Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies

Katarzyna B. Koziara · Martin Stroet · Alpeshkumar K. Malde · Alan E. Mark

Received: 15 November 2013 / Accepted: 16 January 2014 / Published online: 30 January 2014

Abstract To test and validate the Automated force field Topology Builder and Repository (ATB; http://compbio.biosci.uq.edu.au/atb/) the hydration free enthalpies for a set of 214 drug-like molecules, including 47 molecules that form part of the SAMPL4 challenge have been estimated using thermodynamic integration and compared to experiment. The calculations were performed using a fully automated protocol that incorporated a dynamic analysis of the convergence and integration error in the selection of intermediate points. The system has been designed and implemented such that hydration free enthalpies can be obtained without manual intervention following the submission of a molecule to the ATB. The overall average unsigned error (AUE) using ATB 2.0 topologies for the complete set of 214 molecules was 6.7 kJ/mol and for molecules within the SAMPL4 7.5 kJ/mol. The root mean square error (RMSE) was 9.5 and 10.0 kJ/mol respectively. However, for molecules containing functional groups that form part of the main GROMOS force field the AUE was 3.4 kJ/mol and the RMSE was 4.0 kJ/mol. This suggests it will be possible to further refine the parameters provided by the ATB based on hydration free enthalpies.

Keywords SAMPL4 · Automated topology builder · GROMOS · Hydration free enthalpy · Molecular dynamics · Thermodynamic integration

Introduction

While well-optimized and validated force field parameters exist for common biomolecules such as amino acids, nucleic acids, lipids and certain sugars, the parameterization of heteromolecular ligands such as substrates, inhibitors, co-factors and potential drug molecules remains a major challenge and a limiting factor in computational drug design [1–3]. Historically, the generation of parameters for novel molecules compatible with a given biomolecular force field has involved searching for similar chemical groups and assigning parameters based on analogy. This approach is impracticable when dealing with large sets of molecules and a number of automated parameterization protocols have been proposed. Programs and web-servers that can be used to obtain parameters suitable for both atomistic simulations and computational drug design include: Antechamber which can provide parameters compatible with the Generalized Amber Force Feld (GAFF) [4, 5]; the YASARA AutoSMILES server [6] which also generates GAFF compatible topologies; SwissParam [7] which provides topologies and parameters for small organic molecules for use with the simulation packages CHARMM [8, 9] and GROMACS [10, 11] based on the Merck Molecular Force Field (MMFF) [12–14]; and the web server ParamChem which uses the CHARMM General Force Field (CGenFF) program [15–17] to assign

Katarzyna B. Koziara and Martin Stroet have contributed equally to the work.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9713-7) contains supplementary material, which is available to authorized users.

K. B. Koziara · M. Stroet · A. K. Malde · A. E. Mark (✉)
School of Chemistry and Molecular Biosciences, University of Queensland, St Lucia, QLD 4072, Australia
e-mail: a.e.mark@uq.edu.au

A. E. Mark
Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia

atom types, bonded parameters and atomic charges primarily by analogy. Recently, a web accessible Automated force field Topology Builder (ATB; http://compbio.biosci.uq.edu.au/atb/) and Repository [18] has been developed to provide interaction parameters for a wide range of molecules compatible with the GROMOS force field [19]. The ATB can provide topologies and parameters for use in molecular simulations, computational drug design and X-ray refinement. While each of these automated protocols is widely used, the degree to which the parameters proposed have been tested and validated remains an open question. Automating empirical force field parameterization is challenging for many reasons: (a) one must attempt to describe what is a very complex potential energy surface using a limited number of terms (b) the parameters associated with different types of interactions are usually correlated and (c) some terms commonly used in molecular force fields (such as partial atomic charges) do not directly represent physically observable properties. In addition, the number of parameters required to describe a given molecule is large compared to the range of experimental data against which any given model can be validated, meaning that force field development is an under-determined problem. Different assumptions in the fitting of partial charges, for example, can lead to very different sets of parameters being proposed [18].

The ATB attempts to overcome these difficulties by using a combination of quantum mechanical calculations and a knowledge-based approach when generating parameters and topologies. Given that factors such as the geometry, stereochemistry, protonation and tautomeric state of a given molecule can affect the parameterization, the ATB requires that users fully stipulate the state of a molecule. This information is provided in the form of a coordinate file in Protein Data Bank [20] (PDB) format (which must contain all hydrogen atoms), a connectivity record in PDB format listing all interatomic bonds and the net charge on the molecule. Submitted molecules are initially optimised at the HF/STO-3G (or AM1 [21] or PM3 [22, 23]) level of theory, then re-optimised at the B3LYP/6-31G* level of theory [24–26] in conjunction with the *Polarizable Continuum Model* (PCM) implicit solvent (water) as implemented in GAMESS-US [27]. The Hessian and initial estimates of the partial charges are then calculated based on this geometry. Atom types are assigned on the basis of the local environment as determined by connectivity, as has been described previously [18]. Initial charges are estimated by fitting to the electrostatic potential using Kollmann-Singh [28] scheme as implemented in GAMESS-US. The symmetry is analysed and the partial charges modified to ensure assignment of identical parameters to equivalent atoms. Optimised charge groups are assigned using a graph-based algorithm [29]. The charges are then further refined (within the uncertainty in the charge assignment) to obtain neutral charge groups. The parameters are not further optimised or adjusted to match those within the rest of the GROMOS force field. The ATB provides an indication of the reliability of individual parameters and alternative parameters in those cases where a definitive assignment is not possible. The ATB also provides a number of validation tools. The parameters are passed through the GROMOS topology validation tool *check_top* in which the bond, angle and dihedral energies are analysed. The root mean square positional deviation (RMSD) between the quantum mechanical (QM) optimized structure and the structure obtained after minimizing this structure in vacuum using the parameters generated by the ATB is also provided. The ATB output includes building block files (all atom and united atom), interaction parameter files for the corresponding force field and optimized geometries. The ATB also acts as a repository for molecules that have been parameterized as part of the GROMOS family of force fields and for pre-equilibrated systems that can be used as starting configurations in molecular dynamics simulations (e.g. solvent mixtures, lipid systems). Building block and interaction parameter files are provided in GROMOS, GROMACS, CIF and CNS formats.

One of the main limitations of the ATB is that the van der Waals atom type parameters are limited to those currently present in the GROMOS force field. Atom types in the GROMOS force field have been parameterized primarily for simulations of common biomolecules (peptides, lipids and sugars). Parameters for some atom types, such as halogens, are known to be non-optimal. Another current limitation is that the size of the molecules that can be fully processed by the ATB is limited. High-level QM calculations, including those required to determine the Hessian, are only performed for molecules containing up to 40 atoms (including hydrogens).

The GROMOS family of force fields aims to reproduce the thermodynamic properties of biomolecular and related systems. The estimation of solvation free enthalpies is thus central to the on-going refinement and validation of this family of force fields and, in turn, the ATB. Here we present results for the hydration free enthalpies in SPC [30] water for a total of 214 organic molecules including the 47 molecules that formed part of the SAMPL4 challenge. Topologies were generated automatically using the ATB version 2.0. Hydration free enthalpies were calculated using thermodynamic integration in conjunction with a fully automated protocol designed to return final values within a given uncertainty, based on a dynamic analysis of the statistical and integration error. The effectiveness of this protocol is discussed together with an analysis of the differences between the calculated and experimental values for the 214 test molecules.

Methods

Free enthalpy calculations

Solvation free enthalpies were calculated using the thermodynamic integration (TI) approach [31]. Using this approach the difference in free enthalpy between two states of a system A and B can be expressed as:

$$\Delta G_{AB} = {}_{\lambda_A}^{\lambda_B}\left\langle \frac{\partial V(\boldsymbol{r})}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{1}$$

where $V(\boldsymbol{r})$ is the potential energy of the system as a function of the coordinate vector $\boldsymbol{r}$ and λ is a parameter that couples the two states A and B. In this case the coupling parameter λ was used to scale the inter- and intramolecular non-bonded interactions involving the solute from 0 to 1 (where 0 represents the full interaction and 1 no interaction). To avoid sampling singularities in the potential energy function and in the derivative with respect to λ (as well as numerical instabilities during the simulations) the non-bonded interactions were scaled using the λ-dependent soft-core interaction function of Beutler et al. with $\alpha_{LJ} = 0.5$ and $\alpha_{electrostatic} = 0.5$ nm^2 [32, 33]. Note that when using the λ-dependent soft-core interaction function of Beutler et al. as implemented in GROMOS, there is no requirement or advantage in performing the removal of the charge and LJ interactions in separate stages as is sometimes required by other codes.

Equation 1 was evaluated by calculating the ensemble average of the derivative $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ at a series of discrete λ-values. The values of $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ were then integrated using the Trapezoidal approximation. The Trapezoidal approximation was used simply for consistency with the method used to estimate the integration error. Note that due to the number of intermediate λ-values calculated, the final answer was insensitive to the integration method used with the difference between the values obtained using Simpsons rule or the Trapezoidal approximation being negligible (<0.1 kJ/mol). The solvation free enthalpy was calculated as:

$$\begin{aligned}\Delta G_{solvation} &= \Delta G_{0\to 1}(vacuum) - \Delta G_{0\to 1}(water)\\ &= {}_{\lambda_0}^{\lambda_1}\left\langle \frac{\partial V(\boldsymbol{r})}{\partial \lambda} \right\rangle_{\lambda}^{vac} d\lambda - {}_{\lambda_0}^{\lambda_1}\left\langle \frac{\partial V(\boldsymbol{r})}{\partial \lambda} \right\rangle_{\lambda}^{water} d\lambda\\ &= {}_{\lambda_0}^{\lambda_1}\left[\left\langle \frac{\partial V(\boldsymbol{r})}{\partial \lambda} \right\rangle_{\lambda}^{vac} - \left\langle \frac{\partial V(\boldsymbol{r})}{\partial \lambda} \right\rangle_{\lambda}^{water}\right] d\lambda\end{aligned} \tag{2}$$

Note that for the combined integral in Eq. 2 to hold as written the same λ-values must be sampled in water and in vacuum.

Integration protocol

Initially the value of $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ at 9 equally spaced points between $\lambda = 0$ and $\lambda = 1$ was determined to obtain a first estimate of shape of the underling curve. The locations of potential turning points in this curve were then identified by taking the first derivative of a series of cubic splines fitted to the initial 9 points. A new point was then added at the estimated location of the turning point for which the absolute value of the 2nd derivative was the highest. The value of $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ at this point was then determined and the location of the turning points reassessed. This procedure was repeated until all turning points had been located to within a specified distance (0.05λ).

Once the turning points had been identified additional λ-points were added or simulations extended in those regions of the curve with the largest uncertainty until the estimate of the total error fell below a target threshold. The threshold in this study was 1 kJ/mol. This was achieved by decomposing the curve into a series of overlapping regions each consisting of three consecutive λ-values. The uncertainty in each region was estimated from a) the change in the total integral following the exclusion of the central point, and b) by taking the difference in the integral calculated using either the upper or the lower bound of the uncertainty (calculated using block averaging) in the central point [34]. The total integration error Er_{total} was calculated from the sum of the errors in each region, normalized for any overlap between regions as:

$$Er_{total} = \sqrt{\sum\left(\frac{Er_i}{n}\right)^2}$$

where Er_i is the maximum error for a particular region and n is the normalization factor. If the exclusion of the central λ-value in a given region gave rise to the largest error, a new point was added at the midpoint of the largest subinterval within that region. If the three points that made up a given region were evenly spaced then two new points, one on either side of the midpoint, were added. If the maximum error was due to the uncertainty within a particular point the sampling at that point was extended by 200 ps. The error in the affected regions was then updated and the total integration error recalculated. This procedure was repeated until the total error fell below the target threshold. As an additional test of convergence the systems in water were simulated twice at each λ-value using different initial configurations. Systems in vacuum were simulated using four sets of initial conditions. Where possible the initial configurations for the two water simulations at each λ-point were taken as the final frames of the two neighbouring λ-values. The initial configurations for λ_0 (full interaction with water) and λ_1 (no interaction) were taken from the optimized geometry at the B3LYP/6-31G* level of theory [24–26]. The initial configurations used for the four vacuum simulations were taken from the middle and final frames of each of the two water runs. This was done to ensure a more complete sampling of the available configurational space.

Convergence

Simulations were run for 200 ps at each λ-value or until the ensemble average of the derivative $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ had been deemed to have converged. To determine whether the ensemble average of the derivative $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ had converged the distribution of the variance for successive time periods during the simulation was analyzed. The two-sample Kolmogorov–Smirnov (KS) statistic was used to quantify the degree of similarity between the distributions. $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ was considered to have converged if the KS statistic for adjoining regions was less than 0.05.

Topology generation

The parameters and topology files for all molecules were generated automatically by the ATB version 2.0 and used without modification unless otherwise noted. The parameters were generated based on a single initial conformation. In the case of molecules that formed part of the various SAMPL challenges this conformation was taken directly from the structural coordinates provided. The structures of other molecules were generated using a range of molecule building programmes.

Simulation setup

All calculations were performed using the GROMOS11 simulation package [35] in conjunction with the GROMOS 53A6 force field [19] as implemented in the ATB version 2.0. The starting structures for the simulations in water were taken from the QM optimized geometries generated by the ATB. To generate the water systems each molecule was placed at the centre of a cubic periodic box. The size of the box was chosen such that the minimum distance between the solute and the box wall was 1.4 nm. The solute was then solvated using an equilibrated configuration of SPC [30] water. The system was energy minimized using a steepest decent algorithm. Initial velocities were taken from a Maxwell–Boltzmann distribution at 298 K. Bond lengths were constrained using SHAKE [36] with a geometric tolerance of 10^{-4}. The equations of motion were integrated using a time step of 2 fs. All simulations were performed at constant temperature (298 K) and pressure (1 atm) using a Berendsen thermostat and barostat [37]. The coupling times were 0.1 and 0.5 ps, respectively. The isothermal compressibility was 4.575×10^{-4} ($\mathrm{kJ}^{-1}\mathrm{mol}^{-1}\mathrm{nm}^{-3}$). Non-bonded interactions were calculated using a triple-range scheme. Interactions within a shorter-range cutoff of 0.8 nm were calculated every time step. Interactions between 0.8 and 1.4 nm were updated every 5 steps together with the update to the pairlist. A reaction field was applied to correct for the truncation of electrostatic interactions beyond the long-range cutoff using a relative dielectric permittivity of 61 [38]. The vacuum systems were generated from a given configuration in water by simply deleting all water molecules within the simulation box. In this case pressure coupling was not applied and the temperature was maintained by using stochastic dynamics with a reference temperature of 298 K and an atomic friction coefficient of 91 ps^{-1}.

Results and discussion

The aim of this study was to test and validate topologies and parameters generated by the ATB version 2.0 against experimental hydration free enthalpies. This was achieved using a test set of 214 molecules, of these, 167 had been used to test previous versions of the ATB. The set of 167 reference compounds used previously contained a combination of alcohols, alkanes, cycloalkanes, alkenes, alkynes, alkyl benzenes, amines, amides, aldehydes, carboxylic acids, esters, ketones, thiols and sulphides and included molecules from the earlier SAMPL0/CUP8, SAMPL1 and SAMPL2 challenges [39–41]. A full list of the molecules considered is provided as supplementary material (Supplementary material Table S1). The other 47 molecules formed part of the SAMPL4 challenge and are listed in Table 1. Topologies for all the molecules used in this study are publicly available via the ATB repository. In all cases the topology generation and the calculation of the hydration free enthalpies were fully automated with no manual intervention. No attempt was made to either optimize the ATB parameters based on knowledge of the chemical properties of a particular molecule or to force the system to sample a specific conformational or tautomeric state. This said, during the testing of the SAMPL4 molecules a problem with the algorithm that assigned exclusions was detected. Three molecules 033, 034 and 037 from SAMPL4 that contained an aromatic ring with a hydroxyl group ortho- to a methoxy group were not stable during energy minimization. This was due to the fact that the hydroxyl hydrogen and the oxygen of the methoxy group have high opposing partial charges and are constrained to lie in close proximity. To avoid problems due to the high forces associated with this interaction these atoms were excluded.

Molecular geometry

As an initial validation of the topologies and parameters generated by the ATB, each molecule was energy minimized in vacuum and the resulting structure was compared

Table 1 A comparison between the experimental and calculated free enthalpies of hydration for drug and drug-like molecules from SAMPL4 using parameters assigned by the Automated Topology Builder (ATB)

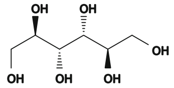
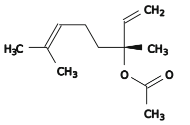
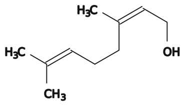
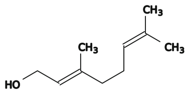
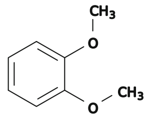
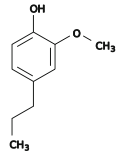
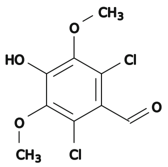
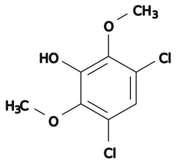
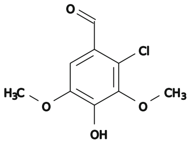
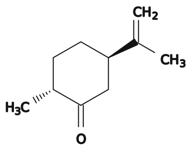
SAMPL ID[a]	FE UA[b]	FE Exp.[c]	Diff[d]	Name	Structure
001	−86.0 ± 1.4	−98.9 ± 1.3	12.9 ± 1.9	Mannitol	
002	−9.5 ± 0.6	−10 ± 4	0.5 ± 4.0	Linalyl acetate	
003	−20.8 ± 0.6	−20.0 ± 1.0	0.8 ± 1.2	Nerol	
004	−16.6 ± 0.5	−18.6 ± 1.0	2.0 ± 1.1	Geraniol	
005	−12.0 ± 0.6	−22.3 ± 0.4	10.3 ± 0.7	1,2-Dimethoxybenzene	
006	−43.4 ± 0.7	−22.0 ± 0.8	21.4 ± 1.1	4-Propylguaiacol	
009	−43.9 ± 0.5	−34 ± 3	9.9 ± 3.0	2,6-Dichlorosyringaldehyde	
010	−30.8 ± 0.5	−26.1 ± 1.6	4.7 ± 1.7	3,5-Dichlorosyringol	
011	−40.3 ± 0.6	−33 ± 3	7.3 ± 3.1	2-Chlorosyringaldehyde	
012	−17.0 ± 0.6	−15.7 ± 0.9	1.3 ± 1.1	Dihydrocarvone	

Table 1 continued

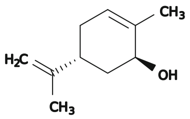
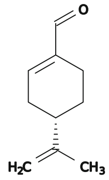
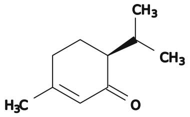
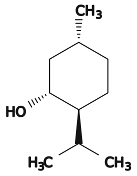
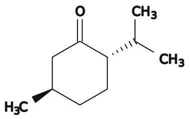
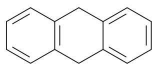
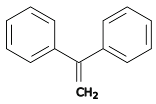
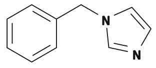
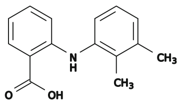
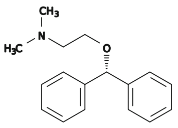
013	−13.5 ± 0.5	−18.6 ± 1.8	5.1 ± 1.9	Carveol	
014	−19.6 ± 0.6	−17.1 ± 0.7	2.5 ± 0.9	*l*-Perillaldehyde	
015	−18.8 ± 0.5	−18.9 ± 0.4	0.1 ± 0.6	Piperitone	
016	−16.7 ± 0.7	−13.4 ± 1.1	3.3 ± 1.3	Menthol	
017	−12.1 ± 0.6	−10.6 ± 1.0	1.5 ± 1.2	Menthone	
019	−3.8 ± 0.5	−15.8 ± 0.4	12.0 ± 0.6	9,10-Dihydroanthracene	
020	−3.8 ± 0.4	−11.6 ± 0.4	7.8 ± 0.6	1,1-Diphenylethene	
021	−27.6 ± 0.5	−31.9 ± 0.5	4.3 ± 0.7	1-Benzylimidazole	
022	−36.9 ± 0.7	−28.4 ± 0.4	8.5 ± 0.8	Mefenamic acid	
023	−12.0 ± 0.8	−39.1 ± 2.6	27.1 ± 2.7	Diphenhydramine	

Table 1 continued

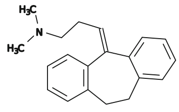
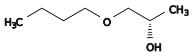
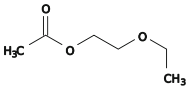
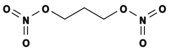
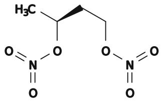
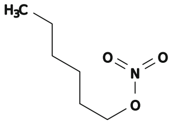
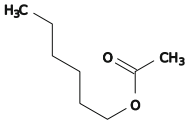
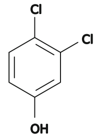
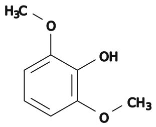
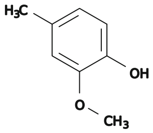
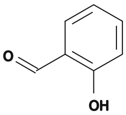
024	−2.9 ± 0.6	−31.1 ± 2.5	28.2 ± 2.6	Amitriptyline	
025	−18.3 ± 0.5	−24.0 ± 0.6	5.7 ± 0.8	1-Butoxy-2-propanol	
026	−23.1 ± 0.6	−22.2 ± 0.4	0.9 ± 0.7	2-Ethoxyethyl acetate	
027	−15.9 ± 0.5	−20.1 ± 1.6	4.2 ± 1.7	1,3-Bis-(nitrooxy)propane	
028	−13.7 ± 0.5	−18.0 ± 1.6	4.3 ± 1.7	1,3-Bis-(nitrooxy)butane	
029	−5.0 ± 0.6	−7.0 ± 0.4	2.0 ± 0.7	Hexyl nitrate	
030	−19.5 ± 0.5	−9.6 ± 0.5	9.9 ± 0.7	Hexyl acetate	
032	−32.6 ± 0.4	−30.5 ± 0.4	2.1 ± 0.6	3,4-Dichlorophenol	
033	−30.3 ± 0.6	−29.1 ± 0.4	1.2 ± 0.7	2,6-Dimethoxyphenol	
034	−36.6 ± 0.6	−24.3 ± 0.4	12.3 ± 0.7	4-Methyl-2-methoxyphenol	
035	−23.9 ± 0.7	−19.6 ± 0.4	4.3 ± 0.8	2-Hydroxybenzaldehyde	

Table 1 continued

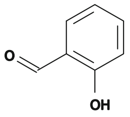
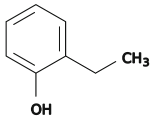
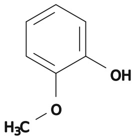
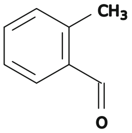
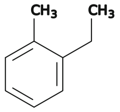
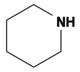
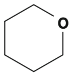
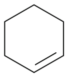
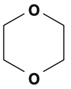
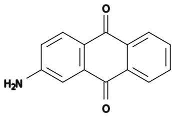
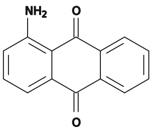
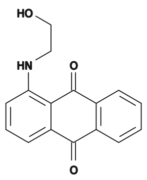
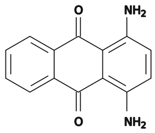
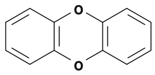
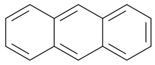
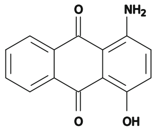
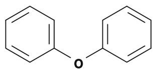
035	−23.9 ± 0.7	−19.6 ± 0.4	4.3 ± 0.8	2-Hydroxybenzaldehyde	
036	−24.1 ± 0.5	−23.7 ± 0.4	0.4 ± 0.6	2-Ethylphenol	
037	−41.6 ± 0.7	−24.9 ± 0.4	16.7 ± 0.8	2-Methoxyphenol	
038	−19.4 ± 0.5	−16.5 ± 0.4	2.9 ± 0.6	2-Methylbenzaldehyde	
039	1.4 ± 0.5	−3.6 ± 0.4	5.0 ± 0.6	1-Ethyl-2-methylbenzene	
041	−5.4 ± 0.5	−21.1 ± 0.4	15.7 ± 0.6	Piperidine	
042	−4.1 ± 0.5	−13.1 ± 0.4	9.0 ± 0.6	Tetrahydropyran	
043	7.6 ± 0.4	0.6 ± 0.4	7.0 ± 0.6	Cyclohexene	
044	−7.6 ± 0.5	−21.3 ± 0.4	13.7 ± 0.6	1,4-Dioxane	
045	−62.0 ± 0.6	−48.3 ± 1.2	13.7 ± 1.3	2-Amino-9,10-anthraquinone	
046	−43.7 ± 0.6	−40 ± 3	3.7 ± 3.1	1-Amino-9,10-anthraquinone	

Table 1 continued

047	−51.6 ± 1	−59 ± 5	7.4 ± 5.1	1-(2-Hydroxyethylamino)-9,10-anthraquinone	
048	−61.0 ± 0.6	−49.6 ± 1.5	11.4 ± 1.6	1,4-Diamino-9,10-anthraquinone	
049	−11.4 ± 0.5	−13.2 ± 0.4	1.8 ± 0.6	Dibenzo-*p*-dioxin	
050	−9.1 ± 0.4	−17.3 ± 0.4	8.2 ± 0.6	Anthracene	
051	−52.9 ± 0.7	−39.9 ± 1.2	13.0 ± 1.4	1-Amino-4-hydroxy-9,10-anthraquinone	
052	−8.0 ± 0.5	−12.0 ± 2.9	4.0 ± 2.9	Diphenyl ether	

All values are in kJ/mol

[a] Molecule identification number as assigned in the SAMPL4 challenge

[b] Hydration free enthalpy calculated using united atom parameters from ATB 2.0

[c] Experimental hydration free enthalpy

[d] The unsigned difference between calculated and experimental free enthalpies

to that obtained after geometry optimization at the B3LYP/6-31G* level of theory [24–26] in implicit solvent (water) using GAMMES-US [27]. The RMSD after performing a least squares fit on all atoms was calculated for each of the 214 molecules. The maximum value of the RMSD was 0.073 nm. Over 66 % of molecules had an RMSD value below 0.01 nm. Approximately 94 % had an RMSD value below 0.03 nm. This suggests that the geometry of the molecules is well maintained in all cases.

Hydration free enthalpies

The hydration free enthalpies calculated using united atom (UA) and all atom (AA) topologies for the 167 test molecules used previously are provided as supplementary material (Supplementary material Table S1). Values for the other 47 molecules that formed part of the SAMPL4 challenge are listed in Table 1 with additional information being provided as supplementary material (Supplementary material Table S2). The results for all 214 molecules are also presented graphically in Fig. 1, which shows a plot of the values calculated using UA parameters versus the experimental values. The 167 molecules (Table S1) are shown as blue crosses while the SAMPL4 molecules (Table 1) are indicated by yellow triangles. The solid line has a slope of one and represents a one-to-one agreement between the calculated and experimental numbers. The two dotted lines represent a 5 kJ/mol deviation from the ideal line. As can be seen in Fig. 1 the points are approximately equally distributed about the line corresponding to a one-to-one agreement between the calculated and experimental values. The overall statistics for the comparison to the available experimental data are given in Table 2. For the UA topologies the average error (AE) was 0.29 kJ/mol, the root mean square error (RMSE) was 9.5 kJ/mol, the average unsigned error (AUE) was 6.7 kJ/mol, the Kendall tau statistic (Tau) was 0.75, the Pearson correlation coefficient (R) was 0.91 and the slope of a line of best fit using linear regression was 1.12. Given the fact that the GROMOS 53A6 is a united atom force field, it is to be expected that the results for the UA topologies are slightly better than for the AA topologies. It should also be noted that the results obtained with ATB version 2.0 are essentially identical to those obtained using version 1.0. Values for 167 molecules calculated using version 1.0 are provided as supplementary material (Supplementary material Table S1). The

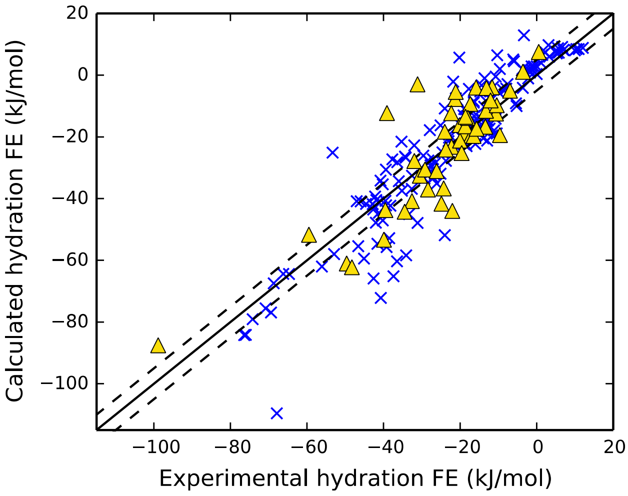


Fig. 1 A plot of the calculated versus experimental hydration free enthalpies (FE) for 214 molecules. Values were calculated using united atom ATB 2.0 topologies. SAMPL4 molecules (Table 1) are indicated by *yellow triangles*. The remaining 167 molecules indicated by *blue crosses* are described in the supplementary material (Table S1). The *solid line* has a slope of one and represents a one-to-one agreement between the calculated and experimental numbers. The *two dotted lines* represent a 5 kJ/mol deviation from the ideal line

differences in the versions relevant to this study are primarily related to the treatment of symmetry in the molecules and the assignment of charge groups. Namely, it is ensured that chemically equivalent groups are assigned identical partial charges and where possible atoms are grouped into neutral charge-groups in-line with the design of the GROMOS force field. This involved small rearrangements in the assignment of partial charges. However, as these changes were small, no significant change in the hydration free enthalpies was expected. A full description of the ATB version 2.0 will be presented elsewhere. The statistics for the SAMPL4 molecules were similar to those obtained for the whole data set and are discussed in more detail later.

Table 2 Statistics for the comparison of the experimental and calculated free enthalpies of hydration for different sets of molecules calculated using parameters generated by the Automated Topology Builder (ATB)

	AE[a]	RMSE[b]	AUE[c]	Tau[d]	R[e]	Slope[f]
Total						
UA[g]	0.29	9.49	6.71	0.75	0.91	1.12
AA[h]	−3.43	10.31	7.21	0.77	0.92	1.17
Small organic	0.56	3.98	3.37	0.74	0.95	1.04
SAMPL0	3.85	8.50	7.16	0.76	0.91	1.21
SAMPL1	−0.03	13.34	9.63	0.64	0.82	1.33
SAMPL2	−2.78	10.51	8.53	0.83	0.95	1.23
SAMPL4						
Sub.[i] UA	1.09	10.02	7.61	0.60	0.85	0.99
Rev.[j] UA	0.96	9.96	7.53	0.60	0.85	1.00
AA	−3.60	9.82	7.75	0.62	0.89	1.11

Values are in kJ/mol

[a] Average error

[b] Root mean square error

[c] Average unsigned error

[d] Kendall's tau statistic

[e] Pearson's correlation coefficient

[f] Slope of linear regression

[g] United atom force field

[h] All atom force field

[i] Values submitted as part of the SAMPL4 challenge

[j] Revised values incorporating an analysis of convergence at each λ-value

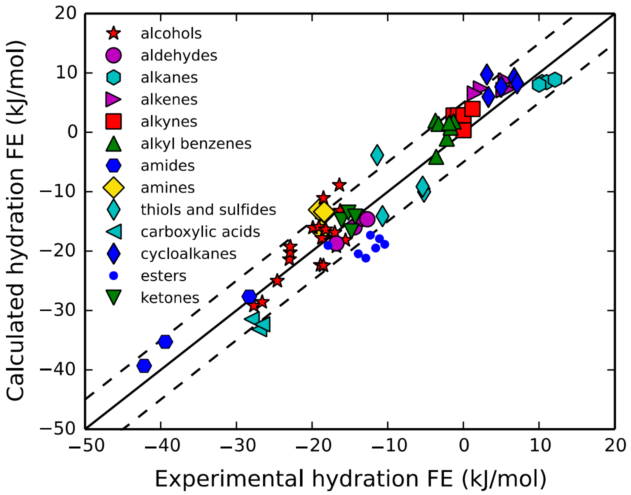


Fig. 2 Comparison of calculated with experimental hydration free enthalpy (FE) values for 75 small organic molecules classified by a characteristic functional group

A set of 75 small organic molecules for which high quality solvation free enthalpy data is available was used as an initial test of the ATB. This test set consisted of alcohols, alkanes, cycloalkanes, alkenes, alkynes, alkyl benzenes, amines, amides, aldehydes, carboxylic acids, esters, ketones, thiols and sulphides. The AUE for these molecules was 3.4 kJ/mol, the RMSE was 4.0 kJ/mol and 77 % of the molecules lay within 5 kJ/mol of the experimental value. The largest deviation from experiment was 8.5 kJ/mol. What is clear from this result is that while the ATB parameters perform well for the majority of molecules, certain functional groups lead to systematic deviations from experiment. This is illustrated graphically in Fig. 2 which shows a plot of the calculated versus experimental hydration free enthalpies for molecules containing a single identifiable functional group. While alcohols, thiols/sulphides, ketones and aldehydes are on average evenly distributed around the experimental values (average signed error <2 kJ/mol), the hydration free enthalpies of esters and carboxylic acids are systematically underestimated by 6 and 5 kJ/mol respectively. In contrast, alkyl benzene groups and alkenes as well as amides and primary amines are overestimated by between 3 and 5 kJ/mol on average.

Of the set of 167 molecules, 92 were taken from previous SAMPL challenges. The AUE for molecules in the SAMPL0, SAMPL1 and SAMPL2 data sets was 7.2, 9.6 and 8.5 kJ/mol respectively. The RSME for molecules in the SAMPL0, SAMPL1 and SAMPL2 data sets was 8.5, 13.3 and 10.5 kJ/mol respectively. These are significantly larger than for other molecules in the data set and dominate the statistics. Approximately 40 % of the molecules in the SAMPL data sets still lay within 5 kJ/mol of the experimental value but, the largest deviation from experiment was 42 kJ/mol. This is in part a reflection of the uncertainty in the hydration free enthalpies of molecules contained in SAMPL challenges (which were as large as 8 kJ/mol) and in part a reflection of the fact that these molecules contained a range of functional groups not commonly found in biomolecular systems. For example, molecules containing multiple halogens showed the largest deviations from experiment. This suggests that it will be possible to greatly improve the overall performance of the ATB by optimizing the parameters for a small number of atom types. Indeed, sulphur-containing compounds, carboxylic acids, esters, amides and amines are known to be not optimal within the GROMOS force field [19].

Analysis of the SAMPL4 data set

The hydration free enthalpies that were submitted as part of the SAMPL4 challenge (id 529) were obtained using UA topologies and calculated over 2 days using an initial iteration of the automated pipeline described in the methods. The values and overall statistics for ATB 2.0 UA topologies using an updated version of our automated pipeline with improved convergence checking are provided in Tables 1 and 2. The hydration free enthalpies for 23 of the 47 molecules were predicted within 5 kJ/mol of the experimental value. The largest deviations from experiment, 27 and 28 kJ/mol, were for two aliphatic tertiary amines diphenhydramine (023) and amitriptyline (024), respectively. Other molecules for which the calculated hydration free enthalpy deviated significantly from experiment included piperidine (041), which contains a secondary amine, and 4-propylguaiacol (006), 4-methyl-2-methoxyphenol (034) and 2-methoxyphenol (037) each of which contains a methoxy group.

Automated TI protocol

The automated protocol to obtain the hydration free enthalpies based on thermodynamic integration (TI) proved highly effective. TI was the method of choice because the convergence of the overall integral can be effectively monitored and systematically improved. In TI, the convergence does not rely on the degree of overlap of two ensembles and is not dependent on an exponentially weighted function. To maximise efficiency of the method the λ-values were preferentially placed in regions of high curvature and the convergence at each point monitored independently. This ensured sampling was concentrated in those regions that had the greatest impact on the overall hydration free enthalpy. Plots of $\langle \partial V(\mathbf{r})/\partial\lambda \rangle_\lambda$ versus λ for 3 example molecules are shown in Fig. 3. The individual lines in each panel represent the change in free enthalpy in water, vacuum and the difference between vacuum and water. Note, an implicit assumption in the estimation of the total error is that existing points represent, to some degree, the highest order feature of the underlying curve. In some sense this aspect of the problem is irreducible, as the form of the underlying function is not known. However in practice, given the shape of the curves illustrated in Fig. 3, 9 equally spaced points were sufficient to identify the turning point in all cases.

Overall statistics for two sets of calculated values for molecules in the SAMPL4 challenge using UA topologies are listed in Table 2. For the values submitted as part of the

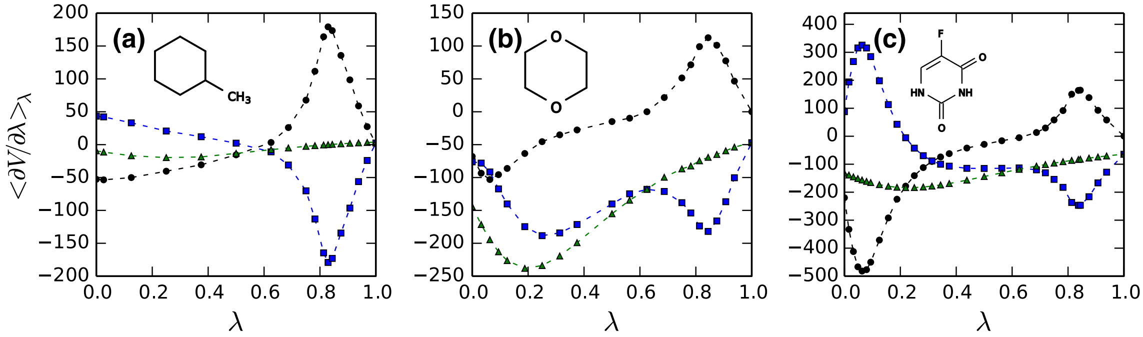


Fig. 3 Example thermodynamic integration curves generated by the automated protocol for methylcyclohexane **a** 1,4-dioxane **b** 5-fluorouracil **c**. Values of $\langle \partial V(\mathbf{r})/\partial\lambda \rangle_\lambda$ in water and vacuum are shown in squares and triangles respectively, the final free enthalpy curve (Eq. 2) is shown in *circles*

Table 3 Average simulation lengths and computational times required to achieve a given statistical error

	1 kJ/mol error	2 kJ/mol error	3 kJ/mol error	4 kJ/mol error
Simulation time (ns)	14 ± 4	12 ± 3	10 ± 3	9 ± 2
CPU (h)	175 ± 50	150 ± 37	125 ± 37	112 ± 25
FE difference (kJ/mol)	–	0.6 ± 0.6	1.1 ± 0.9	1.2 ± 0.9

The free enthalpy (FE) difference is the average difference between the free enthalpies obtained for a given statistical error calculated with respect to an error of 1 kJ/mol

challenge (sub. UA) the system was simulated two times for 200 ps at each λ-value. The revised values (rev. UA) were obtained after the values of $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ had been deemed to have converged at each λ-value based on the criteria described above. Overall, the difference between the two sets is negligible. However, by ensuring the convergence of $<\partial V(\boldsymbol{r})/\partial\lambda>_\lambda$ at each λ-value the overall number of points required to achieve a specific integration error could be greatly reduced resulting in a twofold increase in computational efficiency with no loss of precision. Note, in all but one case the systems were simulated until the uncertainty in the integration was ≤1 kJ/mol. In one case the algorithm was terminated once a total time limit was reached (SAMPL4 ID 001). In many cases the integration error was significantly less than one. For these cases the computational efficiency of the algorithm could be improved further by lowering the default initial sampling values.

Computational cost

The computational costs associated with the analysis of the SAMPL4 results comprised of two parts: the generation of the parameters and the calculation of the free enthalpy values themselves. The time required to generate the parameters is dominated by the time needed to optimize the molecules at the B3LYP/6-31G* level of theory which is highly dependent on the size of the molecule. The average time required for the optimization of SAMPL4 molecules was 26 central processing unit (CPU) hours. The average simulation length and computational time used to obtain the values listed in Table 1 for the SAMPL4 compounds are shown in Table 3. As can be seen, to achieve a statistical uncertainty of 1.0 kJ/mol or less, the mean total simulation length per molecule was 14 ± 4 ns and the mean time per molecule was 175 ± 50 CPU hours. Table 3 also shows how the average simulation time and final result vary with the statistical uncertainty. The last row in Table 3 illustrates the average difference in the results compared to that obtained using a tolerance of 1.0 kJ/mol. Note the actual difference between the results is much less than the statistical uncertainty.

Conclusions

A set of 214 molecules including those of the SAMPL0, 1, 2 and 4 challenges has been used to test and validate the all atom and united atom topologies generated using the latest version of the Automated Topology Builder (ATB version 2.0) against structural (optimised geometries) and thermodynamic (hydration free enthalpies) data. Very good agreement between the QM optimized structures and the energy minimized structures was obtained. There was also good overall agreement between the predicted and experimental hydration free enthalpies for the majority of molecules investigated. For 117 of 214 molecules examined, the predicted hydration free enthalpy was within 5 kJ/mol of the experimental value, with the AUE between the calculated and experimental values of 6.7 kJ/mol and the RMSE of 9.5 kJ/mol. The AUE for a set of small organic molecules with high quality hydration free enthalpy data was only 3.4 kJ/mol and the RMSE was 4.0. The AUE for SAMPL0, 1, 2 and 4 ranged between 7.2 and 9.6 kJ/mol with the RMSE between 8.5 and 13.3 kJ/mol reflecting both the intrinsic uncertainty in some of the experimental values included in the SAMPL data sets, as well as the fact that the GROMOS force field is primarily intended for biomolecular systems and has yet to be optimized for certain functional groups. This suggests that further significant improvements in the predictive ability of the ATB will be possible. Finally, it should be noted that the values presented are based on fully automated protocols that require no manual intervention. The actual values submitted as part of the SAMPL4 challenge itself were generated over a period of 48 h using a distributed computing resource. The implementation of robust parameterization and validation protocols within the ATB combined with the increasing availability of distributed computing resources provides the potential to perform free enthalpy calculations in a high throughput manner and undertake large-scale optimization of molecular force fields.

Acknowledgments The development of the ATB has been supported by the Australian Research Council (ARC; Grant No.: DP0987043, ARC DP110100327, ARC DP130102153) AKM acknowledges the award of an Australian Post-Doctoral (APD) fellowship. KBK acknowledges the award of the University of Queensland International Scholarship (UQI). Computational resources were provided by the National Computational Infrastructure (NCI, Australia) National Facility (projects m72 and n63). Support from the Queensland Cyber Infrastructure Foundation and NeCTAR is gratefully acknowledged. The authors also wish to thank W. Chen, M. El-Kebir, G. Klau and D. Warne for assistance in the development of the ATB version 2.0 and associated infrastructure.

References

1. Verlinde CLMJ, Hol WGJ (1994) Structure-based drug design: progress, results and challenges. Structure 2(7):577–587
2. Tollenaere JP (1996) The role of structure-based ligand design and molecular modelling in drug discovery. Pharm World Sci 18(2):56–62
3. Ooms F (2000) Molecular modeling and computer aided drug design. Examples of their applications in medicinal chemistry. Curr Med Chem 7(2):141–158
4. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graphics Model 25(2):247–260
5. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25(9):1157–1174
6. Krieger E, Koraimann G, Vriend G (2002) Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. Proteins 47(3):393–402
7. Zoete V, Cuendet MA, Grosdidier A, Michielin O (2011) SwissParam: a fast force field generation tool for small organic molecules. J Comput Chem 32(11):2359–2368
8. Patel S, Brooks CL 3rd (2004) CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. J Comput Chem 25(1):1–15
9. Patel S, Mackerell AD Jr, Brooks CL 3rd (2004) CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. J Comput Chem 25(12):1504–1514
10. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comp Phys Comm 91(1–3):43–56
11. van der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast, flexible, and free. J Comput Chem 26(16):1701–1718
12. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17(5–6):490–519
13. Halgren TA (1996) Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comput Chem 17(5–6):520–552
14. Halgren TA (1999) MMFF VII. Characterization of MMFF94, MMFF94 s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. J Comput Chem 20(7):730–748
15. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD (2010) CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem 31(4):671–690
16. Vanommeslaeghe K, MacKerell AD (2012) Automation of the CHARMM general force field (CGenFF) I: bond perception and atom typing. J Chem Inf Model 52(12):3144–3154
17. Vanommeslaeghe K, Raman EP, MacKerell AD (2012) Automation of the CHARMM general force field (CGenFF) II: assignment of bonded parameters and partial atomic charges. J Chem Inf Model 52(12):3155–3168
18. Malde AK, Zuo L, Breeze M, Stroet M, Poger D, Nair PC, Oostenbrink C, Mark AE (2011) An Automated Force Field Topology Builder (ATB) and Repository: version 1.0. J Chem Theory Comput 7(12):4026–4037
19. Oostenbrink C, Villa A, Mark AE, van Gunsteren WF (2004) A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem 25(13):1656–1676
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
21. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 107(13):3902–3909
22. Stewart JJP (1989) Optimization of parameters for semiempirical methods I. Method. J Comput Chem 10(2):209–220
23. Stewart JJP (1989) Optimization of parameters for semiempirical methods II. Applications. J Comput Chem 10(2):221–264
24. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98(7):5648–5652
25. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37(2):785–789
26. Perdew JP, Wang Y (1992) Accurate and simple analytic representation of the electron-gas correlation energy. Phys Rev B 45(23):13244–13249
27. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH, Koseki S, Matsunaga N, Nguyen KA, Su S, Windus TL, Dupuis M, Montgomery JA (1993) General atomic and molecular electronic structure system. J Comput Chem 14(11):1347–1363
28. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. J Comput Chem 5(2):129–145
29. Canzar S, El-Kebir M, Pool R, Elbassioni K, Malde AK, Mark AE, Geerke DP, Stougie L, Klau GW (2013) Charge group partitioning in biomolecular simulation. J Comput Biol 20(3):188–198
30. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) Interaction models for water in relation to protein hydration. In: Pullman B (ed) Intermolecular forces. Springer, The Netherlands, pp 331–342
31. van Gunsteren WF, Weiner PK, Wilkinson T, Wilkinson AJ (1997) Computer simulation of biomolecular systems: theoretical and experimental applications. Springer, Leiden
32. Beutler TC, Mark AE, van Schaik RC, Gerber PR, van Gunsteren WF (1994) Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. Chem Phys Lett 222(6):529–539
33. Zacharias M, Straatsma TP, McCammon JA (1994) Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. J Chem Phys 100(12):9025–9031
34. Allen P, Tildesley DJ (1989) Computer simulation of liquids. Oxford University Press Inc, New York
35. Schmid N, Christ CD, Christen M, Eichenberger AP, van Gunsteren WF (2012) Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. Comput Phys Commun 183(4):890–903
36. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23(3):327–341
37. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81(8):3684–3690
38. Heinz TN, van Gunsteren WF, Hünenberger PH (2001) Comparison of four methods to compute the dielectric permittivity of liquids from molecular dynamics simulations. J Chem Phys 115(3):1125–1136
39. 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
40. Guthrie JP (2009) A blind challenge for computational solvation free energies: introduction and overview. J Phys Chem B 113(14):4501–4507
41. 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