JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Implicit Solvation Parameters Derived from Explicit Water Forces in Large-Scale Molecular Dynamics Simulations

Jens Kleinjung,[†] Walter R. P. Scott,[‡] Jane R. Allison,[¶] Wilfred F. van Gunsteren,[¶] and Franca Fraternali*[,§]

[†]Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom

[‡]Chemistry Department, University of British Columbia, 2036 Main Mall, Vancouver, British Columbia, Canada V6T 1Z1

[¶]Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich, ETH Hönggerberg, CH-8093 Zürich, Switzerland

[§]Randall Division of Cell and Molecular Biophysics, King's College London, New Hunt's House, London SE1 1UL, United Kingdom

Supporting Information

ABSTRACT: Implicit solvation is a mean force approach to model solvent forces acting on a solute molecule. It is frequently used in molecular simulations to reduce the computational cost of solvent treatment. In the first instance, the free energy of solvation and the associated solvent−solute forces can be approximated by a function of the solvent-accessible surface area (SASA) of the solute and differentiated by an atom−specific solvation parameter σ_i^{SASA}. A procedure for the determination of values for the σ_i^{SASA} parameters through matching of explicit and implicit solvation forces is proposed. Using the results of Molecular Dynamics simulations of 188 topologically diverse protein structures in water and in implicit solvent, values for the σ_i^{SASA} parameters for atom types i of the standard amino acids in the GROMOS force field have been determined. A simplified representation based on groups of atom types σ_g^{SASA} was obtained *via* partitioning of the atom−type σ_i^{SASA} distributions by dynamic programming. Three groups of atom types with well separated parameter ranges were obtained, and their performance in implicit versus explicit simulations was assessed. The solvent forces are available at http://mathbio.nimr.mrc.ac.uk/wiki/Solvent_Forces.

INTRODUCTION

Proteins have evolved to function within the water−rich environment of the cell. Adaptation to the particular solvation properties of water, such as strong electrostatic shielding, hydrogen−bond donor/acceptor saturation, and entropic effects, led to the known segregation of predominantly hydrophobic residues in the core and polar/charged amino acid residues on the protein surface. The distribution of residue types on the protein surface determines its interaction with the surrounding bulk solvent and with other solute molecules.[1] These interactions define to a large extent the conformational equilibria and biological function of a protein. The range of accessible conformations under physiological conditions is the result of a delicate balance between competing forces: *(i)* highly anisotropic intraprotein interactions and *(ii)* approximately isotropic bulk−solvent interactions. It is therefore not surprising that the presence of water has become an integral part of protein folds by stabilizing secondary structure elements and their association.[2−4]

Biomolecular simulations account for the presence of water in the native environment either explicitly, by inclusion of water molecules, or implicitly, by approximating the mean force exerted by the water on the biomolecule. The latter is considerably faster to compute, because the implicit solvent does not contribute any degrees of freedom to the simulation, although it comes at the expense of a neglect of specific features such as water dipole orientation and hydrogen bond fluctuations at the surface of the solute. The extent to which a solvent model (explicit or implicit) can realistically reproduce the dominant physical forces in dynamic protein structures is therefore crucial to its success in describing conformational equilibria. Implicit solvation may be the best (or sole) choice for systems with a large number of degrees of freedom, for systems whose reference experiment spans a time scale inaccessible to current state-of-the-art explicit solvent Molecular Dynamics (MD) and also for enhanced sampling simulations, where conformational changes induced in the solute would lead to clashes with explicit water molecules.[5,6]

Because of their computational efficiency, implicit solvent models have been used in a large variety of computational studies, e.g. folding simulations,[7] energy scoring of protein structures,[8] structure prediction and design,[9,10] and membrane simulations.[11,12] Each model relies on approximations of the mean force contribution of the solvent to the overall energy of the solute molecule.

The starting point of most models is the use of a first-shell approximation of the solvent effect, i.e. the assumption that the force on a solute atom exerted by the solvent is on average proportional to the solvent-accessible surface area (SASA) of the solute atom. This simple approximation can be complemented by long-range electrostatic solvent contributions based on the approximation that the bulk solvent behaves as a dielectric continuum, leading to Poisson−Boltzmann (PB) or Generalized Born (GB) energy terms to describe the electrostatic interactions between the solvent and the partial charges of the solute.[13] Mixed models with GB/SASA terms are now widely used[14−17] and are

Received: June 9, 2011
Published: June 12, 2012

successful in predicting binding free energies. Comparisons of the performance of different implicit solvent models[18−22] revealed the lack of an accurate implicit treatment of nonpolar solvation in most models.[23] Complementing the surface term with a volume term improves the description of long-range solute−solvent interactions[24] and nonpolar contributions.[25,26]

In the past we have presented an efficient implicit solvent model for use in MD simulations based on a fast analytical approximation to the SASA.[27] The energy of solvation in this and other SASA-based models is expressed simply as $V_{sol} = \Sigma_i \sigma_i A_i$, where A_i denotes the SASA of atom i and σ_i^{SASA} an atom-specific solvation energy per surface area parameter, which is empirical in nature. The analytical formula used in the model for the fast evaluation of SASA is based on nearest-neighbor distances and was published by Hasel et al.[28] This model was incorporated into the GROMOS simulation package[29,30] and appropriate atomic σ_i^{SASA} parameter values compatible with the GROMOS force field parameter set 43A1 for biomolecular solutes were proposed.[27] The same model with a virtually identical parametrization was later used in conjunction with the CHARMM force field,[31] showing the validity of the solvation parameters independent of the solute force field employed.

Here we describe the derivation of values for the σ_i^{SASA} parameters for the atom types of the GROMOS force field 43A1 *via* force matching within the framework of large-scale explicit solvent MD simulations, where the time-averaged explicit forces on each solute atom type exerted by the explicit solvent molecules are transformed into an implicit mean solvation force and used to derive solvation parameters. The forces on the protein atoms due to the explicit water molecules averaged over the trajectories have been made available on our Web server to encourage further development of the implicit solvation model and the refinement of parameters in protein force fields other than the GROMOS one.

METHODS

A Mean Plus Stochastic Force Representation of the Solvent. If the solvent degrees of freedom are not explicitly simulated, one may approximate the force $\mathbf{f}_i^{solv}$ exerted by the solvent on atom i of the solute by a mean force $\mathbf{f}_i^{mean}$, which can be derived from a potential energy term, a potential of mean force $V^{mean}(\mathbf{r}^N)$

$$\mathbf{f}_i^{mean}(\mathbf{r}^N) = -\frac{\partial V^{mean}(\mathbf{r}^N)}{\partial \mathbf{r}_i} \tag{1}$$

which represents the averaged effect of the omitted solvent degrees of freedom on the solute. A solute configuration is represented by $\mathbf{r}^N \equiv (\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_N)$, the Cartesian coordinates of all N solute atoms. A higher-order approximation of the force $\mathbf{f}_i^{solv}$ exerted by the solvent on the solute is obtained by considering not only its mean effect but also its fluctuations in time and its frictional effect[32]

$$\mathbf{f}_i^{solv}(\mathbf{r}^N) = \mathbf{f}_i^{mean}(\mathbf{r}^N) + \mathbf{f}_i^{stoch} - m_i\gamma_i\mathbf{v}_i \tag{2}$$

The stochastic force is denoted by $\mathbf{f}_i^{stoch}$, and the frictional force is proportional to the solute atom velocity $\mathbf{v}_i$ with proportionality factor $m_i\gamma_i$, in which γ_i is the atomic friction coefficient and m_i is the mass of solute atom i. A stochastic force introduces energy into the system and a frictional force removes energy from it. The condition of zero energy loss on average will relate the two forces. If the stochastic force $\mathbf{f}_i^{stoch}$ obeys a Gaussian probability distribution with zero mean, if it is not correlated with prior velocities or forces, and if the friction coefficient is independent of time, this condition reads

$$\langle(\mathbf{f}_i^{stoch})^2\rangle = 6m_i\gamma_i k_B T_{ref} \tag{3}$$

where a (time) average is denoted by $\langle\cdots\rangle$, k_B is Boltzmann's constant, and T_{ref} is the reference temperature of the system. Combination of eq 2 with Newton's equation of motion leads to the stochastic Langevin equation of motion

$$d\mathbf{v}_i(t)/dt = m_i^{-1}(\mathbf{f}_i^{\mathrm{int}}(\mathbf{r}^N(t)) + \mathbf{f}_i^{mean}(\mathbf{r}^N(t)) + \mathbf{f}_i^{stoch}) - \gamma_i\mathbf{v}_i(t) \tag{4}$$

in which $\mathbf{f}_i^{int}$ represents the intrasolute forces exerted by the atoms of the solute on solute atom i. Numerical integration of such a stochastic equation of motion is called stochastic dynamics (SD) simulation.[33] More sophisticated forms of SD than eqs 3 and 4 can be obtained by incorporating temporal and spatial correlations in the description of the stochastic force $\mathbf{f}_i^{stoch}$. $\mathbf{f}_i^{int}(\mathbf{r}^N)$ are calculated from the solute configuration $\mathbf{r}^N$ using the GROMOS biomolecular force field. The mean force $\mathbf{f}_i^{mean}(\mathbf{r}^N)$ due to the solvent is obtained from its energetic contribution to solvation of a solute molecule, which is here treated as approximately proportional to the molecular SASA, given by the sum of the atomic SASA contributions A_i

$$V^{impl}(\mathbf{r}^N) = \sum_{i=1}^{N} \sigma_i^{SASA} A_i(\mathbf{r}^N) \tag{5}$$

The derivative of eq 5 with respect to $\mathbf{r}_i$ yields the implicit solvent force $\mathbf{f}_i^{impl}$ on atom i

$$\mathbf{f}_i^{impl} = -\sigma_i^{SASA}\frac{\partial A_i(\mathbf{r}^N)}{\partial \mathbf{r}_i} \tag{6}$$

The implicit solvent force is proportional to the atomic SASA change ∂A_i that is associated with a change $\partial\mathbf{r}_i$. An analytical formula for the SASA computation has been described elsewhere;[27,28] it is recapitulated here for completeness of the methodological procedure. The SASA of atom i is given by the analytical formula

$$A_i(\mathbf{r}^N) = S_i \prod_{\substack{j=1 \\ j\neq i}}^{N} \left[1 - p_i p_{ij}\frac{b_{ij}(r_{ij})}{S_i}\right] \tag{7}$$

where the parameter S_i denotes the surface of the isolated atom i, and the terms p_i, p_{ij}, and b_{ij} are geometric parameters that describe the reduction of SASA depending on the atom type i and the neighbor atom types j and $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$. The derivative $(\partial A_i)/(\partial\mathbf{r}_i)$ as required for eq 6 is given in the Appendices of Hasel et al.[28] and Allison et al.[24] Geometric parameters are reported in Hasel et al.,[28] and specific p_i values for the GROMOS[29] atom types are reported in Fraternali and van Gunsteren.[27]

The stochastic force $\mathbf{f}_i^{stoch}(t)$ and the atomic friction coefficient γ_i will only be sizable for solute atoms at the surface. Therefore, they are taken dependent on the number of neighbor atoms[32]

$$\gamma_i(t) = \gamma_{solv}\omega_i(t) \tag{8}$$

with

$$\omega_i(t) = \max(0,\ 1 - N^{nb}(t)/N^{nbref}) \tag{9}$$

where $N_i^{nb}(t)$ denotes the number of non-hydrogen neighbor atoms of the solute within 0.3 nm radius, and N^{nbref} was defined as

an upper limit of 6 neighbor solute atoms at which solvent forces on solute atom i are assumed to vanish.

A Procedure To Determine the Implicit Solvation Parameters σ_i^{SASA}. Previously,[27] the parameters σ_i^{SASA} of the model were derived by a calibration of the radius of gyration, and the hydrophobic and hydrophilic SASA obtained in MD simulations with the implicit solvent model against these quantities obtained in MD simulations with explicit water molecules, for three proteins of different sizes. Here we propose an alternative procedure in which the σ_i^{SASA} are determined such that the implicit solvation forces $\mathbf{f}_i^{impl}$ on the solute atoms i match as closely as possible the corresponding average forces $\langle \mathbf{f}_i^{expl} \rangle$ that are exerted on the solute atoms i by the solvent molecules in an explicit solvent MD simulation.

This matching of $\mathbf{f}_i^{impl}$ to $\langle \mathbf{f}_i^{expl} \rangle$ is not straightforward though.

1 For a given solute configuration $\mathbf{r}^N$, the direction of $\mathbf{f}_i^{impl}$ is determined by eq 6, i.e. it is roughly perpendicular to the SASA A_i in the neighborhood of atom i, while the direction of $\langle \mathbf{f}_i^{expl} \rangle$ is determined by an average of the configurations of many explicit solvent molecules (Figure 1), which means that $\mathbf{f}_i^{impl}$ cannot faithfully represent $\langle \mathbf{f}_i^{expl} \rangle$. In order to minimize noise in the calibration data set we omit those atoms for which the projection of $\langle \mathbf{f}_i^{expl} \rangle$ onto $\mathbf{f}_i^{impl}$ is smaller than its component orthogonal to $\mathbf{f}_i^{impl}$.

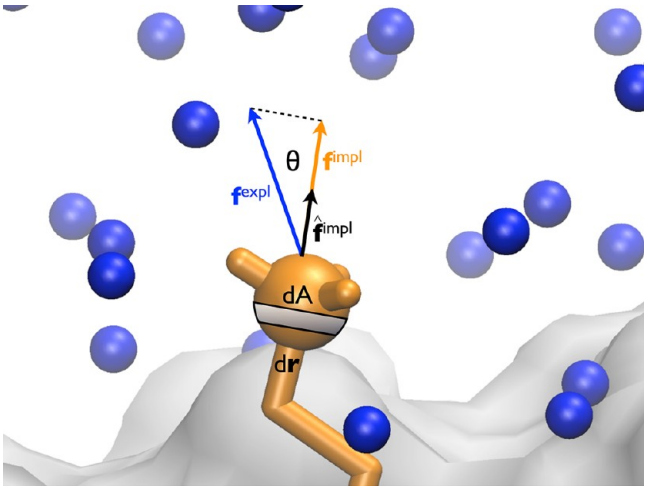


Figure 1. Force matching. Projection of the explicit solvent force $\mathbf{f}_i^{expl}$ on a solute atom onto the unit vector of the implicit solvent force $\hat{\mathbf{f}}_i^{impl}$ direction yields the implicit solvent force $\mathbf{f}_i^{impl}$. The implicit force direction is determined by the derivative $-\partial A_i/\partial \mathbf{r}_i$.

2 The forces $\mathbf{f}_i^{impl}$ $(\mathbf{r}^N)$ and $\langle \mathbf{f}_i^{expl} \rangle$ are straightforwardly matched only when the solute configuration $\mathbf{r}^N$ is the same for all the solvent configurations over which the average in $\langle \mathbf{f}_i^{expl} \rangle$ is carried out. Therefore, we keep the solute configuration $\mathbf{r}^N$ fixed in the MD simulations that were used to obtain the average forces due to the explicit solvent molecules.

3 The $\langle \mathbf{f}_i^{expl} \rangle$ are obtained from MD simulations of proteins in explicit solvent in which the solute force field is compatible with the solvent one, in the present case the GROMOS force field[29,34] 43A1, which is compatible with the simple-point charge (SPC) water model.[35] When omitting the solvent degrees of freedom in an implicit solvent simulation, their dielectric screening effect should somehow be retained. In the GROMOS 43B1 solute force field for *in vacuo* simulations, this is achieved by adapting the partial charges of groups of atoms that bear a total charge of $\pm 1e$, such as in Asp, Glu, Arg, and Lys side chains, such that their net charge becomes zero, while their hydrogen-bonding capacity is retained.[29]

Determination of the σ_i^{SASA} Parameter Values. The implicit solvation force $\mathbf{f}_i^{impl}$ on each atom i should match the strength of the average explicit force $\langle \mathbf{f}_i^{expl} \rangle$ exerted by the surrounding solvent along the direction of $\mathbf{f}_i^{impl}$

$$|\mathbf{f}_i^{impl}| = \hat{\mathbf{f}}_i^{impl} \cdot \left\langle \mathbf{f}_i^{expl} \right\rangle \tag{10}$$

where the hat ^ indicates a vector of unit length, the ensemble average $\langle\rangle$ is over the MD configurations from a simulation in explicit solvent, and the index i denotes a particular atom in a given molecule. This projection is illustrated in Figure 1. Solving eq 6 and eq 10 for σ_i^{SASA} yields

$$\sigma_i^{SASA} = -\frac{\frac{\partial A_i}{\partial \mathbf{r}_i}}{\left|\frac{\partial A_i}{\partial \mathbf{r}_i}\right|^2} \cdot \langle \mathbf{f}_i^{expl} \rangle \tag{11}$$

which means that the σ_i^{SASA} parameter values can be obtained from MD simulations: *(i)* the SASA derivative $\partial A_i/\partial \mathbf{r}_i$ from an implicit solvation simulation and *(ii)* the mean solvation force $\langle \mathbf{f}_i^{expl} \rangle$ from an explicit solvation simulation, under the constraint that the protein conformation is identical in both MD simulations.

The angle between $\mathbf{f}_i^{impl}$ and $\langle \mathbf{f}_i^{expl} \rangle$ is given by

$$\cos(\theta) = \frac{-\frac{\partial A_i}{\partial \mathbf{r}_i} \cdot \left\langle \mathbf{f}_i^{expl} \right\rangle}{\left|\frac{\partial A_i}{\partial \mathbf{r}_i}\right| \cdot |\langle \mathbf{f}_i^{expl} \rangle|} \tag{12}$$

Selection of the Reference Protein Set. The protein set for the parametrization of σ_i^{SASA} should contain a range of different folds to reflect the variation of protein structures. To this end a topological alphabet was devised to capture in simple terms the structural composition of protein folds, in loose analogy to the methods of Martin[36] and Kamat and Lesk.[37] The protein topology was described as the sequence of pairs of secondary structure elements (supersecondary structures), i.e. $\beta-\beta$, $\beta-\alpha$, $\alpha-\beta$, or $\alpha-\alpha$. To increase the resolution of the alphabet, the angle between the secondary structure elements was included as a geometric parameter, and three angle ranges 0−60°, 60−120°, and 120−180° were combined with the four topological states to yield 12 states of the alphabet (Table S1 in the Supporting Information). Using this alphabet, we translated a selected set of 2559 well-resolved protein domains of the SCOP ASTRAL40 database[38] to topological strings by assigning a topological alphabet character to each supersecondary structure. The concatenated characters form a topological string that characterizes basic features of the protein fold.

The selected SCOP set was reduced to less than 10% of its size by applying the MinSet method,[39] so as to derive a database subset that was amenable to MD simulations but maximally informative in terms of topological composition. Within the framework of a genetic algorithm, random domain subsets were created, their topological strings concatenated and assessed in terms of the overall string entropy, where the entropic score takes the original database composition into account. The subset with the highest entropy was chosen. Among all the created random subsets, this subset was the most informative with respect to the

topological composition, which does not exclude the presence of topologically similar proteins. The final list of 188 protein domains is given in Supporting Information Table S2.

Simulations. MD simulations were performed using the GROMOS biomolecular simulation software.[29,30] The employed force fields were GROMOS 43A1 for simulations in explicit solvent (water) and GROMOS 43B1 for implicit solvent. The integration time step was set to 2 fs. The temperature was set to 298 K and controlled by weak coupling to a temperature bath[40] with a coupling constant $\tau_T = 0.1$ ps. Simulations in explicit water were kept at a pressure of 0.061020 kJ mol^{-1} nm^{-3} (1 atm) with a coupling time of $\tau_P = 0.5$ ps and an isothermal compressibility of 5.575×10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$.

Bond lengths were constrained by the SHAKE algorithm.[41] Both simulation types (i.e., in explicit and implicit solvent) were run for 2.5×10^5 steps (500 ps), and configurations were saved at intervals of 500 steps (1 ps). Explicit solvent forces and SASAs were saved with each configuration. Since water equilibration around solutes occurs on the time scale of 10−20 ps, explicit solvent simulations of 500 ps are sufficiently long to sample representative force distributions.

Simulations in Explicit Solvent (Water). Initial protein structures (taken from the PDB database) were energy minimized using 100 steps of steepest descent. Energy minimized protein conformations were solvated in a periodic water box of either rectangular or truncated-octahedral shape, whichever was smaller and therefore matched the overall protein shape better. The dimensions of all periodic boxes were larger than twice the nonbonded cutoff radius of 1.4 nm (the shortest box axis length was 5.7 nm), and the distance between solute and box was set to 0.75 nm (rectangular box) or 0.85 nm (octahedral box). Systems were electrostatically neutralized by replacing water molecules with sodium or chloride ions to compensate the net charge of the protein at neutral pH value. The neutralized systems were energy minimized using 100 steps of steepest descent, while the solutes (protein domains) were harmonically positionally restrained using a force constant of 2.5×10^4 kJ mol^{-1} nm^{-2}. The systems were run for 2.5×10^5 steps of MD while keeping the solute positionally constrained. Twin-range cutoff radii of 0.8/1.4 nm were used to compute nonbonded interactions. The nonbonded pair list was updated every time step for pairs within 0.8 nm and every fifth time step for the range 0.8−1.4 nm. Long-range electrostatic interactions were approximated by a reaction-field force, using a dielectric constant of 54.

Simulations in Implicit Solvent. The GROMOS 43B1 force field is derived from the 43A1 force field by neutralizing the $\pm 1e$ charges of charged side chains (Arg, Lys, Asp, Glu) and the charged termini of the polypeptide backbone and by reducing their repulsive van der Waals parameters.[29] Accordingly, the dielectric constant was set to 1.0 and the electrostatic cutoff to 100 nm. The POPS parametrization of Fraternali and Cavallo[42] was used for the implicit solvation. POPS parameters were derived specifically for proteins and DNA/RNA molecules using the SASA model given in eq 7 (see below). Initial protein conformations were energy minimized using 500 steps of steepest descent in the presence of the implicit solvent force field term. The system was run for 2.5×10^5 steps of stochastic dynamics using $\gamma_{solv} = 91\ ps^{-1}$, and $\omega_i(t)$ was updated every 100 steps during the simulation.

Trajectory Analysis and Force Matching. Values for the atomic σ_i^{SASA} parameters were derived using eq 11. Explicit solvent forces on each atom along the trajectory were averaged over 10 configurations (10 ps), yielding 50 averaged σ_i^{SASA} parameters per atom per 500 ps trajectory. Area derivatives $(\partial A_i)/(\partial \mathbf{r}_i)$ were calculated for the solute configuration of the explicit water simulation, which guarantees that the implicit and explicit forces are referring to identical atom positions. For atoms with covalently bound polar hydrogen atom(s) (e.g., the OH group), the explicit force on the hydrogen atom(s) was added to that on the main atom, because the hydrogen atoms are not considered in the calculation of the SASA. The resulting force represents the cumulative solvent force on the atom group. Several subsets of atoms were created to select atoms for fitting the parameters of the SASA model: *(i)* ⟨ALL⟩, all atoms; *(ii)* ⟨SA⟩, atoms with A_i within the range of 0.2−0.5 nm^2 ; *(iii)* ⟨SA & θ⟩, with the additional criterion that the implicit-explicit solvent force angle θ lies between 0−45° (hydrophilic) or 135−180° (hydrophobic); *(iv)* ⟨SA & θ+⟩, as before, but exclusively only one of the two angle ranges, i.e. either hydrophilic or hydrophobic. The angle range 0−45° corresponds to 'outward' and 'hydrophilic', and accordingly, the 135−180° range corresponds to 'inward' or 'hydrophobic'.

Data analysis was performed using the R-project software (R Development Core Team[43]). Maximum-likelihood fitting was performed with the function 'fitdistr' of the 'MASS'[44] package. Subsamples of observed force distributions were obtained with the 'sample' function; subsamples of theoretical distributions were generated using the 'rlnorm' function.

Resampling was undertaken to transform a source distribution into the form of a target distribution, specifically the ⟨SA & θ+⟩ distribution into the form of the ⟨SA⟩ distribution. This was achieved by resampling points from the source distribution (⟨SA & θ+⟩) with a suitable probability to reconstruct the density of the target distribution (⟨SA⟩). The probability densities of each data point of ⟨SA & θ+⟩ to be found *(i)* in the ⟨SA⟩ distribution and *(ii)* in the ⟨SA & θ+⟩ distribution were computed using the 'dlnorm' function. The ratio of these probabilities (per data point) was used as a probability vector to resample the source distribution (⟨SA & θ+⟩), which generated the data points of the resampled distribution with a density matching that of the target ⟨SA⟩ distribution.

Partitioning of the Range of σ_i^{SASA} Values into Groups *via* Dynamic Programming. Atom grouping according to the distribution of σ_i^{SASA} parameters of all atom types is a partitioning problem: considering the entire range of σ_i^{SASA} values divided into n bins, one seeks to find the best partition of the bins into k groups, which is equivalent to finding the best locations for $k-1$ dividers. An optimal partitioning for a given number of bins n can be found *via* dynamic programming[45] as sketched below.

The first requirement is the definition of a score by which the obtained partition is judged. We used the Mutual Information between two sets of variables: *(i)* the GROMOS atom types (here '1', '2', ..., '16') and *(ii)* the atom groups that are to be defined (for example 'charged', 'polar', and 'hydrophobic'). The Mutual Information is generally defined as[46]

$$I(X;\ Y) = \sum_{x \varepsilon X} \sum_{y \varepsilon Y} p(x,\ y) \log \frac{p(x,\ y)}{p(x)p(y)} \tag{13}$$

X denotes the set of atom types and x a single atom type, while Y denotes the set of atom groups and y a single atom group. Since each σ_i^{SASA} value is assigned to an atom type, depending on the partitioning to an atom group, the probability $p(x,y)$ is the relative frequency of observing the combination x and y for the given σ_i^{SASA} values, while $p(x)$ and $p(y)$ are the marginal probabilities of these variables.

Table 1. Solvation Parameters $\sigma^{SASA}_{\langle SA\ \&\ \theta+\rangle}$ for Each GROMOS Atom Type[a]

atom type			solvation parameter			
id.	type	description	$\overline{\sigma^{SASA}_{\langle SA\ \&\ \theta+\rangle}}$ (iqr) (kJ mol^{-1} nm^{-2})	sd_{bs}	$n(\sigma^{SASA}_{\langle SA\ \&\ \theta+\rangle})$	resampled $\overline{\sigma^{SASA}_{r\langle SA\ \&\ \theta+\rangle}}$ (iqr) (kJ mol^{-1} nm^{-2})
1	O	carbonyl oxygen (C═O)	−7.2 (5.1)	0.04	11246 [1184750]	−7.5 (4.0)
2	OM	carboxyl oxygen (CO^-)	−21.7 (14.4)	0.1	17346 [316400]	−21.7 (16.6)
3	OA	hydroxyl oxygen (OH)	−7.0 (5.0)	0.1	5942 [156050]	−7.1 (4.9)
5	N	peptide nitrogen (NH)	− (−)		− [1105150]	− (−)
6	NT	terminal nitrogen (NH_2)	−4.0 (3.0)	0.05	3803 [89000]	−3.8 (3.4)
7	NL	terminal nitrogen (NH_3^+)	−26.1 (22.5)	0.1	20384 [79450]	−26.0 (21.6)
8	NR	aromatic nitrogen (−N═)	−4.5 (4.5)	0.1	1589 [64150]	−4.4 (4.3)
9	NZ	Arg amino nitrogen (NH_2^+)	−13.3 (12.9)	0.2	1908 [111500]	−13.5 (13.4)
10	NE	Arg imino nitrogen (NH)	− (−)		− [55750]	− (−)
11	C	bare carbon (C)	− (−)		− [1360950]	− (−)
12	CH1	methine carbon (CH)	3.8 (3.0)	0.03	11347 [1473500]	4.0 (3.1)
13	CH2	methylene carbon (CH_2)	5.0 (4.3)	0.02	55292 [1408200]	4.3 (3.2)
14	CH3	methyl carbon (CH_3)	3.3 (2.9)	0.01	57976 [664450]	3.7 (3.2)
16	CR1	aromatic carbon (−CH═)	4.5 (4.5)	0.04	10793 [644900]	5.1 (5.6)

[a]Atom types 4 (water oxygen) and 15 (CH4) were not included in this parametrization. Atom types with unassigned data (−) were underrepresented in the data subset because of their small SASA values. Numbers in square brackets show the total number of atoms (per atom type) in the reference proteins. The 'resampled' columns show the $\sigma^{SASA}_{r\langle SA\ \&\ \theta+\rangle}$ parameters derived from the resampled force distribution (see text). $\overline{\sigma^{SASA}_{\langle SA\ \&\ \theta+\rangle}}$, median value; (iqr), interquartile range; sd_{bs}, standard deviation of the property in 1000 bootstrap (with replacement) samples; $n(\sigma^{SASA}_{\langle SA\ \&\ \theta+\rangle})$, number of data points.

Table 2. Solvation Parameters σ_g^{SASA} of the Three Atom Groups Derived by Partitioning *via* Dynamic Programming, As Shown in Figure 7[a]

group		atom		solvation parameter		
id.	description	id.	type	$\overline{\sigma_g^{SASA}}$ (iqr) (kJ mol^{-1} nm^{-2})	sd_{bs}	$n(\sigma_g^{SASA})$
1	charged	2, 7, 9	OM, NL, NZ	−23.3 (19.0)	0.1	38325 [507350]
2	polar	1, 3, 6, 8	O, OA, NT, NR	−7.3 (5.9)	0.05	20090 [1493950]
3	hydrophobic	12, 13, 14, 16	CH1, CH2, CH3, CR1	4.1 (3.6)	0.01	143152 [4191050]

[a]$\overline{\sigma_g^{SASA}}$, median value; (iqr): interquartile range; sd_{bs}, standard deviation of the property in 1000 bootstrap (with replacement) samples; n: number of data points. See Table 1 for a description of the atom types.

In the partitioning process, by placing a new divider, the σ_i^{SASA} range is separated into a left and a right part. Assuming that the left part already contains dividers (partitions), the score for placing the new divider is the sum of the Mutual Information of the partitions on the left side (including that of the already existing partitions) and the right side. Systematic variation of the divider position yields the one with the maximal Mutual Information. This procedure is iterated until the positions of all $k-1$ dividers are found.

Without knowing *a priori* the best number of groups k, a safe procedure is to run the partitioning for a range of k-values, here 2 to 20, and to evaluate the results of all runs. The other free parameter is the width of the bins. A width of 1 kJ mol^{-1} nm^{-2} was chosen here. To compare the Mutual Information between the partitionings resulting from different k-values (and therefore from a different number of free parameters), the Mutual Information of each partitioning was normalized by the Joint Entropy, which is defined as[46]

$$H(X, Y) = -\sum_{x\varepsilon X}\sum_{y\varepsilon Y} p(x, y)\log p(x, y) \quad (14)$$

The Joint Entropy is the maximal Mutual Information that could be theoretically achieved given the number of discrete states in the variables X and Y. The ratio $I_{norm} = I/H$ is therefore a measure of the actual versus the maximal information gain.

Error Estimation by Bootstrapping. The statistical basis for the estimation of the σ_i^{SASA} parameters in the previous section is solid with about 10^3−10^4 data points per parameter. However, some σ_i^{SASA} parameter distributions are skewed, and it is instructive to evaluate the variability of the distribution measures (median $\overline{\sigma_i^{SASA}}$, interquartile range $iqr(\sigma_i^{SASA})$) with respect to variation of the input data. This was performed here by bootstrapping, i.e. computation of distribution measures on artificial subsamples of the input data. The 'boot' package[47] of the R-project (R Development Core Team[43]) was used to compute the median and iqr value on 1000 subsamples of the input data of each atom type and atom group. Sampling was performed with replacement. The median and iqr values obtained from the bootstrap procedure were indistinguishable from the ones reported in the σ_i^{SASA} result tables, confirming the robustness of the obtained parameters from variation of the input data; therefore, only standard deviations of the bootstrap values are reported (this refers to Table 1, Table 2, and Supplementary Table S4).

RESULTS

The parametrization of σ_i^{SASA} by force matching is based on a projection of the explicit solvent force onto the implicit force direction (Figure 1 and eq 11). To obtain meaningful solvation parameter estimates, the direction of the explicit force and the implicit force should not be too different and the size of the explicit force should not be very small. Moreover, the solvent-accessible area A_i of an atom should not be very small. These

conditions are not always met as the data analysis presented below reveals.

Atoms having small SASAs or near-orthogonal projection angles behave like 'noise' when deriving σ_i^{SASA} values because they hardly contribute to the implicit force vector. Therefore, we select data that match the characteristics of the SASA model to be included in the calibration set of atoms, i.e. atoms showing a partial exposure to solvent and explicit solvent forces roughly aligned with the direction of the implicit solvation force. Starting from the set of all data ⟨ALL⟩, three data subsets with increasing degree of selectivity were created and tested for the determination of σ_i^{SASA}: ⟨SA⟩ (selection on area range 0.2−0.5 nm^2), ⟨SA & θ⟩ (selection on area range 0.2−0.5 nm^2 and two angle ranges 0−45° (hydrophilic) and 135−180° (hydrophobic)), and ⟨SA & θ+⟩ (selection on area range 0.2−0.5 nm^2 and on angle range 0−45° (hydrophilic) or 135−180° (hydrophobic)). The subset short names will be given in the following as subscripts to indicate the underlying data set. The final parametrization was performed with the ⟨SA & θ+⟩ subset.

Distribution of SASA for Different Atom Types. Before embarking on the examination of solvent forces, it is instructive to view the distribution of SASA per atom type (Figure S1). Charged atoms (NH_3^+ (a), CO^- (b)) show a median exposure of 0.2−0.3 nm^2, while polar atoms (OH (c), C═O (d)) are less exposed, but usually more than the hydrophobic carbon atoms (CH_2 (e), −CH═ (f)). The low exposure of the carbonyl oxygen C═O is caused by its presence in the peptide backbone. More exposed C═O atoms are located in the amido groups of the Asn and Gln side chains (data not shown).

Distribution of the Force on the Solute Atoms Due to the Solvent. The explicit water force density distributions depend on the SASA and polarity of the respective atom type (Figure 2). This and the following plots show typical atoms of the set of GROMOS atom types, with two examples of each charged, polar, and hydrophobic atom types. The color coding illustrates the data subsets: gray denotes the entire data set (<ALL>), orange the selection of atoms within the SASA range 0.2−0.5 nm^2 (⟨SA⟩), and blue the additional selection of force angle in the ranges 0−45° or 135−180° (⟨SA & θ+⟩), where the angle is measured between the direction of the implicit force and the direction of the explicit force.

It is apparent that most atoms are excluded from the selected parametrization sets because of their small SASA. The insets show the finally selected data set ⟨SA & θ+⟩ scaled to 10^4 data points. The shape of all force density distributions is log−normal, as demonstrated by quantile−quantile (Q−Q) plots of a data sample over a random sample from a theoretical log−normal density function with identical mean and sd values as the fitted data (Supplementary Figure S6 − Figure S8). The forces on charged atoms show a median at about 200−300 kJ mol^{-1} nm^{-1}; those on polar atoms are about half as strong, and most hydrophobic atoms experience forces of about 10 kJ mol^{-1} nm^{-1}.

The distributions of angles between the explicit and implicit force are shown in Figure 3. Small angles (close to 0°) indicate hydrophilicity, because the solvent force points in the direction of the area derivative of the implicit force, which points generally toward the solvent, and large angles (close to 180°) are accordingly associated with hydrophobicity. This is clearly visible for the most hydrophilic NH_3^+ (a) and the most hydrophobic CH_2 (e) atom types.

The blue distributions of ⟨SA & θ+⟩ reflect the angle restrictions required in order to fit a SASA-based implicit solvation model to explicit forces.

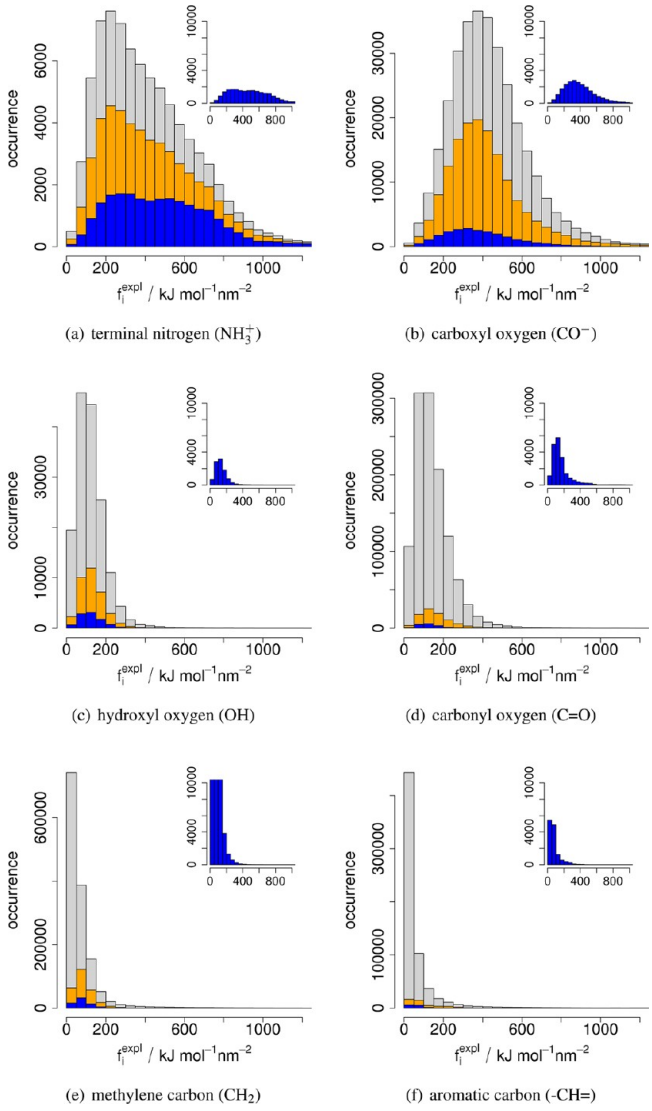


Figure 2. Distribution of the size of the explicit solvent (water) forces $\mathbf{f}_i^{expl}$ for selected GROMOS atom types and for different data sets. The gray distributions show the forces on ⟨ALL⟩ atoms. The overlaid colored distributions show subsets: the ⟨SA⟩ forces in orange, the ⟨SA & θ+⟩ forces in blue. Insets show the ⟨SA & θ+⟩ data in a uniform scale of 10^4 data points for a quantitative comparison of the atom type frequency. The selected atom types are the same as in Figure S1.

Distribution of σ_i^{SASA} Values. The distributions of values of the σ_i^{SASA} parameters derived from the explicit forces in the area range 0.2−0.5 nm^2 (⟨SA⟩) is reasonably constant as a function of the SASA (Figure 4). This is reassuring, because σ_i^{SASA} is assumed to be a constant, independent of the SASA. However, one can observe an increased scatter of σ_i^{SASA} toward low SASA values, as mentioned above. This scatter can be understood from eq 11: as the atom is near complete burial or exposure, small fluctuations in its position lead to large fluctuations in the area derivatives and the derived σ_i^{SASA} values.

$\sigma_{\langle SA\,\&\,\theta+\rangle}^{SASA}$ Parameters from Data Selected To Be Relevant to the SASA Model. The distributions of $\sigma_{\langle SA\,\&\,\theta+\rangle}^{SASA}$ parameters (denoting σ_i^{SASA} parameters derived from the ⟨SA & θ+⟩ subset) are illustrated in Figure 5. Charged atoms show a larger spread of the $\sigma_{\langle SA\,\&\,\theta+\rangle}^{SASA}$ distribution than polar and hydrophobic atoms.

The derived $\sigma_{\langle SA\,\&\,\theta+\rangle}^{SASA}$ values are given in Table 1. The median and interquartile range (iqr) were chosen instead of mean and

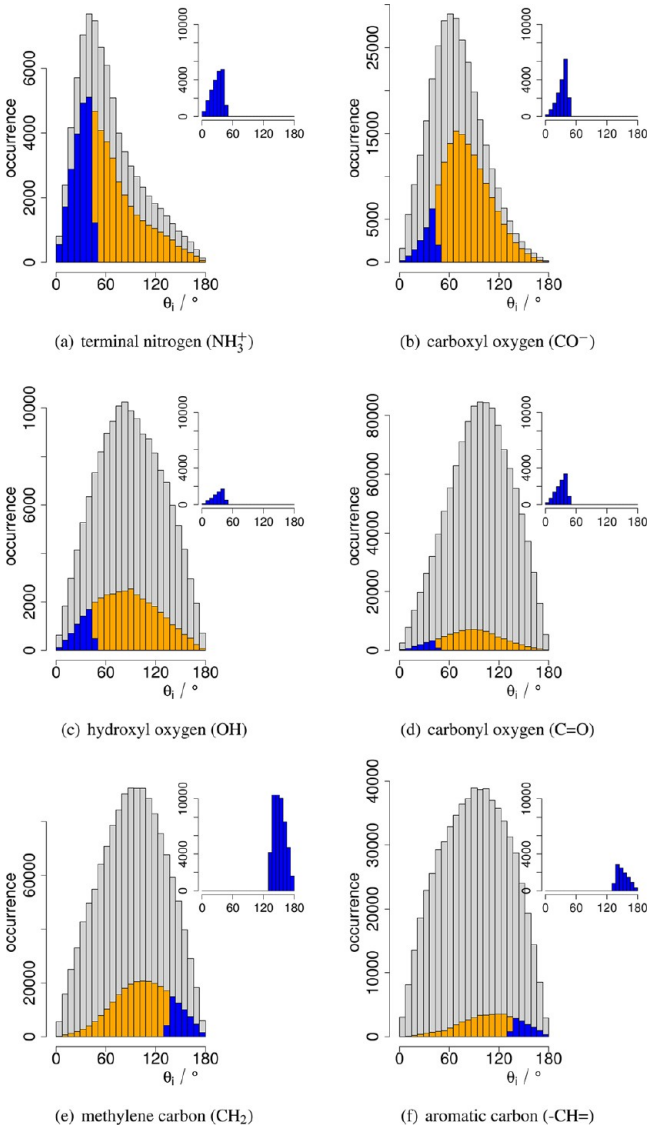


Figure 3. Distribution of the angle θ_i between the explicit and implicit force vectors for selected GROMOS atom types. The atom type selection, data subsets, and color scheme are the same as in Figure 2.

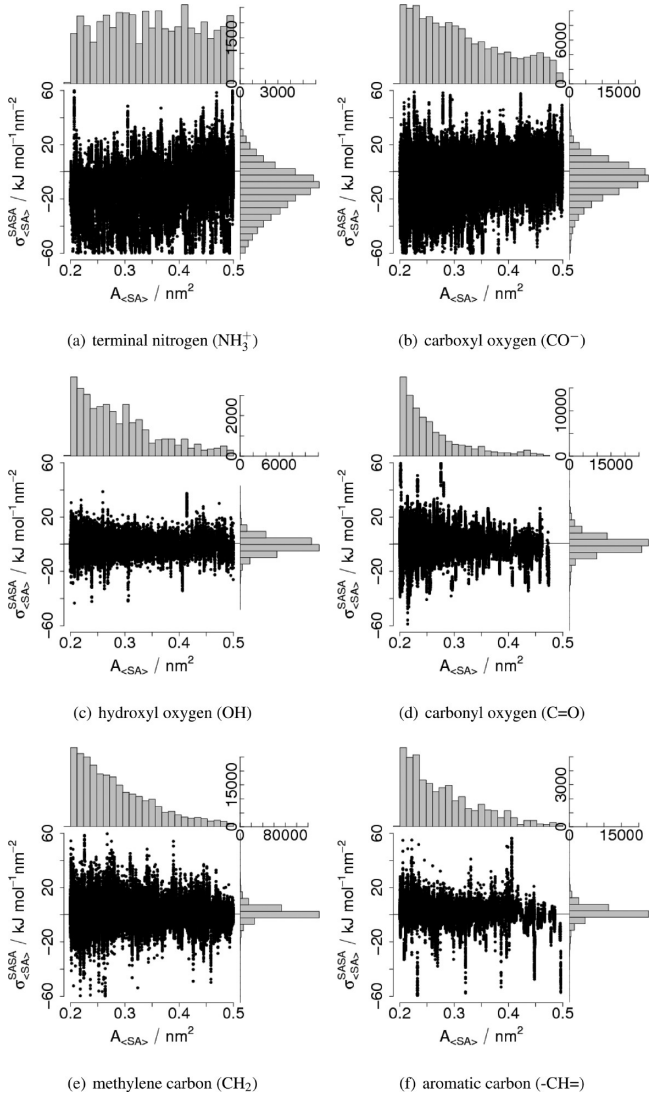


Figure 4. Distribution of the subset $\sigma^{SASA}_{\langle SA\rangle}$ values as function of the atomic SASA value A_i for selected GROMOS atom types.

standard deviation (sd) because of the asymmetry of the $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ distributions (assuming normality, the conversion between iqr and sd is $sd = iqr/1.349$). Despite the relatively large iqr values, the $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ estimates are robust as shown by the small bootstrap 'sd' values, which is due to the sound statistical basis of the data. A graphical illustration of the $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ value distributions is provided by the box plots in Figure 6.

Resampling According to the ⟨SA⟩ Distribution. The intention of the determination of the σ_i^{SASA} parameters *via* eq 11 is a direct transformation of observed solvent forces into a mean force. The selection of forces relevant to the SASA model by requiring a minimum size and projection on the area derivative direction biases the original force distribution. To provide a quantitative estimate of this bias, the distributions of the raw and selected data were approximated with a maximum-likelihood fit using a 'log−normal' density function. We find that the original forces and all subsets closely follow log−normal distributions as shown by the Q−Q-plots in Supplementary Figure S6 − Figure S8, albeit with different mean and spread (Supplementary Table S3).

The $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ values in the 'solvation parameter' columns of Table 1 were derived from the ⟨SA & θ+⟩ subset of explicit water forces of minimal size and with implicit force direction, and it is not clear at this point how this selection on projection angle influences the force distribution and the derived parameters. To reconstruct the ⟨SA⟩ force distribution without compromising the selection on projection angle, the ⟨SA & θ+⟩ distribution was resampled with a probability vector that generated the mean and spread of the ⟨SA⟩ distribution (see Supplementary Figure S9 and Methods).

The 'resampled' values in the right columns of Table 1 were computed from this distribution. The results are very similar for most atom types, the largest deviations being 0.7 (CH2) and 0.6 (CR1). In comparison to the iqr ranges, these deviations are small enough to justify the selection on projection angle as originally performed.

Atom Grouping *via* Dynamic Programming Partitioning. The similarity of the $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ parameters between several atom types suggests a simplified parametrization by partitioning the atom types into atom groups. A good partition should preserve a maximal amount of the information contained in the input data (i.e., the pairing of atom types and $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ values) at a

dx.doi.org/10.1021/ct200390j | J. Chem. Theory Comput. 2012, 8, 2391−2403

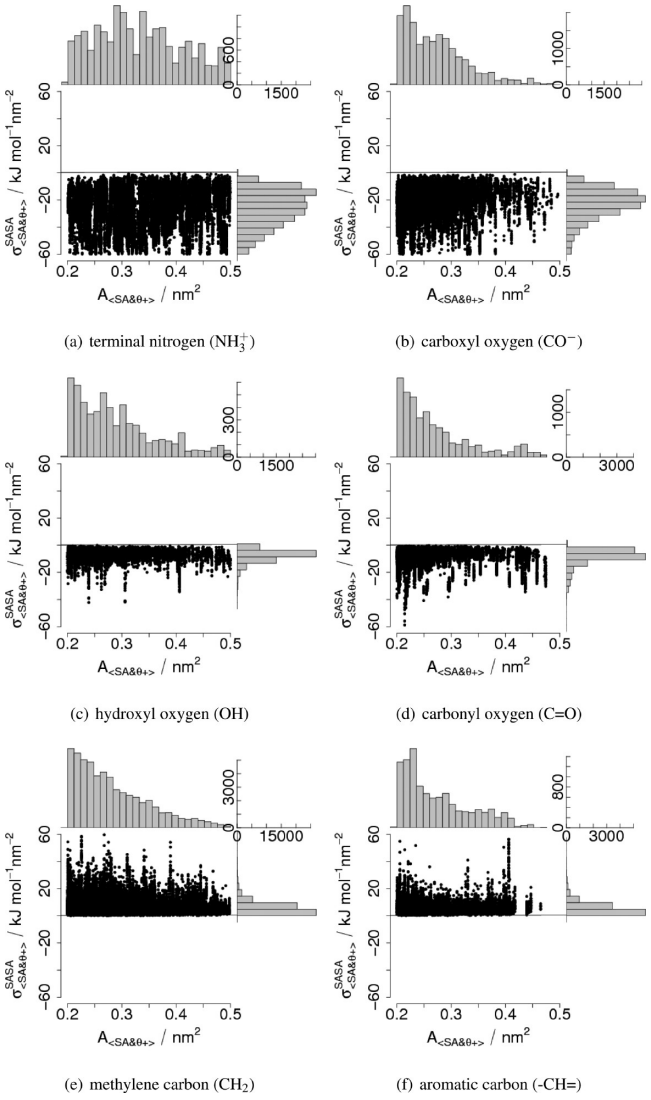


Figure 5. Distribution of the subset $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ values as a function of the atomic SASA value A_i for selected GROMOS atom types.

lower number of parameters (i.e., groups of atom types instead of atom types). In Figure 6, one can easily discern the difference between atom types that are charged (OM, NL, and NZ), polar

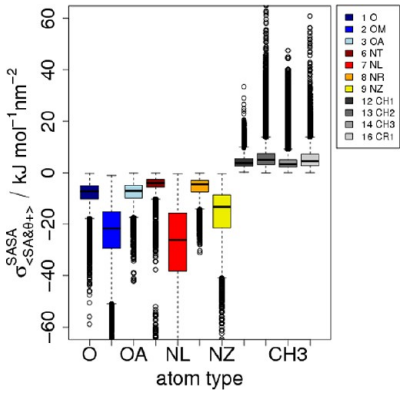


Figure 6. Box plots of the $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ value distributions of the atom types. Distribution features are characterized by symbols; box: 50%, whiskers: 99%, circles: outliers.

(O, OA, NT, and NR), or hydrophobic (CH1, CH2, CH3, and CR1), and intuitively one might choose to combine these into groups. A more systematic and quantitative approach is provided by information theory. An estimate of the fraction of preserved information can be obtained from the ratio of the Mutual Information I(type; group) and the Joint Entropy H(type, group).[48] The Mutual Information increases with the number of groups up to a maximum, at which the two variables (type, group) are equivalent with respect to the attributed σ_i^{SASA} values. The Joint Entropy is the maximally achievable Mutual Information, and it is dependent on the number of groups. Therefore it can be used as a normalization term to compare the Mutual Information between partitions based on different group numbers k. In Supplementary Figure S10, the normalized Mutual Information I_{norm} = I/H is plotted over the number k of groups for partitioning into $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ value bins of width 1 kJ mol^{-1}nm^{-2}. The corresponding entropy values are given in Supplementary Table S5. The curve shows a maximum at which the gain of information is highest in comparison to the theoretically achievable information. Details of the partitioning algorithm and the associated information measures are given in the Methods section.

The results show a maximal I_{norm} for three groups in accordance with the intuitive grouping described above (Figure 7). Table 2 contains the statistics for the resulting σ_g^{SASA} values given to this grouping for force angles 0−45°, and box plots of the distributions are shown in Figure 8.

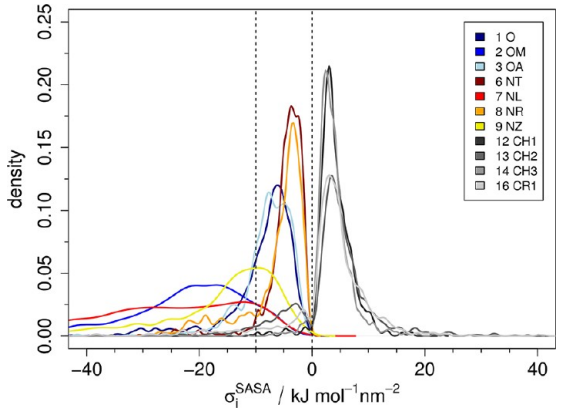


Figure 7. Distribution of $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ parameters of GROMOS atom types, color coded with blue for oxygen atoms, yellow-red for nitrogen atoms, and gray for carbon atoms. The data set contains 5000 sampled (with replacement) data of the $\sigma^{SASA}_{\langle SA\,\&\,\theta+\rangle}$ distribution of each atom type. This data set served as input for the partitioning into atom groups. The two dashed vertical lines show the optimal partitioning obtained *via* dynamic programming.

The hydrophilicity decreases from group 1 to 3, and the spread of the σ_g^{SASA} values is comparable to that for the atom types in Table 2. Overall, the σ_g^{SASA} values of the groups are well separated.

Performance of the Implicit Solvation Model. The performance of the new implicit solvent parametrization described above was evaluated on a test set of six proteins[24] that were not included in the parametrization set. The size of the test proteins ranges from 20 to 189 residues. Simulations of 10 ns length were performed in implicit solvent using the old (Supplementary Table S6) and new (Table 1) parametrization. Reference values were taken from simulations in explicit solvent as described by Allison et al.[24] A comparative summary of relevant protein properties is given in Table 3. The most sensitive

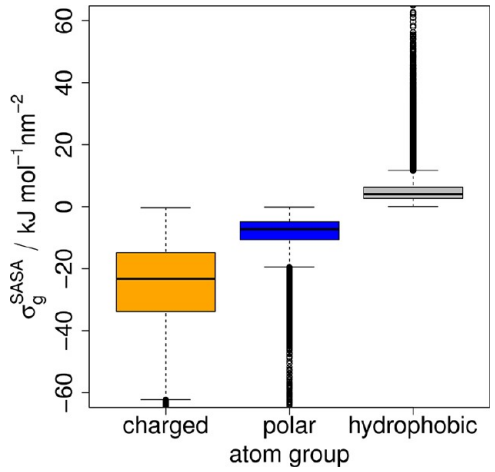


Figure 8. Box plots of the σ_g^{SASA} value distributions of groups of atom types derived by dynamic programming partitioning. Symbols are as in Figure 6.

Table 3. Comparison between 10 ns Simulations in Explicit Solvent (expl), Implicit Solvent Using the New Parametrization Derived Here (impl.n), and Implicit Solvent Using the Old Parametrization (impl.o)[a]

protein	$SASA_{total}$ (nm^2)	$SASA_{phob}$ (nm^2)	$SASA_{phil}$ (nm^2)	R_{gyr} (nm)	rmsd (nm)	rmsf (nm)
trp^{expl}	14.37	8.56	5.81	0.75*	0.34*	0.17*
$trp^{impl.n}$	16.97	10.09	6.88	0.73	0.23	0.12
$trp^{impl.o}$	18.24	10.52	7.73	0.77	0.31	0.12
drk^{expl}	42.55	21.95	20.61	1.11*	0.31*	0.13*
$drk^{impl.n}$	42.25	22.67	19.58	1.06	0.27	0.11
$drk^{impl.o}$	45.88	23.14	22.74	1.07	0.23	0.08
ubq^{expl}	53.27	29.96	23.31	1.21*	0.26*	0.12*
$ubq^{impl.n}$	51.00	27.63	23.37	1.15	0.38	0.17
$ubq^{impl.o}$	55.90	29.26	26.64	1.19	0.28	0.12
$if3c^{expl}$	61.64	35.67	25.97	1.34*	0.18*	0.11*
$if3c^{impl.n}$	64.47	34.34	30.14	1.28	0.27	0.11
$if3c^{impl.o}$	67.87	35.53	32.34	1.30	0.29	0.15
lys^{expl}	68.00	37.18	30.82	1.41*	0.23*	0.15*
$lys^{impl.n}$	68.96	35.50	33.46	1.35	0.29	0.12
$lys^{impl.o}$	77.93	38.73	39.20	1.40	0.33	0.12
$talin^{expl}$	115.24	67.45	47.80	1.92*	0.48*	0.19*
$talin^{impl.n}$	107.53	61.08	46.44	1.90	0.39	0.14
$talin^{impl.o}$	115.04	63.77	51.27	1.91	0.38	0.12

[a]The test proteins (PDB code) are trp (1l2y), drk (2a36), ubq (1ubq), if3c (1tig), lys (1aki), and talin (2jsw). Values marked with an asterisk and trajectories underlying the expl SASA values were taken from Allison et al.[24] R_{gyr}: radius of gyration; rmsd: root mean square deviation from the X-ray or NMR model structure; rmsf: root mean square fluctuation.

and important parameters for the current work are the SASA values. The difference between the parametrizations is discernible in Figure 9. The hydrophilic SASA of the new parametrization (medium-blue solid) is closer to the reference values (dark-blue solid) than the old parametrization (light-blue dashed). The old parametrization leads to an overexposure of hydrophilic atoms. The average deviation from the reference values improved from 19.3% error to 8.5% error. The hydrophobic SASA is nearly unchanged, which is expected from the small difference between the old and new σ_g^{SASA} values. The new parametrization underestimates the hydrophobic SASA

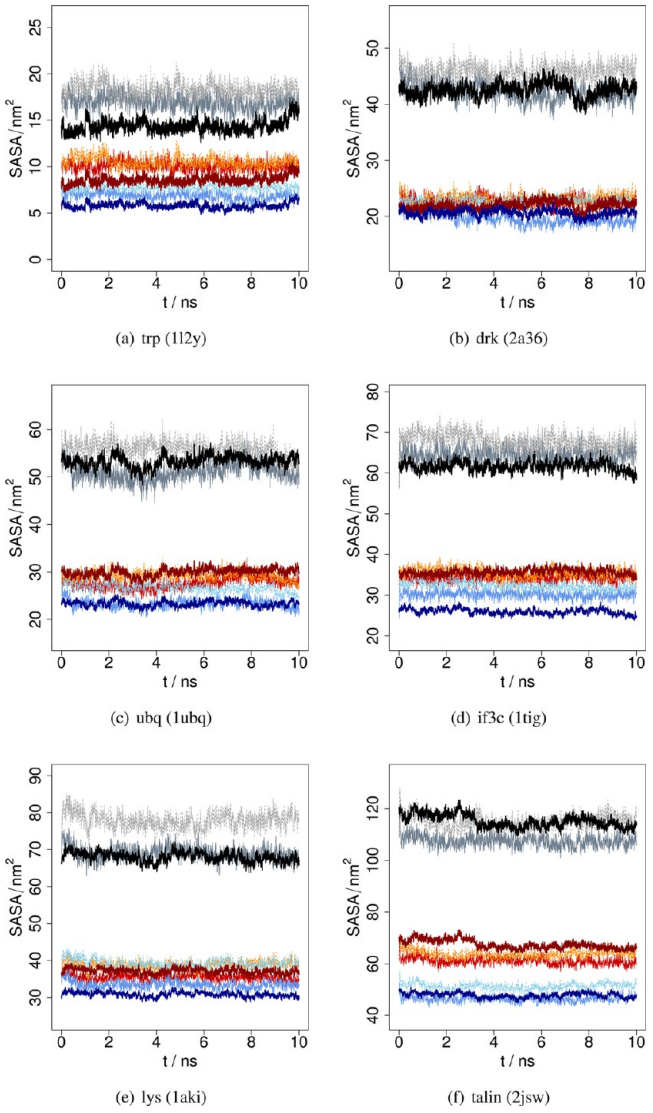
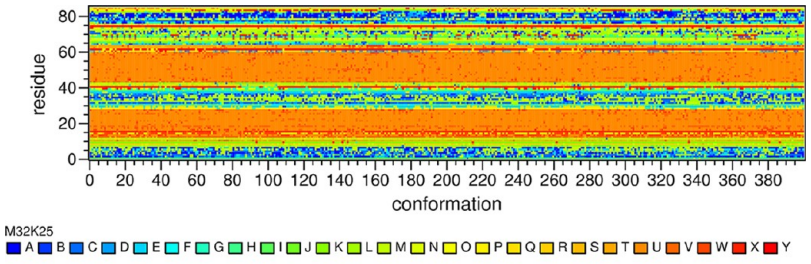


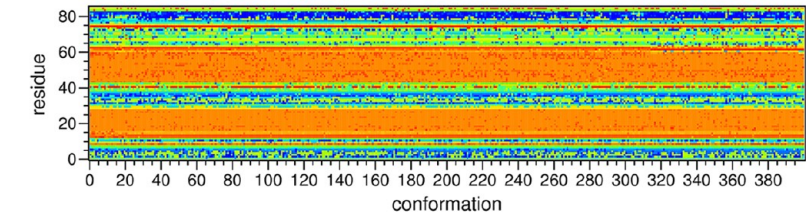
Figure 9. Comparison between the SASA values of six test proteins (abbreviation (PDB code) below figure) obtained under different simulation conditions: explicit water (solid lines, dark colors), implicit solvent using the parametrization derived here (solid lines, medium colors), and implicit solvent using the previous parametrization (dashed lines, light colors). The color code is blue for hydrophilic SASA, red for hydrophobic SASA, and black for total SASA. Conformations were plotted every 25 ps.

of ubq, lys, and talin by about 7%. This difference could become smaller with the addition of a volume term that represents solute−solvent interactions for mostly buried atoms.[24]

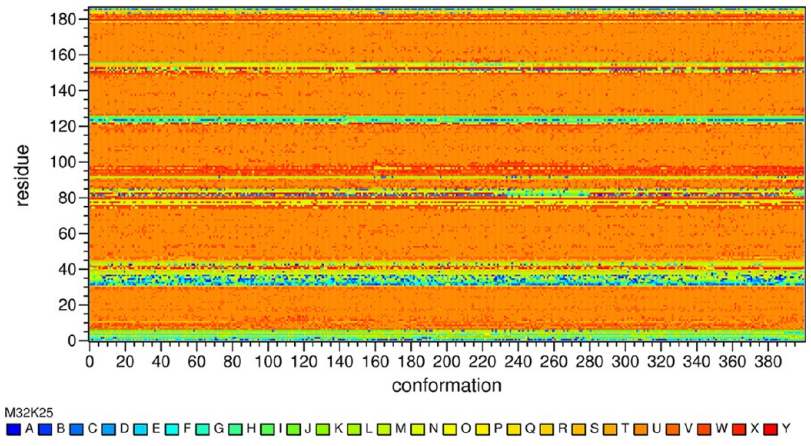
A comparison of local structural properties along the simulations in implicit solvent and in water is shown in Figure 10 for two exemplary test molecules. The data of the remaining four test molecules are shown in the Supporting Information (Figure S11). Local conformations are represented by a structural fragment alphabet (M32K25). This alphabet was derived to describe local conformational states in protein dynamics, and it provides a more comprehensive set of loop and turn states than the conventional secondary structure tools.[49,50] The implicit solvent reproduces the structural properties of the water simulations. Proteins with low conformational fluctuations (Figure 10a,b) show virtually the same profile of structural states. Proteins containing regions with significant



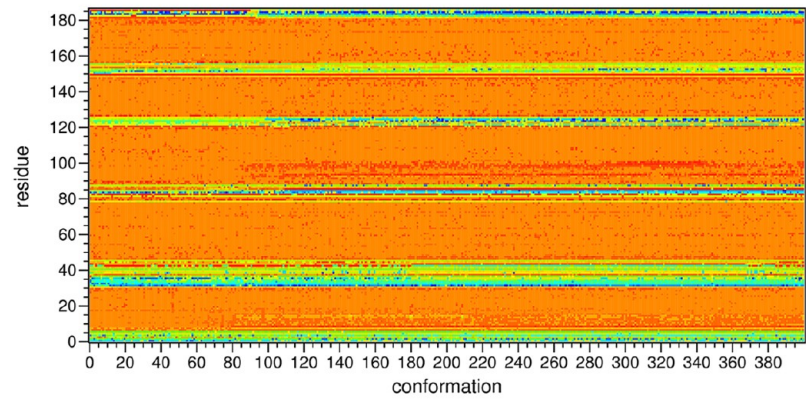
(a) if3c (1tig) in implicit solvent



(b) if3c (1tig) in explicit solvent



(c) talin (2jsw) in implicit solvent



(d) talin (2jsw) in explicit solvent

Figure 10. Local structural properties of the test proteins if3c (a,b) and talin (c,d) in implicit solvent (a,c) and water (b,d). Coloring scheme of conformational states: red, helical (including α-helix); blue: extended (including β-strand); green-yellow: turns and loops. Conformations were plotted every 25 ps.

conformational fluctuations (Figure 10c,d) show slight shifts in the conformational profiles between implicit and explicit solvation, probably as a result of the approximations in the implicit solvation model. However, the range of fluctuations is similar, and the sampled conformations are closely related.

Scope and Limitation of the Parametrization. The SASA model and its parametrization described here is based on several assumptions that have been mentioned throughout the text. Here we summarize these assumptions and discuss their implications for the scope and limitation of the parametrization. From a theoretical point of view, the force matching formula eq

Table 4. Solvation Parameters σ_i^{SASA} for Each GROMOS Atom Type Derived for Angle Ranges $[0°, \theta°]$ and $[(180\text{-}\theta)°, 180°]$[a]

atom type		solvation parameter						
id.	type	$\overline{\sigma_{15}}$	$\overline{\sigma_{30}}$	$\overline{\sigma_{45}}$	$\overline{\sigma_{60}}$	$\overline{\sigma_{75}}$	$\overline{\sigma_{90}}$	$\overline{\sigma_{180}}$
1	O	−7.7 (4.9)	−7.7 (5.3)	−7.2 (5.1)	−6.4 (4.7)	−5.2 (4.4)	−3.8 (4.7)	−0.5 (7.5)
2	OM	−22.4 (15.4)	−23.4 (15.7)	−21.7 (14.4)	−18.4 (13.5)	−14.1 (12.6)	−10.3 (13.1)	−4.6 (16.7)
3	OA	−7.9 (5.2)	−7.6 (5.2)	−7.0 (5.0)	−6.1 (4.6)	−5.0 (4.3)	−3.8 (4.7)	−0.8 (7.0)
5	N	− (−)	− (−)	− (−)	− (−)	− (−)	− (−)	− (−)
6	NT	−4.0 (2.8)	−4.1 (2.9)	−4.0 (3.0)	−3.6 (2.9)	−2.9 (2.6)	−2.0 (2.7)	4.6 (4.2)
7	NL	−28.6 (27.2)	−28.7 (24.9)	−26.1 (22.5)	−22.4 (20.7)	−19.1 (20.0)	−16.9 (20.6)	−12.6 (23.0)
8	NR	−5.1 (4.9)	−4.7 (4.8)	−4.5 (4.5)	−4.1 (4.5)	−3.4 (4.1)	−2.7 (3.9)	−0.5 (5.3)
9	NZ	−17.7 (15.0)	−18.9 (20.9)	−13.3 (12.9)	−10.3 (7.5)	−7.1 (6.0)	−4.0 (6.0)	2.1 (8.5)
10	NE	− (−)	− (−)	− (−)	− (−)	− (−)	− (−)	− (−)
11	C	− (−)	− (−)	− (−)	− (−)	− (−)	− (−)	− (−)
12	CH1	4.6 (3.3)	4.2 (3.1)	3.8 (3.0)	3.3 (2.8)	2.8 (2.8)	2.5 (2.9)	2.1 (3.1)
13	CH2	5.8 (4.8)	5.4 (4.5)	5.0 (4.3)	4.3 (3.9)	3.5 (3.7)	2.7 (3.7)	1.2 (4.3)
14	CH3	3.9 (3.2)	3.0 (3.0)	3.3 (2.9)	2.9 (2.6)	2.5 (2.5)	2.2 (2.6)	1.9 (2.7)
16	CR1	5.0 (4.5)	4.9 (4.6)	4.5 (4.5)	4.1 (4.3)	3.5 (4.0)	2.8 (3.8)	1.6 (4.4)

[a]Atom types 4 (water oxygen) and 15 (CH4) were not included in this parameterization. Atom types with unassigned data (−) are underrepresented in the selected data. $\overline{\sigma}$, median value of $\sigma^{SASA}_{\langle SA\ \&\ \theta+\rangle}$ in units kJ mol^{-1} nm^{-2}, θ value as subscript; the error is given as interquartile range in parentheses (assuming normality, the conversion between iqr and sd is $sd = iqr/1.349$).

11 is applicable to all recorded data pairs $\partial A_i/\partial \mathbf{r}_i$ and $\langle \mathbf{f}_i^{expl} \rangle$ and the determined σ_i^{SASA} parameters would represent all cumulative solvent forces on atom i. In practice, forces on atoms with less than 0.2 nm^2 SASA are very noisy, and therefore they were excluded. Using all remaining explicit forces leads to unphysical σ_i^{SASA} parameters, as shown in the last column of Table 4 for θ = 180°. At this angle, polar atom types 1, 3, and 8 appear to be barely hydrophilic; parameters of atom types 6 and 9 become incorrectly hydrophobic. Therefore, as described in Methods, the parametrization in Table 1 was performed using a data subset of the explicit forces that matches the SASA model. In this subset, explicit forces point in the same direction as the implicit forces. This selection on the force angle introduces a bias on the resulting σ_i^{SASA} parameters as shown in Table 4 for the angle range 15−90° in 15° increments. As expected, σ_i^{SASA} values decrease toward larger angles, but from 15−60° the variation lies within the iqr range. The σ_i^{SASA} values of the angle range 75−90° become progressively similar to the unphysical values of the last column (180°). We chose the σ_i^{SASA} values resulting from θ = 45° as the preferred parametrization, because they represent a reasonable middle course between including all data and selecting data that match best the framework of the SASA model.

A feature of the SASA model is the independence of σ_i^{SASA} from the atomic SASA value (eq 6). However, charged atoms (Figure 4a,b) show a positive correlation betweeen σ_i^{SASA} and SASA values. This correlation became apparent in this study through the use of explicit solvent forces on individual atom types. Since this study is based on the eq 6, this correlation is necessarily neglected, and the range of σ_i^{SASA} values is represented by an averaged σ_i^{SASA} value. This approximation could be remedied by an additional term that complements the SASA derivative of eq 6 with a term that is proportional to the atomic SASA. Exploration of such an extended SASA model will be performed in future studies.

DISCUSSION

A method for the determination of the parameters of an implicit solvation model has been proposed. It is based on a particular form of force matching: the mean solvation force on an atom in explicit solvent is projected onto the direction of the implicit force, as given by the derivative of the solvent-accessible surface area term of the solute potential energy function. This allows for the extraction of the solvation parameters σ_i^{SASA} of the implicit solvation model directly from the observed solvation forces in MD simulations using explicit solvation for a set of proteins in their native structure.

In the original description of the implicit SASA model,[27] the implicit solvation parameters σ_i^{SASA} were estimated by comparing the preservation of characteristic geometric properties of proteins between MD simulations with explicit and implicit solvation. The resulting values for σ_i^{SASA} were −25 kJ mol^{-1} nm^{-2} for hydrophilic atoms and 5 kJ mol^{-1} nm^{-2} for hydrophobic atoms (see Supplementary Table S6). Here we used MD simulations of 188 protein domains with diverse topology and applied a force matching formula to derive the σ_i^{SASA} parameter values directly from the simulations of the rigid proteins in implicit and explicit water. Analysis of the solvent forces revealed that only a fraction of the explicit solvent forces are suitable for the parametrization of the SASA model: those that act on atoms having a surface exposure over 0.2 nm^2 and that are roughly aligned with the implicit force. Therefore it was imperative to start from a large data set to arrive at a final data set of sufficient statistical weight. While the selection based on a sizable exposed area can be viewed as an intrinsic part of a SASA-based model, it is less obvious how selection based on a small angle between explicit and implicit forces modifies the explicit solvent force distribution. We showed that the latter follows a log−normal function like all other solvent force distributions explored here and that the derived σ_i^{SASA} parameters would be the same if the distribution was that of the SASA-only selection.

The derived σ_i^{SASA} values show a good correspondence between the original and current parametrization. Charged atoms adopt σ_i^{SASA} values of −26.1 (NH_3^+) and −13.3 (Arg NH_2^+) kJ mol^{-1} nm^{-2}, hydrophobic atoms values between 3.3 (CH_3) and 5.0 (CH_2) kJ mol^{-1} nm^{-2}. Additionally we observed a group of polar atom types (e.g., OH and C=O) that adopt intermediate values between −7.0 (OH) and −4.0 (NH_2) kJ mol^{-1} nm^{-2}. We partitioned the resulting σ_i^{SASA} distributions of the 16 GROMOS atom types into three groups using dynamic programming. The partitioning maximizes the Mutual Information between the (discretized) σ_i^{SASA} distributions and the assigned group labels. The resulting three groups can be labeled

as 'charged' (σ_g^{SASA} = −23.3 kJ mol^{-1} nm^{-2}), 'polar' (σ_g^{SASA} = −7.3 kJ mol^{-1} nm^{-2}), and 'hydrophobic' (σ_g^{SASA} = 4.1 kJ mol^{-1} nm^{-2}).

The current parameter values were tested on a set of 6 proteins for which data of long and independent (from this study) simulations were available. The new parameters improve the hydrophilic SASA, indicating that the average solvent forces on hydrophilic atoms are well reproduced by the implicit solvent model.

■ ASSOCIATED CONTENT

Supporting Information

Distribution of atomic SASA values A_i for all GROMOS atom types included in this study (Figure S1). Distribution of the size of the explicit solvent (water) forces $\mathbf{f}_i^{expl}$ for the GROMOS atom types and for different data sets (Figure S2). Distribution of the angle θ_i between the explicit and implicit force vectors for the GROMOS atom types (Figure S3). Distribution of the subset $\sigma_{\langle SA\rangle}^{SASA}$ values as a function of the atomic SASA value A_i for the GROMOS atom types (Figure S5). Q−Q plots showing the log-normal behavior of ⟨ALL⟩ water forces (Figure S6). Q−Q plots of water forces of the ⟨SA⟩ subset (Figure S7). Q−Q plots of the ⟨SA & θ+⟩ water forces (Figure S8). Q−Q plots of resampled ⟨SA & θ+⟩ forces over ⟨SA⟩ forces (Figure S9). Ratio I_{norm} of the Mutual Information and the Joint Entropy between the GROMOS atom type classification and the atom groups derived by partitioning (Figure S10). Local structural properties of the test proteins trp (a,b), drk (c,d), ubq (e,f), and lys (g,h) (Figure S11). Topological alphabet defined by the supersecondary structure of two successive α or β elements and the angle between their central axis (Table S1). Characteristics of the protein domains used in this study for the implicit solvent parametrization (Table S2). Parameters of central tendency of log-likelihood log-normal fits of various explicit solvent force distributions (Table S3). Solvation parameters σ_g^{SASA} of atom groups derived by partitioning *via* dynamic programming (Table S4). Mutual Information (I), Joint Entropy (H), and the normalized Mutual Information I_{norm} = I/H of the partitioning *via* dynamic programming into k groups for 1 kJ mol^{-1} nm^{-2} binning of the σ_i value range (Table S5). Original implicit solvation parameters[27] for GROMOS atom types[29] (Table S6). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: franca.fraternali@kcl.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.K. acknowledges support by the MRC National Institute for Medical Research (U117581331). F.F. and J.K. acknowledge visiting professorships to the van Gunsteren lab at the ETH Zürich in May-June 2011.

■ REFERENCES

(1) Schiffer, C. A.; Dötsch, V. *Curr. Opin. Biotechnol.* **1996**, *7*, 428−432.
(2) De Simone, A.; Dodson, G. G.; Verma, C. S.; Zagari, A.; Fraternali, F. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7535−7540.
(3) De Simone, A.; Spadaccini, R.; Temussi, P. A.; Fraternali, F. *Biophys. J.* **2006**, *90*, 3052−3061.
(4) Autore, F.; Bergeron, J. R. C.; Malim, M. H.; Fraternali, F.; Huthoff, H. *PLoS One* **2010**, *5*, e11515.
(5) Kleinjung, J.; Bayley, P. M.; Fraternali, F. *FEBS Lett.* **2000**, *470*, 257−262.
(6) Kleinjung, J.; Fraternali, F.; Martin, S. R.; Bayley, P. M. *Proteins* **2003**, *50*, 648−656.
(7) Gaudreault, M.; Viñals, J. *Phys. Rev. E* **2009**, *80*, 021916.
(8) Arnautova, Y. A.; Vorobjev, Y. N.; Vila, J. A.; Scheraga, H. A. *Proteins* **2009**, *77*, 38−51.
(9) Lazaridis, T.; Karplus, M. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139−145.
(10) am Busch, M. S.; Lopes, A.; Amara, N.; Bathelt, C.; Simonson, T. *BMC Bioinf.* **2008**, *9*, 148.
(11) Lazaridis, T. *Proteins* **2005**, *58*, 518−527.
(12) Ulmschneider, M. B.; Ulmschneider, J. P.; Sansom, M. S. P.; Di Nola, A. *Biophys. J.* **2007**, *92*, 2338−2349.
(13) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(14) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.
(15) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712−3720.
(16) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479−499.
(17) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265−284.
(18) Shimizu, S.; Chan, H. S. *Proteins* **2002**, *48*, 15−30.
(19) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.
(20) Prabhu, N. V.; Zhu, P.; Sharp, K. A. *J. Comput. Chem.* **2004**, *25*, 2049−2064.
(21) Wagoner, J.; Baker, N. A. *J. Comput. Chem.* **2004**, *25*, 1623−1629.
(22) Vitalis, A.; Pappu, R. V. *J. Comput. Chem.* **2009**, *30*, 673−699.
(23) Chen, J.; Brooks, C. L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471−481.
(24) Allison, J. R.; Boguslawski, K.; Fraternali, F.; van Gunsteren, W. F. *J. Phys. Chem. B* **2011**, *115*, 4547−4557.
(25) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523−9530.
(26) Wagoner, J.; Baker, N. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8331−8336.
(27) Fraternali, F.; van Gunsteren, W. F. *J. Mol. Biol.* **1996**, *256*, 939−948.
(28) Hasel, W.; Hendrickson, T.; Still, W. C. *Tetrahedron Comput. Methodol.* **1988**, *1*, 103−116.
(29) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.: 1996.
(30) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719−1751.
(31) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24−33.
(32) Yun-yu, S.; Lu, W.; van Gunsteren, W. F. *Mol. Simul.* **1988**, *1*, 369−383.
(33) van Gunsteren, W. F.; Berendsen, H. J. C. *Mol. Simul.* **1988**, *1*, 173−185.
(34) Daura, X.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **1998**, *19*, 535−547.
(35) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*, 1st ed.; Pullman, B., Ed.; Reidel; Dordrecht, 1981; Chapter Interaction models for water in relation to protein hydration, pp 331−342.
(36) Martin, A. C. *Protein Eng.* **2000**, *13*, 829−837.
(37) Kamat, A. P.; Lesk, A. M. *Proteins* **2007**, *66*, 869−876.
(38) Brenner, S. E.; Koehl, P.; Levitt, M. *Nucleic Acids Res.* **2000**, *28*, 254−256.
(39) Pandini, A.; Bonati, L.; Fraternali, F.; Kleinjung, J. *Bioinformatics* **2007**, *23*, 515−516.
(40) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(41) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Chem. Phys.* **1977**, *23*, 327−341.
(42) Fraternali, F.; Cavallo, L. *Nucleic Acids Res.* **2002**, *30*, 2950−2960.
(43) R Development Core Team, R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2009.
(44) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*, 4th ed.; Springer: 2002.
(45) Skiena, S. S. *The Algorithm Design Manual*, 2nd ed.; Springer: 2008; Chapter 8.5, pp 294−298.
(46) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley Series in Telecommunication; Wiley: 1991.
(47) Canty, A. S. An S-Plus Library for Resampling Methods. 30th Symposium on the Interface: Computing Science and Statistics, 1998.
(48) Martin, L. C.; Gloor, G. B.; Dunn, S. D.; Wahl, L. M. *Bioinformatics* **2005**, *21*, 4116−4124.
(49) Pandini, A.; Fornili, A.; Kleinjung, J. *BMC Bioinf.* **2010**, *11*, 97.
(50) Pandini, A.; Fornili, A.; Fraternali, F.; Kleinjung, J. *FASEB J.* **2012**, *26*, 868−881.