs (ITP and PDB files) were deposited in the Ligandbook repository `https://ligandbook.org` under this accession number.

The RMSE was 3.95 ± 0.05 for the *NPT* calculation; the RMSE for the *NVT* calculations was marginally better with 3.56 ± 0.05. The *NVT* results had been submitted to the SAMPL5 challenge as entries *#68* (see Table 3) and *#32* (Table 4) and were analyzed by the SAMPL5 organizers in the context of all other submissions [47]. Nevertheless, in the following we primarily discuss the results from the *NPT* calculations because these values should in principle better represent the experimental measurements. Furthermore, *NPT* calculations also yielded good statistical reproducibility of the van der Waals component of the FEP calculations unlike *NVT* calculations, which depended sensitively on the initial simulation system size [12]. The similar RMSE between the *NVT* and *NPT* simulations suggests that in a large data set, the *NVT* van der Waals error averages out and the overall precision in prediction is similar, as indicated by a high degree of correlation between the *NVT* and *NPT* distribution coefficients as measured by the Pearson linear correlation coefficient $r = 0.97$ (Figure S2 and discussion in the Electronic supplementary material).

The correlation between experimental and computed values in Figure 4 also showed a wide spread of values. Although a few compounds like **005**, **020**, **046**, **047**, and **081** were within one log unit, many others were off by three or more units, with a few as far as more than seven units (such as **065** and **090**). The Pearson correlation coefficient r for both the *NPT* and the *NVT* calculations was 0.64 (with $r = 1$ indicating perfect correlation, 0 no correlation, and -1 perfect anticorrelation), summarizing the moderate success in quantitatively predicting $\log D_{cw}$. To quantify the ability to rank-order the data we computed the Kendall rank correlation coefficient τ; a value of $\tau = 1$ indicates that the simulations predict the same ranking of compounds by $\log D_{cw}$ as the experimental data whereas if the rankings were completely reversed τ would obtain the value -1 and if the simulations produced random results, a value close to 0 would be expected. The *NPT* data yielded $\tau = 0.49$, slightly better than the value of 0.47 for the *NVT* predictions. The simulations are moderately successful at rank-ordering compounds, with the *NPT* protocol being slightly better despite a worse RMSE.

For a number of compounds (**037**, **042**, **085**) we also explored alternative parameterizations but without any clear improvements (Table 4). Compound **083** is a large and complicated macrocyle that is likely able to undergo slow conformational changes. Initially, we had only been able to sample for one tenth of the simulation time and obtained an error of -2.13 ± 0.98 (1.5 ns instead of 15 ns, see Table 3). However, neither more extensive sampling for 15 ns improved the prediction (error -3.73 ± 0.49) nor alternative starting conformations with 15 ns sampling (error -2.69 ± 0.48; see Table 4).

The overall quality of the prediction is worse than one would have expected from the accuracy that is considered achievable for solvation free energy calculations (1–2 kcal/mol), namely 1–2 log units at $T = 300$ K (estimated from Eq. 6). We did not detect an obvious pattern in the chemical character of the compounds that were predicted well versus the ones that were predicted poorly. However, visual inspection of the correlation plot (Figure 4) indicated that most predictions were too positive compared to the experimental values, which was also borne out by the (signed) mean error (ME) of -3.03 ± 0.05 (calculated as experimental value minus computed value,

Table 4: Computed ($\log D_{cw}$) and experimental ($\log D_{cw}^{\mathrm{exp}}$) water-cyclohexane distribution coefficients for selected SAMPL5 compounds with modified simulation parameters. The standard error of the mean in the last significant digits is given in parentheses. See Table 3 for further details.

id	Exp.	Protocol *NVT*		Protocol *NPT*		Ligandbook	Parameterization
	$\log D_{cw}^{\mathrm{exp}}$	$\log D_{cw}$	Δ	$\log D_{cw}$	Δ	ID[a]	
037	−1.50(10)			−4.80(8)	3.30(13)	2844	sulfamide with CM5
042	−1.10(30)	2.27(11)[b]	−3.37(32)[b]	1.75(11)	−2.85(32)	2846	2*H*-pyridazin-3-one with CM5
083	−1.90(40)			1.83(28)	−3.73(49)	2873	(15 ns equilibration MD)
083	−1.90(40)			0.79(27)	−2.69(48)	2873	(15 ns equilibration MD, alternative initial conformation)
085	−2.20(40)	1.52(10)[b]	−3.72(41)[b]	2.69(10)	−4.89(41)	2876	hydantoin with adapted uracil parameters

[a] Parameters for Gromacs (ITP and PDB files) were deposited in the Ligandbook repository `https://ligandbook.org` under this accession number. [b] These results were substituted for the computed values in Table 3 and the new data set comprised our second submission to SAMPL5 (*#32*).

see Table 3 for details). Overall, these results suggested the presence of a systematic error.

If we were to assume that our results could be corrected by systematically shifting the calculated values by the ME then the shifted *NPT* data would have an RMSE of 2.55 instead of 3.95; the shifted *NVT* data would have a similar RMSE of 2.57 instead of 3.56. Even if such an ad-hoc correction were to be considered, the resulting accuracy would remain modest.

It is therefore important to understand the source of the systematic error, with the hope to improve both the systematic shift and the low accuracy. Our previous SAMPL4 hydration free energy results [12] showed a systematically too positive ΔG_w. We therefore hypothesize that primarily the hydration free energy calculations contribute to the systematic error in $\log D_{cw}$. However, the experimental distribution coefficient data do not contain sufficient information to distinguish our hypothesis from the other possibilities of either only ΔG_c being in error or both ΔG_w and ΔG_c contributing similarly. In addition to $\log D_{cw}$, either the experimental hydration free energies or the cyclohexane free energies would be required to directly test our hypothesis. Only for compound **080** (caffeine) hydration free energy data were available [5] and in this case, our prediction of $\log D_{cw}$ was already fairly good with an error of -1.02 ± 0.22. To test our hypothesis, we began to compile an alternative test data set of 92 compounds with known ΔG_w and ΔG_c and calculated the solvation free energies (manuscript in preparation). Preliminary results indicated that for this data set, the cyclohexane solvation free energy can be accurately computed with an RMSE less than 0.8 kcal/mol. The hydration free energy ΔG_w, however, was more difficult to compute (RMSE 1.5 kcal/mol) and was systematically overestimated, in agreement with our previous study [12]. Among the different water models that were evaluated in our preliminary study, TIP3P and SPC provided slightly better predictions than TIP4P, in agreement with the results of a previous report that calculated hydration free energies of amino acid analogues in the OPLS-AA force field with different water models [48]. It follows from Eq. 1 that a more positive ΔG_w leads to a more positive $\log D_{cw}$ and thus our hypothesis is consistent with the results shown here. Taken

together, these results already suggest that the hydration free energy calculations are currently the major source of error in our distribution coefficient calculations.

4 Conclusions

We used explicit solvent all-atom MD simulations with the OPLS-AA force field and the TIP4P water model to predict water-cyclohexane distribution coefficients for the 53 compounds included in the SAMPL5 challenge. We validated cyclohexane parameters for the necessary cyclohexane solvation free energy calculations and introduced a number of new OPLS-AA atom types that were necessary to cover the chemical functionalities in the SAMPL5 compounds. The overall quality of our prediction was worse than expected from what should be theoretically possible with current state-of-the-art absolute solvation free energy calculations. Across the data set, there was no statistical difference between calculations in the NVT and the NPT ensemble. Changes in parameterizations that were tested for a subset of compounds also did not make a difference and the errors did not seem to correspond to any specific chemical functional groups. An overall systematic error was observed whereby predicted distribution coefficients were too positive. Experimental distribution coefficients on their own were not sufficient to determine the source of the error. Based on calculations for an alternative test set of compounds (manuscript in preparation) we hypothesize that the hydration free energy calculations are the main source of the error and future work will focus on addressing this shortcoming in our OPLS-AA parameterization approach, including a critical assessment of the role of the water model itself.

References

1. Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 6(11):881–90, DOI 10.1038/nrd2445
2. Bannan CC, Calabro G, Kyu DY, Mobley DL (2016) Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. J Chem Theory Comput DOI 10.1021/acs.jctc.6b00449
3. Nicholls A, Mobley DL, Guthrie JP, Chodera JD, Bayly CI, Cooper MD, Pande VS (2008) Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. J Med Chem 51(4):769–779, DOI 10.1021/jm070549+
4. Guthrie JP (2009) A blind challenge for computational solvation free energies: Introduction and overview. J Phys Chem B 113(14):4501–4507, DOI 10.1021/jp806724u
5. Geballe MT, Skillman AG, Nicholls A, Guthrie JP, Taylor PJ (2010) The SAMPL2 blind prediction challenge: Introduction and overview. J Comput Aided Mol Des 24(4):259–279, DOI 10.1007/s10822-010-9350-8
6. Geballe MT, Guthrie JP (2012) The SAMPL3 blind prediction challenge: transfer energy overview. J Comput Aided Mol Des 26(5):489–96, DOI 10.1007/s10822-012-9568-8

7. Mobley DL, Wymer KL, Lim NM, Guthrie JP (2014) Blind prediction of solvation free energies from the SAMPL4 challenge. J Comput Aided Mol Des 28(3):135–50, DOI 10.1007/s10822-014-9718-2
8. Rustenburg AS, Dancer J, Lin B, Ortwine DF, Mobley DL, Chodera JD (2016) Measuring experimental cyclohexane/water distribution coefficients for the SAMPL5 challenge. J Comput Aided Mol Des
9. Lin B, Pease JH (2013) A novel method for high throughput lipophilicity determination by microscale shake flask and liquid chromatography tandem mass spectrometry. Comb Chem High Throughput Screen 16(10):817–25, DOI 10.1021/ct200866d
10. Jorgensen WL, Tirado-Rives J (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc Natl Acad Sci USA 102(19):6665–6670, DOI 10.1073/pnas.0408037102
11. Beckstein O, Iorga BI (2012) Prediction of hydration free energies for aliphatic and aromatic chloro derivatives using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 26(5):635–645, DOI 10.1007/s10822-011-9527-9
12. Beckstein O, Fourrier A, Iorga BI (2014) Prediction of hydration free energies for the SAMPL4 diverse set of compounds using molecular dynamics simulations with the OPLS-AA force field. J Comput Aided Mol Des 28(3):265–276, DOI 10.1007/s10822-014-9727-1
13. Kaminski G, Duffy E, Matsui T, Jorgensen W (1994) Free energies of hydration and pure liquid properties of hydrocarbons from the OPLS all-atom model. J Phys Chem 98(49):13,077–13,082, DOI 10.1021/j100100a043
14. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118(45):11,225–11,236, DOI 10.1021/ja9621760
15. Damm W, Frontera A, Tirado-Rives J, Jorgensen W (1997) OPLS all-atom force field for carbohydrates. J Comput Chem 18(16):1955–1970, DOI 10.1002/(SICI)1096-987X(199712)18:16<1955::AID-JCC1>3.0.CO;2-L
16. Jorgensen WL, McDonald NA (1998) Development of an all-atom force field for heterocycles. Properties of liquid pyridine and diazenes. J Mol Struct THEOCHEM 424(1-2):145–155, DOI 10.1016/S0166-1280(97)00237-6
17. McDonald NA, Jorgensen WL (1998) Development of an all-atom force field for heterocycles. Properties of liquid pyrrole, furan, diazoles, and oxazoles. J Phys Chem B 102(41):8049–8059, DOI 10.1021/jp981200o
18. Rizzo RC, Jorgensen WL (1999) OPLS all-atom model for amines: Resolution of the amine hydration problem. J Am Chem Soc 121(20):4827–4836, DOI 10.1021/ja984106u
19. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105(28):6474–6487, DOI 10.1021/jp003919d
20. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput 4(3):435–447, DOI 10.1021/ct700301q

21. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, Hess B, Lindahl E (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics 29(7):845–54, DOI 10.1093/bioinformatics/btt055
22. Watkins EK, Jorgensen WL (2001) Perfluoroalkanes: Conformational analysis and liquid-state properties from ab initio and Monte Carlo calculations. J Phys Chem A 105(16):4118–4125, DOI 10.1021/jp004071w
23. Price M, Ostrovsky D, Jorgensen W (2001) Gas-phase and liquid-state properties of esters, nitriles, and nitro compounds with the OPLS-AA force field. J Comput Chem 22(13):1340–1352, DOI 10.1002/jcc.1092
24. Kony D, Damm W, Stoll S, Van Gunsteren W (2002) An improved OPLS-AA force field for carbohydrates. J Comput Chem 23(15):1416–1429, DOI 10.1002/jcc.10139
25. Kahn K, Bruice T (2002) Parameterization of OPLS-AA force field for the conformational analysis of macrocyclic polyketides. J Comput Chem 23(10):977–996, DOI 10.1002/jcc.10051
26. Thomas L, Christakis T, Jorgensen W (2006) Conformation of alkanes in the gas phase and pure liquids. J Phys Chem B 110(42):21,198–21,204, DOI 10.1021/jp064811m
27. Jorgensen W, Jensen K, Alexandrova A (2007) Polarization effects for hydrogen-bonded complexes of substituted phenols with water and chloride ion. J Chem Theory Comput 3(6):1987–1992, DOI 10.1021/ct7001754
28. Xu Z, Luo HH, Tieleman DP (2007) Modifying the OPLS-AA force field to improve hydration free energies for several amino acid side chains using new atomic charges and an off-plane charge model for aromatic residues. J Comput Chem 28(3):689–697, DOI 10.1002/jcc.20560
29. Domański J, Beckstein O, Iorga BI (2012) Ligandbook – an online repository for small and drug-like molecule force field parameters. In: 244th Meeting of the American Chemical Society – Abstracts of Papers, pp 227–COMP, URL `https://ligandbook.org`
30. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79(2):926–935, DOI 10.1063/1.445869
31. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA Jr, Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT

32. Vilseck JZ, Tirado-Rives J, Jorgensen WL (2014) Evaluation of CM5 charges for condensed-phase modeling. J Chem Theory Comput 10(7):2802–2812, DOI 10.1021/ct500016d
33. Dodda LS, Vilseck JZ, Cutrona KJ, Jorgensen WL (2015) Evaluation of CM5 charges for nonaqueous condensed-phase modeling. J Chem Theory Comput 11(9):4273–82, DOI 10.1021/acs.jctc.5b00414
34. Páll S, Abraham MJ, Kutzner C, Hess B, Lindahl E (2015) Tackling exascale software challenges in molecular dynamics simulations with GROMACS. In: Markidis S, Laure E (eds) Solving Software Challenges for Exascale: International Conference on Exascale Applications and Software, EASC 2014, Stockholm, Sweden, April 2-3, 2014, Revised Selected Papers, Lecture Notes in Computer Science, vol 8759, Springer International Publishing, Switzerland, pp 3–27, DOI 10.1007/978-3-319-15976-8_1
35. Mobley DL, Dumont E, Chodera JD, Dill KA (2007) Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. J Phys Chem B 111(9):2242–2254, DOI 10.1021/jp0667442
36. Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52(12):7182–7190, DOI 10.1063/1.328693, URL `http://link.aip.org/link/?JAP/52/7182/1`
37. Shirts MR, Pitera JW, Swope WC, Pande VS (2003) Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. J Chem Phys 119(11):5740–5761, DOI 10.1063/1.1587119
38. Essman U, Perela L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8592, DOI 10.1063/1.470117
39. Hess B (2008) P-LINCS: A parallel linear constraint solver for molecular simulation. J Chem Theory Comput 4(1):116–122, DOI 10.1021/ct700200b
40. Jorge M, Garrido N, Queimada A, Economou I, Macedo E (2010) Effect of the integration method on the accuracy and computational efficiency of free energy calculations using thermodynamic integration. J Chem Theory Comput 6(4):1018–1027, DOI 10.1021/ct900661c
41. Jones E, Oliphant T, Peterson P, et al (2001–) SciPy: Open source scientific tools for Python. URL `http://www.scipy.org/`, [Online; accessed 2016-06-11]
42. Faber NKM (1999) Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration. Chemometrics and Intelligent Laboratory Systems 49(1):79 – 89, DOI 10.1016/S0169-7439(99)00027-1
43. O'Neil MJ (ed) (2013) The Merck Index — An Encyclopedia of Chemicals, Drugs, and Biologicals. Royal Society of Chemistry, Cambridge, UK
44. Marenich AV, Kelly CP, Thompson JD, Hawkins GD, Chambers CC, Giesen DJ, Winget P, Cramer CJ, Truhlar DG (2009) Minnesota Solvation Database - version 2009, University of Minnesota, Minneapolis. URL `http://comp.chem.umn.edu/mnsol/`
45. Marenich AV, Jerome SV, Cramer CJ, Truhlar DG (2012) Charge Model 5: An extension of Hirshfeld population analysis for the accurate description of molec-

ular interactions in gaseous and condensed phases. J Chem Theory Comput 8(2):527–41, DOI 10.1021/ct200866d
46. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA (2009) Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. J Chem Theory Comput 5(2):350–358, DOI 10.1021/ct800409d
47. Bannan CC, Burley KH, Mobley DL (2016) Blind prediction of cyclohexane-water distribution coefficients from the SAMPL5 challenge. J Comput Aided Mol Des
48. Shirts MR, Pande VS (2005) Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. J Chem Phys 122(13):134,508, DOI 10.1063/1.1877132

Journal of Computer-Aided Molecular Design

Electronic supplementary material

Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field

Ian M. Kenney · Oliver Beckstein · Bogdan I. Iorga

1 Solvation free energies

Solvation free energies ΔG_{solv} in water and cyclohexane for all SAMPL5 components reported in this study are listed in Table S1. Hydration free energies for model compounds used for parameterization are reported in Table S2. Parameters for Gromacs (ITP and PDB files) were deposited in the Ligandbook repository `https://ligandbook.org` under the accession number give in the tables.

OB was supported in part by grant ACI-1443054 from the National Science Foundation. BII was supported in part by grants ANR-10-LABX-33 (LabEx LERMIT) and ANR-14-JAMR-0002-03 (JPIAMR) from the French National Research Agency (ANR).

I. M. Kenney
Department of Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA

O. Beckstein
Department of Physics and Center for Biological Physics, Arizona State University, P.O. Box 871504, Tempe, AZ 85287-1504, USA
Tel.: +1 480 727 9765
Fax: +1 480 965-4669
E-mail: oliver.beckstein@asu.edu

B. I. Iorga
Institut de Chimie des Substances Naturelles, CNRS UPR 2301, Université Paris-Saclay, Labex LERMIT, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette, France
Tel.: +33 1 69 82 30 94
Fax: +33 1 69 07 72 47
E-mail: bogdan.iorga@cnrs.fr

Table S1: Solvation free energies ΔG_{solv} (kcal·mol^{-1}) from all SAMPL5 simulations reported in this study. The standard error of the mean in the last significant digits is given in parentheses. Parameters for Gromacs (ITP and PDB files) were deposited in the Ligandbook repository `https://ligandbook.org` under the ID accession number.

	water		cyclohexane			
id	*NVT*	*NPT*	*NVT*	*NPT*	ID[a]	Parameterization
002	−10.23(11)	−10.16(10)	−14.54(10)	−13.96(10)	2825	
003	−6.03(9)	−5.70(9)	−13.86(9)	−13.54(9)	2826	
004	−9.65(10)	−8.53(11)	−15.40(12)	−15.41(11)	2827	
005	−18.69(15)	−18.15(13)	−16.76(10)	−16.74(11)	2828	
006	−10.15(10)	−8.92(11)	−11.24(11)	−11.67(11)	2829	
007	−11.00(10)	−9.30(13)	−16.28(11)	−16.06(12)	2830	
010	−13.44(10)	−12.72(10)	−13.19(9)	−13.35(12)	2831	
011	−13.73(10)	−13.49(10)	−14.39(9)	−14.21(11)	2832	
013	−15.67(12)	−14.99(11)	−20.93(13)	−20.92(12)	2833	
015	−13.11(13)	−12.53(13)	−12.88(9)	−12.89(8)	2834	
017	−8.78(11)	−6.92(13)	−16.87(12)	−16.67(12)	2835	
019	−13.50(11)	−12.51(11)	−16.36(11)	−16.39(12)	2836	
020	−12.64(10)	−11.73(10)	−14.31(10)	−14.50(10)	2837	
021	−14.67(9)	−14.41(9)	−14.19(9)	−13.88(10)	2838	1,3,4-thiadiazole with CM5
024	−13.05(12)	−12.74(11)	−19.99(13)	−19.91(14)	2839	thiophene with CM5
026	−11.33(10)	−10.15(12)	−11.69(9)	−11.66(9)	2840	
027	−13.30(9)	−13.04(8)	−12.70(8)	−12.49(8)	2841	
033	−9.96(11)	−8.86(12)	−18.05(12)	−17.63(14)	2842	thiophene with CM5
037	−19.89(11)	−20.11(9)	−9.49(9)	−9.48(8)	2843	sulfamide with adapted sulfonamide
042	−11.40(12)	−11.21(11)	−17.42(11)	−16.86(11)	2845	aromatic tautomer form of 2*H*-pyridazin-3-one, with pyridazine parameters
044	−11.68(11)	−10.88(11)	−19.57(11)	−19.56(11)	2847	
045	−11.88(9)	−10.74(9)	−11.88(8)	−11.83(9)	2848	
046	−17.00(23)	−16.03(30)	−17.72(11)	−17.24(12)	2849	thiophene with CM5
047	−14.99(16)	−14.04(11)	−14.78(10)	−14.63(11)	2850	thiophene with CM5
048	−13.77(11)	−12.98(11)	−18.65(11)	−18.82(11)	2851	
049	−10.19(9)	−9.35(9)	−13.69(10)	−13.43(10)	2852	
050	−9.69(9)	−10.16(9)	−13.22(9)	−13.00(8)	2853	
055	−9.27(8)	−8.85(7)	−11.21(8)	−10.87(7)	2854	
056	−10.44(10)	−8.85(10)	−11.68(10)	−11.74(11)	2855	
058	−7.93(8)	−6.95(8)	−14.32(9)	−13.72(10)	2856	
059	−8.04(8)	−7.23(7)	−10.06(8)	−9.86(7)	2857	1,2,4-thiadiazole with CM5
060	−10.82(8)	−10.20(9)	−13.07(8)	−12.81(9)	2858	
061	−10.41(29)	−9.36(20)	−10.07(8)	−9.89(9)	2859	

Table S1 – continued

	water		cyclohexane			
id	*NVT*	*NPT*	*NVT*	*NPT*	ID[a]	Parameterization
063	−18.09(11)	−16.78(10)	−11.17(9)	−10.94(10)	2860	
065	−18.43(16)	−17.32(18)	−29.64(19)	−28.92(18)	2861	
067	−8.87(15)	−6.60(15)	−13.36(10)	−12.85(11)	2862	
068	−9.20(11)	−7.82(11)	−18.13(11)	−18.08(12)	2863	1,2,4-triazine with CM5
069	−13.11(12)	−12.33(13)	−16.01(11)	−15.83(11)	2864	
070	−6.89(11)	−6.15(11)	−14.13(11)	−13.64(11)	2865	
071	−10.07(10)	−9.15(10)	−12.80(9)	−12.26(11)	2866	
072	−5.72(13)	−5.38(11)	−11.81(11)	−11.87(10)	2867	
074	−21.26(11)	−21.25(11)	−15.49(9)	−15.10(10)	2868	
075	−9.00(17)	−7.74(16)	−12.65(10)	−12.84(11)	2869	
080	−13.53(8)	−13.37(8)	−11.57(8)	−11.75(7)	2870	adapted purine and amide parameters
081	−16.71(11)	−17.23(15)	−13.28(10)	−13.77(11)	2871	
082	−5.66(13)	−4.38(19)	−17.35(14)	−16.71(13)	2872	
083	−41.18(72)	−34.96(88)	−34.75(60)	−35.28(87)	2873	(1.5 ns equilibration MD)
084	−13.29(17)	−12.33(25)	−17.54(13)	−17.71(13)	2874	
085	−14.25(9)	−13.24(10)	−14.70(9)	−14.79(10)	2875	hydantoin with adapted urea and amide parameters
086	−14.52(23)	−12.76(18)	−18.49(13)	−18.02(12)	2877	
088	−13.55(11)	−11.99(12)	−13.75(12)	−13.36(12)	2878	
090	−8.01(11)	−7.07(12)	−19.05(12)	−17.98(11)	2879	
092	−21.52(16)	−19.84(20)	−26.73(16)	−26.32(16)	2880	
037		−16.16(8)		−9.57(8)	2844	sulfamide with CM5
042	−15.64(11)	−16.05(12)	−18.75(11)	−18.45(10)	2846	2*H*-pyridazin-3-one with CM5
083		−31.51(25)		−34.02(28)	2873	(15 ns equilibrium MD)
083		−31.78(29)		−32.86(22)	2873	(15 ns equilibrium MD, alternative initial conformation)
085	−12.41(10)	−11.08(10)	−14.49(11)	−14.77(10)	2876	hydantoin with adapted uracil parameters

[a] Parameters for Gromacs (ITP and PDB files) were deposited in the Ligandbook repository `https://ligandbook.org` under this accession number.

Table S2 Hydration free energies ΔG_{hyd} (kcal·mol^{-1}) for compounds used for the validation of thiophene parameterization. In all cases, parameterization of thiophene rings used CM5 charges. The standard error of the mean in the last significant digits is given in parentheses.

id	*NVT*	*NPT*	ID[a]
thiophene	−0.24(4)	0.14(5)	2881
2-Me-thiophene	−0.38(5)	0.01(5)	2882

[a] Parameters for Gromacs (ITP and PDB files) were deposited in the Ligandbook repository `https://ligandbook.org` under this accession number.

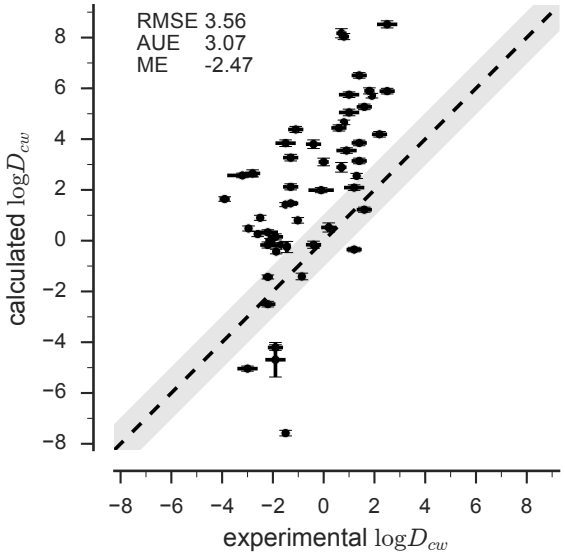


Fig. S1 Correlation between experimental and computed water-cyclohexane distribution coefficients $\log D_{cw}$ for simulations performed in the *NVT* ensemble. The gray band indicates ± 1 log-units from ideal correlation, shown by the dashed line. The root mean square error (RMSE), the absolute unsigned error (AUE), and the (signed) mean error (ME) are indicated. Error bars represent the error in the experiments or the error on the mean, derived from the simulations.

2 $\log D_{cw}$ correlations

Figure S1 shows the correlation of the $\log D_{cw}$ computed in the *NVT* ensemble to the experimental ones. The Pearson correlation coefficient is $r = 0.64$ and the Kendall rank ordering coefficient is $\tau = 0.47$.

In general, the distribution coefficients calculated in the *NVT* and *NPT* ensemble are highly correlated (Figure S2) with a Pearson correlation coefficient $r = 0.97$. The two data sets also rank order almost all compounds in the same way as indicated by

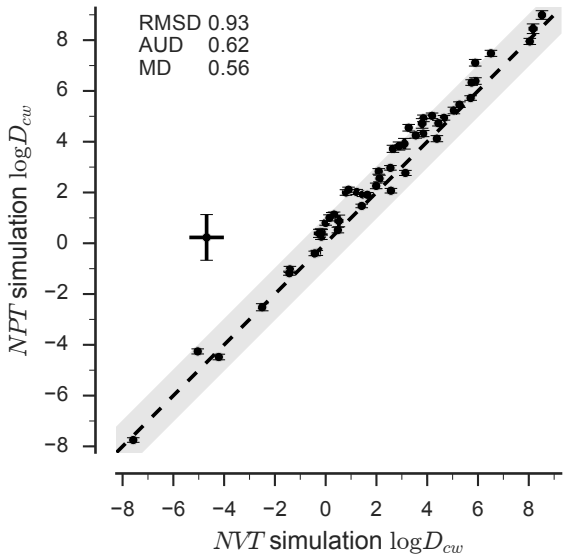


Fig. S2 Correlation between water-cyclohexane distribution coefficients $\log D_{cw}$ for simulations performed in the *NVT* ensemble vs ones computed from the *NPT* ensemble. The obvious outlier is **083**. The gray band indicates ± 1 log-units from ideal correlation, shown by the dashed line. The root mean square deviation (RMSD), the absolute unsigned deviation (AUD), and the (signed) mean deviation (MD) are indicated. Error bars represent the error on the mean, derived from the simulations.

a high Kendall's $\tau = 0.92$. Only compound **083** is very different between the two ensembles but this is almost certainly due to insufficient sampling: The simulations were only 1.5 ns instead of 15 ns and the compound is much larger and more complicated than the other compounds. With additional calculations of 15 ns simulations (*NPT* only) and using different initial conformations, the solvation free energies still differ by about 2 kcal/mol (see Table S1). It is thus likely that the *NVT* value is not converged and the *NPT* value can only be seen as an initial guess.