JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0

Alpeshkumar K. Malde,*[,†] Le Zuo,[†] Matthew Breeze,[†] Martin Stroet,[†] David Poger,[†] Pramod C. Nair,[†] Chris Oostenbrink,[‡,§] and Alan E. Mark*[,†,∥]

[†]School of Chemistry and Molecular Biosciences, University of Queensland, St. Lucia, Australia

[‡]Leiden/Amsterdam Center for Drug Research, Division of Molecular Toxicology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[§]Institute of Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Vienna, Austria

[∥]Institute for Molecular Bioscience, University of Queensland, St. Lucia, QLD 4072, Australia

S *Supporting Information*

ABSTRACT: The Automated force field Topology Builder (ATB, http://compbio.biosci.uq.edu.au/atb) is a Web-accessible server that can provide topologies and parameters for a wide range of molecules appropriate for use in molecular simulations, computational drug design, and X-ray refinement. The ATB has three primary functions: (1) to act as a repository for molecules that have been parametrized as part of the GROMOS family of force fields, (2) to act as a repository for pre-equilibrated systems for use as starting configurations in molecular dynamics simulations (solvent mixtures, lipid systems pre-equilibrated to adopt a specific phase, etc.), and (3) to generate force field descriptions of novel molecules compatible with the GROMOS family of force fields in a variety of formats (GROMOS, GROMACS, and CNS). Force field descriptions of novel molecules are derived using a multistep process in which results from quantum mechanical (QM) calculations are combined with a knowledge-based approach to ensure compatibility (as far as possible) with a specific parameter set of the GROMOS force field. The ATB has several unique features: (1) It requires that the user stipulate the protonation and tautomeric states of the molecule. (2) The symmetry of the molecule is analyzed to ensure that equivalent atoms are assigned identical parameters. (3) Charge groups are assigned automatically. (4) Where the assignment of a given parameter is ambiguous, a range of possible alternatives is provided. The ATB also provides several validation tools to assist the user to assess the degree to which the topology generated may be appropriate for a given task. In addition to detailing the steps involved in generating a force field topology compatible with a specific GROMOS parameter set (GROMOS 53A6), the challenges involved in the automatic generation of force field parameters for atomic simulations in general are discussed.

INTRODUCTION

Computer simulations are widely used to gain insight into dynamic molecular processes at near atomic resolution. In particular, molecular dynamics (MD) methods can be used to model the time evolution of biomolecular systems (proteins, lipids, nucleic acids, and carbohydrates) in order to study dynamic processes such as protein folding or to sample a Boltzmann-weighted ensemble to estimate thermodynamic properties such as ligand binding free energies. The primary challenge in such simulations is to describe the properties of the system in terms of the interactions between atoms. Given the size and complexity of biomolecular systems together with the time scales that must be reached to model processes of interest, the use of classical mechanics in conjunction with a so-called *effective* force field is normally the method of choice. In such MD simulations, the interactions between the various particles in a given system are represented by an empirical energy function, parametrized to reproduce a range of structural, energetic, or thermodynamic properties of model compounds derived from experimental and/or high level quantum mechanical calculations.

A variety of such empirical force fields have been developed for use in biomolecular simulations. These include the GROMOS,[1−5] AMBER,[6,7] CHARMM,[8,9] and OPLS[10,11] force fields. Although the form of the potential energy function used in each of these force fields is very similar, the parameters that are used to describe specific molecules can differ significantly. This in part reflects different parametrization philosophies. However, it also reflects the fact that the range of data that can be used during the parametrization is limited. As a result, parametrization is an underdetermined problem. Force field development is also further complicated by the fact that the parameters themselves are highly correlated. Common force fields such as GROMOS,[1−5] AMBER,[6,7] CHARMM,[8,9] and OPLS[10,11] provide a set of internally consistent parameters for a core set of molecules such as amino acids, nucleotides, simple sugars, common lipids, and common solvents. These are not general force fields per se but have instead been specifically parametrized to reproduce a given set of properties of this core set of molecules. As a consequence, each new heteromolecule such as a substrate, inhibitor, cofactor, or drug molecule must be parametrized individually. A number of more general force fields intended for the description of a wide range of small molecules have also been developed. Examples include MM2,[12] MM3,[13] MM4,[14,15] and MMFF94.[16−18] Such force fields employ more complex potential energy functions expressed in terms of atomic properties and, as such, are

Received: March 22, 2011
Published: October 27, 2011

ACS Publications © 2011 American Chemical Society

incompatible with most of the biomolecular force fields in common use.

The most basic approach that can be used to develop a force field description for a novel molecule compatible with one of the major biomolecular force fields is to manually assign parameters to groups of atoms based on their similarity to groups of atoms in molecules that have already been specifically parametrized for that force field. This is, however, tedious, time-consuming, and error-prone. Also, to be done appropriately, a detailed knowledge of the philosophy underlying a given force field is required. As a consequence, various tools have been developed to facilitate the generation of the force field descriptions for novel molecules for a range of force fields. A number of these generate simplified topologies and parameters that can be used to maintain a particular geometry of a ligand molecule for use in X-ray refinement. For example, the program eLBOW[19] (electronic Ligand Builder and Optimization Workbench), a module of the PHENIX[20] suite of programs, generates a topology, parameters, and geometric restraints for ligand molecules based on the results of semiempirical quantum mechanical (AM1[21]) calculations. It uses a rule-based approach to identify all possible bonds, bond angles, dihedral angles, planar rings, and chiral centers in a molecule. Hydrogen atoms are added and semiempirical QM calculations performed to optimize the geometry of the molecule in a vacuum. The optimized geometry is then used to derive a set of purely geometric parameters for X-ray refinement in the CCP4[22] monomer library format. Note that this format does not contain information regarding the force constants used to impose these geometric restraints. The program LIBCHECK,[23] distributed as part of the CCP4[22] package, has a database of molecules that can be searched for known ligands based on element type and bonding patterns. In cases when no appropriate ligand is found, the parameters are generated using a rule-based approach using the geometry supplied by the user. The new molecule is then added to the database. LIBCHECK is also used by the refinement programs Refmac[24] and Coot.[25] XPLO2D, part of the X-PLOR and CNS[26,27] packages, generates a highly simplified topology for use with the program CNS based on a set of coordinates supplied by the user. Again, a set of rules is used to identify atom type, bonds, bond angles, etc., which are combined with a small set of default values for the force constants to generate a set of geometric restraints. For example, 4184 kJ mol^{-1} Å^{-2} is used for all bond lengths, 2092 kJ mol^{-1} rad^{-2} for all bond angles, and 3140 kJ mol^{-1} for all proper and improper dihedral angles of constrained systems and 0 kJ mol^{-1} in all other cases. The HIC-Up[28] (Heterocompound Information Centre—Uppsala, http://xray.bmc.uu.se/hicup/) Web server provides parameter and topology files generated using XPLO2D for all heteromolecules found in the PDB and is updated once or twice a year. The program Hess2FF[29] can also be used to generate a topology and geometric constraints in CNS format. Hess2FF uses a Hessian (force constant matrix) of a molecule as input, which can be obtained from molecular mechanics, semiempirical or quantum mechanical calculations. The molecular descriptions generated by the tools listed above are, however, intended for use in X-ray refinement and are not suitable for molecular simulations. Many do not provide terms to model the nonbonded atoms (electrostatic and van der Waals interactions) or only provide parameters for heavy, non-hydrogen atoms. While these simplified descriptions are widely used, it is increasingly apparent that errors in the description of ligand molecules, often linked to the use of these very simplified representations, lead to errors and uncertainties in structural databases such as the Protein Data Bank (PDB). This potentially has severe implications for structure-based drug design, interpretation of biochemical mechanisms, and validation of virtual screening.[30]

A number of programs and Web servers have been developed to facilitate the generation of force field descriptions of novel molecules for use in MD simulations. MKTOP[31] (http://labmm.iq.ufrj.br/mktop/) implements a rule-based approach to generate an OPLS-AA like force field in a GROMACS-readable format. A distance criterion of 0.3 Å is used to identify bonded partners. This information is used subsequently to assign atom types based on chemical environment and to derive other bonded parameters. The program does not assign partial charges to the atoms. Instead, these must be supplied by the user together with a set of coordinates for atoms in the ligand. Antechamber[32] (http://ambermd.org/antechamber/antechamber.html) and YASARA AutoSMILES Server (http://www.yasara.org/autosmilesserver.htm) generate topologies compatible with the GAFF[7] (Generalized AMBER force field) based on a set of coordinates supplied by the user. In the program Antechamber, first the atom types and atomic charges are assigned. If the user supplies a QM output file containing the electrostatic potential (ESP), then the partial atomic charges are assigned on the basis of the RESP[33] method; otherwise AM1-BCC[34] [Bond Charge Correction (BCC) applied to the AM1 atomic charges] charges are assigned. The bonded parameters are assigned on the basis of the GAFF parameters subsequently. In the YASARA server, an AM1 calculation is performed to optimize the geometry of the ligand. Initially, AM1-BCC parameters (including charges) are assigned to the molecule. However, if fragments within the molecule match fragments within a standard GAFF database, the charges are replaced by RESP charges, and the bonded parameters replaced by standard GAFF parameters. GENRTF[35] (http://a.cmm.ki.si/genrtf/) generates parameters and a topology compatible with the CHARMM force field using a rule-based approach. SwissParam (http://swissparam.ch/) generates a force field description for small organic molecules by combining bonded parameters extracted from the Merck Molecular Force Field (MMFF) and nonbonded terms from the CHARMM22 force field. Paramchem server (https://www.paramchem.org) also aims to generate parameters consistent with the CHARMM force field. PRODRG[36] (http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrg/prodrg.cgi) is a Web server which generates force field descriptions of ligand molecules based roughly on the GROMOS 43A1 force field. The molecular topologies generated by PRODRG are intended for use in X-ray refinement, docking, and MD simulations. Nevertheless, PRODRG has a number of serious limitations, including that the ligand protonation states are assigned automatically by PRODRG, the assignment of critical 1−4 exclusions is inappropriate in some cases, and atomic charges and charge groups are not assigned in a manner consistent with the GROMOS force field.[37] In fact, it is not possible to assign the protonation/tautomeric state and/or the overall charge of the molecule using many of the tools and Web servers described above. Although the YASARA AutoSMILES Server does consider the pH when adding hydrogen atoms to a given structure, the tautomeric state cannot be assigned.

Here, we describe the Automated force field Topology Builder (ATB, http://compbio.biosci.uq.edu.au/atb) and repository. The ATB is a Web-accessible server that can provide topologies and parameters for a wide range of molecules compatible with the GROMOS family of force fields. The topologies and parameters provided are designed for use in molecular simulations, computational drug design, and X-ray refinement. The ATB has three primary functions: first, to act as a repository for molecules that have been parametrized as part of the GROMOS family of force fields; second, to act as a repository for pre-equilibrated systems for use as starting configurations in molecular dynamics simulations (solvent mixtures, lipid systems pre-equilibrated to adopt a specific phase, etc.); and third, to generate force field descriptions of novel molecules compatible with the GROMOS force field in a variety of formats. This is done in a multistep process in which results from a series of quantum mechanical (QM) calculations are combined with a knowledge-based approach to ensure compatibility with the GROMOS family of force fields. The ATB differs from other topology builders in that it requires the user to stipulate the protonation and tautomeric states of the molecule. The symmetry of the molecule is also analyzed to ensure equivalent atoms are assigned identical parameters irrespective of molecular geometry. Charge groups, which are used in the GROMOS force field to ensure that compensating groups of charges are always considered simultaneously, thus reducing artifacts in the calculation of the long-range electrostatic interactions, are assigned automatically. Importantly, in cases where the assignment of a given parameter is ambiguous, a range of possible alternatives is provided. Finally, the ATB provides several validation tools to assist the user to assess the degree to which the topology generated may be appropriate for a given task.

The remaining sections of the manuscript are organized as follows. First, the challenges associated with the automatic generation of molecular force fields are discussed briefly. Then, the basic structure of the ATB pipeline is presented. This is followed by a discussion on the limitations of the current version of the ATB along with the steps taken to validate the final topologies.

■ CHALLENGES IN THE AUTOMATIC GENERATION OF MOLECULAR FORCE FIELDS

The empirical potential energy functions used in biomolecular force fields such as GROMOS,[1−5] AMBER,[6,7] CHARMM,[8,9] and OPLS[10,11] are crude in the sense that they attempt to represent the potential energy surfaces of a wide range of molecules using a very limited set of parameters. Differences in chemical environment are encoded by assigning different sets of van der Waals interaction parameters for a given atom type depending on the neighboring atoms. Pair interactions are based on simple combination rules. Electrostatic interactions are modeled by assigning fixed partial charges to atoms. In addition, the parameters are often correlated. For example, the choice of van der Waals parameters is correlated with the partial charge, and dihedral terms are correlated with both the other bonded as well as nonbonded parameters. This leads to a number of challenges when attempting to generate a molecular force field description automatically. To select an appropriate atom type, one must first be able to identify the local chemical environment, determine if the atom is aromatic or aliphatic, if needed determine whether the atom may be a hydrogen-bond donor or acceptor, and assign a partial charge to an atom. In particular, assigning appropriate partial atomic charges is challenging. The most common approach to deriving partial atomic charges is to fit to the electrostatic potential of a molecule as obtained from a QM calculation. However, while the electrostatic potential around an atom can be calculated to high precision, the net charge on an atom is not itself an observable property. As a consequence, charges proposed on the basis of alternative charge assignment methods can differ significantly.[38] In addition, QM-derived charges depend on the level of theory used in the calculations, the precise conformation of the molecule, and the symmetry (point group) used when performing the calculations on the molecule. These differences can be significant even in very simple systems. Figure 1 shows the partial atomic charges obtained by fitting to the electrostatic potential of a molecule of benzene in implicit water calculated using the method of Kollmann−Singh[39] starting from the optimized geometry at the B3LYP/6-31G* level of theory as a function of the molecular symmetry (point group). As can be seen, there are dramatic differences in the charges assigned to the atoms using C_1 as opposed to enforcing D_{6h} symmetry. Symmetry must also be considered in the assignment of atom types and bonded parameters if the topologies are to be used in simulations where the molecules can change conformation. Last but not the least, the assignment of bonded parameters also represents a major challenge. Most biomolecular force fields contain bonded parameters for only a narrow range of molecular architectures, and new parameters must be derived for novel cases. However, the actual value of a bond angle in an optimized geometry may differ significantly from the relevant ideal value due to the effect of the local environment captured by other terms in the force field. Likewise, while force constants for specific internal degrees of freedoms can be derived from the Hessian (the derivative of the force with respect to the coordinates) as obtained from QM calculations, these will also contain contributions due to nonbonded interactions within the molecule.

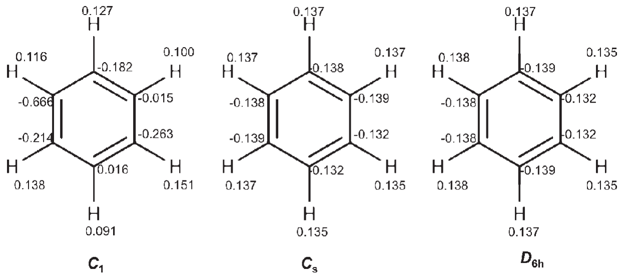

Figure 1. The partial atomic charges derived by fitting to the electrostatic potential at B3LYP/6-31G* level of theory for benzene optimized assuming C_1, C_s, and D_{6h} symmetry using the method of Kollmann and Singh.[39]

■ ATB PIPELINE

A flow-chart outlining the steps taken in the generation of a molecular topology using the ATB is outlined in Figure 2. The user is required to provide three pieces of information:

(i) A set of three-dimensional coordinates of all atoms (including all hydrogen atoms) in PDB format

(ii) A set of atomic connectivity data (PDB CONECT records)

(iii) The overall formal charge on the molecule

The combination of coordinates, connectivity, and overall charge provided by the user uniquely defines the stereochemistry as well as the protonation and tautomeric state of the molecule. In addition, the user is requested to provide a common name or a description of the molecule to assist when searching the database. The user is also asked to provide the IUPAC name of the molecule, any experimental data related to the free energy of hydration of the molecule, and a unique three to four-letter/-digit

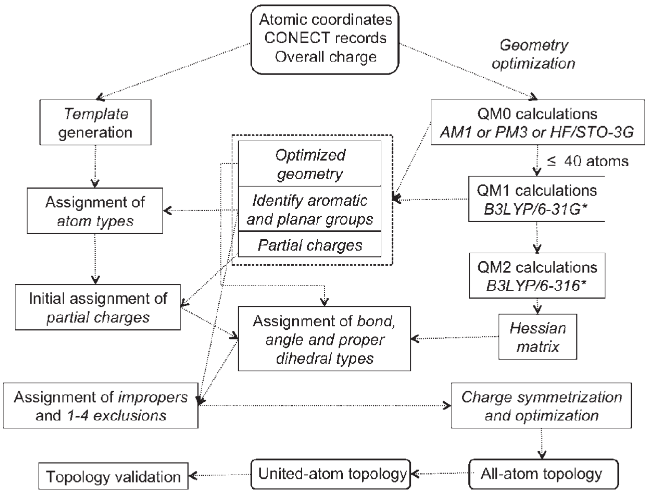

Figure 2. A flow-chart summarizing the primary steps in the generation of a molecular topology using the Automated Topology Builder (ATB).

residue name (RNME). If not provided by the user, a unique residue name will be generated automatically, and the IUPAC name is obtained using the "IUPAC name generator" utility of the program Jchem 5.2.03.1 (2009, ChemAxon, http://www.chemaxon.com). The user can also assign the molecule to a particular class: amino acid, nucleic acid, lipid, sugar, solvent, or heteromolecule. The topology builder uses a knowledge-based approach in combination with QM calculations to select parameters consistent with a given version of the GROMOS force field. An important feature of the ATB is that alternate bonded parameters are listed as comments in cases where the assignment is ambiguous. The user can then select the most appropriate parameter from the alternative choices.

The topology is constructed as follows:

1. Generation of the Initial Template. The first stage involves the generation of an initial template for the topology. This initial template contains only information that can be assigned unambiguously on the basis of the sequence of the atoms and the CONECT records. It does not incorporate information based on external rules or information extracted from the QM calculations. The standard GROMOS topology building block file contains various blocks. The TITLE, TOPPHYSCON, LINKEXCLUSIONS, and MTBUILDBLSOLUTE blocks are common to all molecules. The TITLE block can contain arbitrary text. Within the TITLE block, the ATB provides the date and time the file was generated, the IUPAC name of the molecule, a description of the molecule, and the revision date of the ATB program. The ATB also provides additional information as comments. These are placed before the TITLE block in the topology file. This includes general information regarding the ATB, warnings pertaining to any parameters that could not be assigned by the ATB, information on how to choose appropriate type codes when alternate bonded parameters are listed, the level of QM calculations performed, and the method used to calculate the initial partial atomic charges. The TOPPHYSCON block contains a set of standard physical constants. The LINKEXCLUSIONS block contains information on which atoms are to be excluded when linking individual monomeric units when building a polymeric chain in certain versions of the GROMOS force field, for example, when building a protein from a set of amino acid building blocks. Since the ATB generates topologies for complete molecules, the LINKEXCLUSIONS block is not used. The number of exclusions has been set arbitrarily to 2.

The MTBUILDBLSOLUTE block contains information concerning the molecule itself. The first line contains a unique identifier residue name (RNME). If the user does not supply a unique name, one will be generated by the ATB. This is followed by the number of atoms (NMAT) and the number of preceding exclusions (NLIN). NMAT is derived from the PDB file and is set equal to the number of lines containing atomic coordinates. NLIN is set to zero. NLIN is relevant only when linking more than one monomer in a polymeric chain. Various parameters are then given for each atom in the molecule. These include the atomic serial number (ATOM), the atom name (ANM), the atom type code (IACM), the atomic mass type code (MASS), the partial atomic charge (CGM), the charge group (ICGM), the number of exclusions (MAE), and a list of the excluded atoms (MSAE). ATOM and ANM are derived from columns 7−11 and 13−16 of the PDB coordinate file, respectively. The assignment of IACM depends on the chemical environment and is done at a later stage. The MASS types are assigned on the basis of the elemental symbol in columns 77−78 of the PDB file. The corresponding mass information is taken from the MASSATOMTYPECODE block of the appropriate interaction parameter file (e.g., IFP53A6.dat of the GROMOS 53A6 force field). As no charge has yet been assigned, CGM and ICGM are set to zero. In the GROMOS force field, first and second covalently bound neighbors (1−2 and 1−3 bonded pairs) are normally excluded from nonbonded interactions. In certain cases, such as within aromatic rings, third neighbors (1−4 bonded pairs) are also excluded. These are assigned later. Initially, all possible 1−2 and 1−3 pairs are identified on the basis of the CONECT records of the PDB file and used to assign the exclusions (MAE and MSAE).

The MTBUILDBLSOLUTE block then provides information on the bonds, bond angles, and the proper and improper dihedral angles. The total number of bonds (NB) is given followed by the list of the atom pairs (IB and JB), each with a corresponding bond type (MCB). The number of bonds and the pairs IB and JB are taken directly form the CONECT records. The MCB codes are assigned later. The CONECT records are also used to generate a list of all possible bond angles and proper dihedral angles. From this list, the total number of bond angles (NBA) and the atoms (IB, JB, and KB) forming a given angle are populated. Likewise, the total number of dihedral angles (NDA) and atoms forming the dihedral (IB, JB, KB, and LB) are populated. The corresponding angle and dihedral type codes (MCB) are not assigned at this stage. The assignment of improper dihedrals requires knowledge of whether certain atoms are part of aromatic rings or planar groups and are not assigned at this stage.

2. Quantum Mechanical Calculations. To assist in the assignment of appropriate parameters, a series of QM calculations is performed. All QM calculations are performed using GAMESS-US.[40] The molecule is initially optimized at the HF/STO-3G, AM1,[21] or PM3[41] level of theory as selected by the user. If the molecule contains ≤40 atoms, it is further optimized at the B3LYP/6-31G* level of theory[42−44] in an implicit solvent described using the polarizable continuum model (PCM[45]). The Hessian of the system is also calculated.

A range of information is extracted from the output of the QM calculations. This includes the following:

(i) **Bonded parameters**. The optimized geometry is used to obtain an indication of the bond lengths, bond angles, and dihedral angles, which are used to help assign appropriate bonds, bond angles, and dihedral type codes (MCB). In addition, the optimized geometry is used to identify any chiral centers within the molecule. The "bond order" is also extracted and used to verify the CONECT records supplied by the user.

(ii) **Identification of aromatic rings and planar groups**. These are identified from the optimized geometry and assist with the assignment of the atom types and improper dihedral angles along with the relevant 1−4 exclusions. To identify the aromatic rings, all of the rings in a given molecule are first identified on the basis of the loop closure from the connectivity records. Fused rings are identified from the shared connectivity between two or more rings. The rings are initially analyzed separately. Any redundancies are then removed so that the fused system is treated as a single unit. All possible proper dihedral angles that involve bonds that form the ring are identified. If all of these dihedral angles have a value between −5.0° and +5.0°, then the ring is classified as an aromatic ring. Next, the atoms (e.g., C), which are connected to three other atoms (e.g., X, O, and N) and are not a part of an aromatic ring, are identified. The improper dihedral between these four atoms as C−X−O−N is determined. If the value is between −10.0° and +10.0°, then the group is classified as planar (e.g., an amide group).

(iii) **Initial partial atomic charges**. The initial charges are estimated using either the MOPAC[46,47] method (for AM1 and PM3) or the method of Mulliken[48] (for HF/STO-3G) for molecules containing >40 atoms. For the molecules with ≤40 atoms, the initial atomic charges are generated by fitting the electrostatic potential (at the B3LYP/6-31G* level of theory) using the Kollmann−Singh[39] scheme. These are then further optimized by ATB.

(iv) **The Hessian**. The Hessian is used to estimate harmonic force constants for bond stretching and angle bending degrees of freedom as described by Seminario.[49]

3. Initial Assignment of Atom (IACM) Types. In the GROMOS force field, the IACM type code determines the Lennard-Jones parameters used to model van der Waals interactions. As in most biomolecular force fields, the appropriate IACM depends on the local chemical environment. For example, oxygen, nitrogen, carbon, and hydrogen can be modeled using multiple, different atom types. Thus, while the element type can be assigned on the basis of the elemental symbol in columns 77−78 of the PDB file, the value of IACM must be inferred on the basis of the connectivity, the chemical environment, and in some cases the net charge on the atom. Possible atom type codes are first extracted from the IAC block of the interaction parameter file (e.g., IFP53A6.dat). The GROMOS 53A6 force field contains 53 unique atom types. For oxygen (O), nitrogen (N), carbon (C), and hydrogen (H), the initial assignment is based on the following rules.

For oxygen, there are five different atom types (1−5). Type 1 is assigned in cases where the oxygen is bound to just one C atom (i.e., a carbonyl group; O═C) and where no other rule applies. Currently, type 2 is assigned in all cases where more than one O atom is attached to the same atom but which themselves are not attached to another atom (e.g., an O atom with double bonds in $-CO_2^-$, $-PO_4^{-2}$, $-SO_4^{-2}$, $-SO_2-$, $-NO_2$, etc. groups). This is appropriate in cases where there is a large partial (negative) charge on the oxygen. In cases where the partial charge is small, type 1 may be more appropriate. Type 3 is assigned in cases where the oxygen is bound to two atoms, at least one of which is not carbon (e.g., C−O−H, H−O−P, etc.). Type 4 is assigned in cases where the oxygen atom is connected to two carbon atoms [e.g., an O atom with two single bonds in ester or ether; C(═O)−O−C]. Type 5 is specific for the oxygen in a SPC water molecule and is not used for heteromolecules. There are six atom types for nitrogen (6−11). Nonaromatic nitrogen atoms connected to three neighbors are assigned either types 6 or 7. Cases where the nitrogen atom is connected to at most one hydrogen atom and a carbonyl carbon are assigned to type 6 [e.g., a N in $-CONHCH_3$]. Other nitrogen atoms connected to three neighbors are assigned to type 7 (e.g., a N in $-CONH_2$, $-NH_2$, etc.). Nitrogen atoms bound to four other atoms (e.g., a quaternary N atom with +1 charge) are assigned to type 8. When the nitrogen atom forms part of an aromatic ring with either two or three bonded neighbors, it is assigned to type 9. Nitrogen atoms that have a single neighbor (e.g., N in −C#N) are also assigned to type 9. The atom types 10 and 11 are specific for the side chain of arginine and are currently not assigned by ATB for heteromolecules. There are eight different atom types for carbon (12−19). Initially, all carbons are assigned to type 12 except if the atom in question is attached to four non-hydrogen atoms. In this case, the atom is assigned to type 13. Types 14−19 correspond to united-atom carbons: CH1, CH2, CH3, CH4, and CH-aromatic. These will be discussed later in relation to the conversion of an all-atom topology into a united-atom topology. Finally, there are two types for hydrogen. Type 20 is used for all hydrogens bound to a carbon, and type 21 is used for all other hydrogen atoms. The other atom type codes either are unique or refer to atoms parametrized for a given solvent. The solvent atom types are not used by ATB, except atom type 42 (sulfur in DMSO solvent in GROMOS), which is used for molecules containing S with more than two connected neighbors.

4. Assignment of Bond and Bond Angle Types (MCB). The bond and bond angles are derived from the BONDTYPECODE and the BONDANGLETYPECOD blocks of the IFP53A6.dat file. The value of the bond, bond angle, and corresponding force constant along with the atom types are matched to the standard GROMOS bond and angle types in the IFP53A6.dat file. The matching threshold is set to ±0.004 nm for the bond length and ±5.0° for the bond angle. For molecules with ≤40 atoms, the Hessian is available from the QM calculations, from which the force constants for the bonds and bond angles can be estimated using the method of Seminario.[49] The matching threshold for the force constant for the bonds is $\pm 1 \times 10^6$ kJ mol^{-1} nm^{-4}, while that for bond angles is ±100.0 kJ mol^{-1}. Note, in the GROMOS 53A6 parameter set, the bond term is quartic in the bond length and the angle term is dependent on the cosine of the bond angle. In cases where a suitable match is not found, a new bond or bond angle type is introduced with the corresponding force constant and value as obtained from the QM calculations. These new parameters are marked as "nonstandard" in the topology file to indicate them appropriately. The angle type 41 in GROMOS is specific for the heme with C#O (Fe−C#O, 180°) and is not used for the heteromolecules. Instead a new (nonstandard) angle type 55 has been introduced with a force constant of 500.0 kJ mol^{-1} and angle of 180° for molecules with linear groups −C#N, −C#C−, −C═C═C−, −N═C═S, etc.

5. Assignment of Proper Dihedral Angles. The proper dihedral angles are derived from the DIHEDRALTYPECODE block of the IFP53A6.dat file. Initially, to build the template building block, all possible dihedral angles for a given bond are listed. Redundant proper dihedrals are then removed. In most cases in the GROMOS force field, only one set of atoms *i*, *j*, *k*, and *l* is chosen to define a dihedral angle around the central bond between atoms *j* and *k*.

The multiplicity (m_n) of a dihedral angle is determined based on the connectivities of atoms *j* and *k* as follows. First, the value of *M* is determined on the basis of the connectivity information:

$$M = A \times B \tag{1}$$

where A = (number of atoms connected to j) − 1 and B = (number of atoms connected to k) − 1.

The possible values for *M* in heteromolecules are 1, 2, 3, 4, 6 and 9:

if M = 1, 2, 3, or 6, then $m_n = M$

if M = 4 or 9, then $m_n = M^{1/2}$, i.e., m_n = 2 or 3, respectively. Thus, possible values of m_n are 1, 2, 3, and 6. The GROMOS force field also contains a dihedral angle with a multiplicity of 4. This is used specifically for heme groups and is not assigned in the heteromolecules.

Next, the phase shift is determined by evaluating the potential energy of the dihedral angle using eq 2:

$$V^{\text{trig}}(r; s) = V^{\text{trig}}(r, K_{\varphi}, \delta_m) = \sum_{n=1}^{N_{\varphi}} K_{\varphi,n}[1 + \cos(\delta_n)\cos(m_n\varphi_n)] \tag{2}$$

where δ_n is the phase shift, which is restricted to 0 or π (i.e., cos $\delta_n = \pm 1.0$), m_n is the multiplicity of the torsion dihedral angle, and φ_n is the actual value of the dihedral angle defined by atoms *i*, *j*, *k*, and *l*. Equation 2 indicates that a phase shift of 180° and a multiplicity of 1 means that the potential reaches the maximum value at 180° (if multiplicity = 3, then at −60°, 60°, and 180°). Equation 2 is evaluated for cos δ_n = +1 or −1 for each dihedral angle in a given molecule with a given multiplicity (m_n) and the dihedral angle value (φ_n). The cos δ_n value that gives lower value for the potential energy is taken as the phase shift (δ_n) for that dihedral angle. Possible dihedral type codes for a given dihedral angle are identified by matching the multiplicity and phase shift to types in the IFP53A6.dat file.

The final selection is then based on the combination of atom types. In ambiguous cases, multiple options are presented. In case no suitable match is found, the standard GROMOS dihedral type with the lowest force constant with the corresponding phase shift and multiplicity is chosen.

Dihedral type 16 (force constant of 0.0 kJ mol^{-1}) is not used for heteromolecules. Since there is no dihedral angle type with a phase shift of −1 and multiplicity of 3 in the GROMOS force field, a new (nonstandard) type 42 was introduced with a force constant of 1.0 kJ mol^{-1}. No dihedral is assigned in cases where one of the bond angles involved is 180°, i.e., in molecules containing a linear group such as an alkyne (−C#C−) or azido (−N=N=N).

6. Assignment of Improper Dihedral Angles and 1,4 Exclusions in the Case of Aromatic Rings, Planar Groups, and Chiral Centers. Aromatic rings, planar groups, and chiral centers are identified on the basis of the QM optimized geometry as described previously. The improper dihedral angles are assigned from the IMPDIHEDRALTYPEC block of the IFP53A6.dat file. To maintain the planarity of an aromatic ring and or planar groups such as $-NO_2$ or $-NH_2$, a type 1 (planar) improper dihedral is assigned. In cases where the chirality of a group cannot be maintained by angle terms alone, such as in the case of a chiral united-atom −CH group, a type 2 (tetrahedral) is assigned. Where a proper dihedral angle involving the central two atoms had been assigned previously, such as for atoms involved in an aromatic ring, the proper dihedral is removed and only the equivalent improper dihedral retained. The number of proper dihedrals (NDA) is adjusted accordingly. For those atoms that either form part of an aromatic ring or are attached directly to the ring, all possible 1−4 pairs are determined. These 1−4 pairs are added to the list of exclusions in MAE and the value of MSAE updated.

7. Optimization of the Partial Atomic Charges and Assignment of Charge Groups (CGM and ICGM). The initial partial atomic charges are extracted from the output of the QM calculations, rounded to three decimal places, and put in the column CGM. As discussed previously, the assignment of partial charges to atoms represents a major challenge in force field development. This is because it is not possible to relate the charge on an atom to a physical observable, and while a number of models have been developed that can be used to infer charges based on the electron density as derived from QM calculations, they are based on subjective assumptions and do not yield a unique answer. In addition, the charges derived from QM calculations are often very sensitive to the geometry of the molecule. For example, the charges assigned to equivalent chemical groups in a molecule can differ significantly, making such QM charges inappropriate for use in molecular dynamics simulations. In an attempt to overcome these limitations and obtain charges compatible with the GROMOS force field, a number of charge optimization steps are performed. First, the molecule is analyzed in order to identify any global and/or local symmetry. The charges on atoms in equivalent chemical environments are then averaged appropriately. The effect of this procedure is illustrated in Figure 3, which shows the initial QM derived charges calculated by fitting them to the electrostatic potential using the Kollmann−Singh scheme at the B3LYP/6-31G* level of theory and the final optimized ATB charges for hydroquinone and isopropanol. Hydroquinone contains two sets of equivalent −O−H groups and four sets of equivalent −C−H groups. These symmetric groups were identified, and the charges on individual atoms were averaged and reassigned (Figure 3a). Isopropanol has both local (within the individual methyl groups) as well as global symmetry (the two equivalent methyl groups). First, the charges on the three equivalent hydrogen atoms of the methyl group are averaged and reassigned. Then, the methyl groups are identified as equivalent and the charge of the corresponding atoms averaged and reassigned (Figure 3b).

Next, charge groups are assigned by grouping small numbers of covalently bound atoms together such that the overall charge in a given charge group is either 0.0 or +1.0 or −1.0. For this, all of the hydrogen atoms attached to a single heavy atom are first merged into a single charge group. Any heavy atoms connected to only one other heavy atom are then identified and merged into the adjoining charge groups. The charge groups containing a residual charge are identified, and the residual charge is transferred to the atom within the charge group with the highest charge such that now the overall charge in the charge group is either 0.0 or +1.0 or −1.0. Next, the total charge of the molecule is calculated and compared to the value supplied by the user. Any residual charge that results from rounding etc. is transferred to the atom with the largest charge. Note, members of a charge group must be numbered sequentially, and thus the atoms in the topology file and the associated coordinate file are reordered if

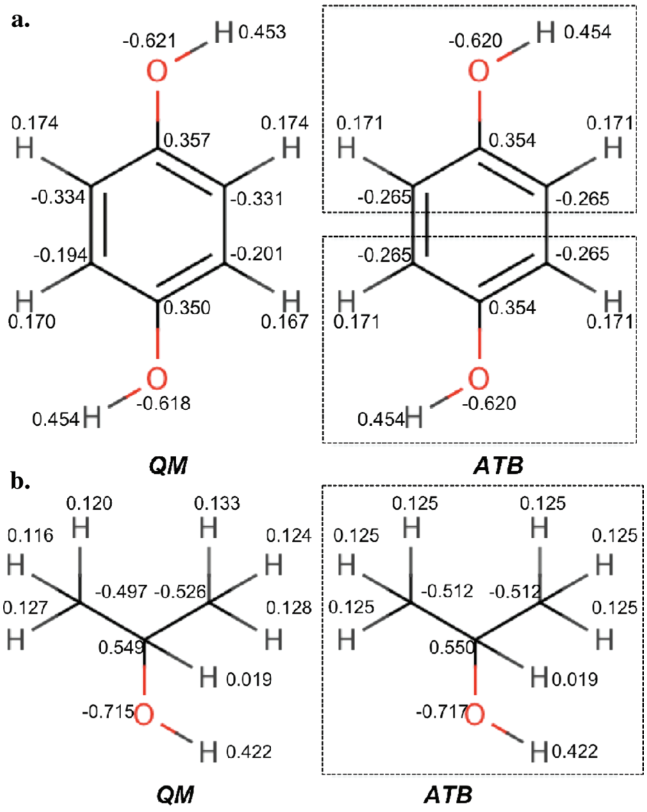

Figure 3. The partial atomic charges derived for molecules (a) hydroquinone and (b) isopropanol from the QM calculations by fitting to the electrostatic potential at the B3LYP/6-31G* level of theory using the method of Kollmann−Singh[39] and final ATB charges, respectively. The atoms within the square formed by dotted lines form a single charge group with overall charge of the charge group being either 0.0 (in these cases) or +1.0 or −1.0.

required. This charge optimization process results in atomic charges that are less sensitive to the geometry of the molecule used to calculate the QM charges. The charges are thus more robust and more in line with the philosophy of the GROMOS force field.

8. Conversion from an All-Atom to a United-Atom GROMOS Topology. In the GROMOS force field, aliphatic (alkanes: $-CH$, $-CH_2$, $-CH_3$ and $-CH_4$) carbon atoms are treated as united atoms. In ATB, only aliphatic carbon atoms are converted to a united-atom model. The alkene (e.g., $-CH{=}CH-$), alkyne (e.g $-C\#C-$), and aromatic $-CH$ groups are retained as all-atom.

A united-atom GROMOS topology building block is generated from the all-atom topology building block in a stepwise manner.

(i) The C atoms attached to four different atoms and at least one H atom is identified from the CONECT records as either $-CH$, $-CH_2$, or $-CH_3$.

(ii) The partial atomic charges of the H atoms are added to the charge of the corresponding C atom.

(iii) The atom type for each of these C atoms is changed from type 12 to the appropriate united-atom C type, 14, 15, and 16 for $-CH$, $-CH_2$, and $-CH_3$, respectively.

(iv) The H atoms that were attached to these groups are deleted from the topology file and the coordinate file and the remaining atoms renumbered accordingly.

(vi) For a united atom $-CH$, the improper dihedral angle between $-C(X)(Y)Z$ is calculated. If the improper dihedral angle is 35.0° ± 5.0°, then an additional improper dihedral angle with type 2 is added to the topology to maintain the chirality of the $-CH$ atom. If the improper dihedral angle is −35.0° ± 5.0°, then one of the pairs from X−Y−Z is swapped (for, e.g., $-C(X)(Z)Y$) to obtain the value of 35.0° ± 5.0°, and the improper dihedral is assigned appropriately. The number of improper dihedrals (NIDA) is updated. This is done for all of the $-CH$ united atoms in the molecule. Note in the case of aliphatic and alicyclic amines, no improper dihedral is assigned, so that pyramidal inversion of the nitrogen atom is possible.

9. Validation of the Force Field Topology. The united-atom GROMOS building block file is validated using the program *check_top* in the GROMOS package.[50] First, a complete force field topology is constructed from the building block file using the program *make_top*. It is this file that is used in the validation. The program *check_top* performs basic checks on the charge groups, exclusions, bonds, angles, and improper dihedrals. For example, *check_top* will identify charge groups with a residual charge; missing dihedral angles, if the overall charge on the molecule is not a whole number; and possible inconsistencies in the exclusions. If the molecule contains 1−4 exclusions, a warning to check for the presence of aromatic rings in the molecule is issued. Any inconsistencies between the atom type and corresponding charge and bonded types assigned by the ATB and those in the standard GROMOS force field along with possible alternatives are then listed. Finally, the potential energy of each of the bonded terms (bonds, bond angles, proper, and improper dihedrals angles) is then calculated on the basis of the optimized geometry. This allows the user to identify whether a given set of parameters is incompatible with the optimized geometry.

10. Conversion to Other Formats. The final GROMOS force field files can in principle be converted to a wide variety of other formats. Currently, force field files that can be read by the GROMACS simulation package and the CNS structural refinement program are provided. The force field (molecule.itp) files provided for use with GROMACS retain all of the information contained in the original GROMOS file. Information concerning 1−4 pairs for which special van der Waals parameters are assigned as part of the GROMOS force fields are included in the "[pairs]" block in the GROMACS molecule.itp file. In GROMACS, exclusions are generated automatically on the basis of the parameter "*nrexcl*". By setting *nrexcl* = 3, the 1−2 and 1−3 exclusions are generated automatically. Additional 1−4 exclusions, such as those found in aromatic rings, are included in the "[exclusions]" block. In the molecule.itp file generated by ATB, the parameters for the bonded terms are included explicitly for each interaction to avoid possible incompatibilities when using type codes in GROMACS.

The CNS output contains two files: a topology file and a parameter file. Both are derived from the GROMOS building block file. The topology file contains atom type information. In addition, all bonds, bond angles, and dihedral angles are listed using atom names rather than atom serial numbers. The parameter file contains the interaction parameters. Again, specific values are listed for each of the bonded terms.

REPOSITORY FOR GROMOS FORCE FIELDS AND PRE-EQUILIBRATED SYSTEMS

The ATB also provides a repository for the topologies of molecules that have been individually optimized as part of the generation of the GROMOS family of force fields, including the GROMOS 43, 45, and 53 parameter sets. The original force field files can be downloaded from "the GROMOS force field" tab on the Web server. A range of equilibrated systems are also provided under the "pre-equilibrated systems" tab. These include a range of common solvents such as methanol, ethanol, and acetone as well as a series of small organic molecules including ethers, esters, alcohols, ketones, and carboxylic acids that were used in the parametrization of the GROMOS 53A6 force field. The pre-equilibrated boxes are available in PDB and GROMOS96 file formats along with the corresponding GROMOS input file, force field topologies in various formats, and details of the simulations (temperature, pressure, step size, time duration for equilibration). Links to relevant literature references are also provided. A range of pre-equilibrated mixed systems such as lipid bilayers is currently being added.

CURRENT LIMITATIONS AND ONGOING DEVELOPMENTS

While the ATB aims to generate topologies for a wide range of heteromolecules, it does have several limitations:

(i) The GROMOS force field does not contain parameters for all possible atom types. Atoms such as boron and iodine are not included in the standard GROMOS force field. When novel atom types are encountered, the current version of ATB terminates after the generation of the initial template.

(ii) The GROMOS force field is primarily intended for biomolecular systems. The rules that have been developed to assign specific atom types are limited to those that can be derived from biomolecules present in the current GROMOS force fields.

(iii) The size of the molecule is limited. Higher-level QM calculations, required to determine the Hessian, are only performed if the molecule contains ≤40 atoms.

(iv) Currently, the Hessian is used only to estimate the force constants for bonds and bond angles. The dihedral angle terms are highly correlated with the nonbonded parameters such as partial charges, Lennard-Jones terms, exclusions, etc. While in principle corrections to the Hessian can be applied, in the current version of the ATB, the force constants for dihedral angles are not estimated on the basis of the Hessian.

(v) The possibility to scale the partial charges of atoms in specific functional groups in order to provide a better match to the existing GROMOS 53A6 force field has been implemented, but charge scaling is currently not applied. For example, primary amines are known to be problematic in classical force fields, and the magnitude of the optimized charges used in the GROMOS force field are significantly larger than those obtained using QM approaches.[51]

(vi) At present, atom types are assigned on the basis of the local environment as determined by connectivity. The incorporation of atomic charge in the assignment of atom types is under development.

(vii) Force field topologies are currently only provided in GROMOS, GROMACS, and CNS formats the ATB.

VALIDATION OF THE ATB

To test the validity of the force field descriptions generated by the ATB, the ability to reproduce the thermodynamic properties of a range of molecules has been examined. The GROMOS 53A6 force field has been parametrized to reproduce the density and heats of vaporization of pure organic liquids as well as to reproduce the free energies of solvation of analogs of the side chains of common amino acids in polar and apolar solvents. As a part of the validation, the topologies generated by the ATB have been used to estimate the free energy of hydration of analogs of the side chains of amino acids. In addition, the free energy of hydration of a set of 90 biologically relevant small organic molecules[52−54] (Table S1, Supporting Information) and 100 chemically diverse drug-like and drug molecules taken from the "Statistical Assessment of the Modeling of Proteins and Ligands" (SAMPL) challenges including the CUP8(SAMPL0),[55] SAMPL1,[56] and SAMPL2[57] data sets has been determined (Table S2, Supporting Information).

As an initial test of the validity of the molecular descriptions generated by the ATB, the all-atom RMSD between the QM optimized structure used in the parametrization and the structure after 1 ns of simulation in SPC water were determined for the 190 test molecules (Tables S1 and S2, Supporting Information). For 156 of these molecules (∼80%), the all-atom RMSD between the simulated and QM optimized structure was <0.1 nm, demonstrating that the QM optimized structure is also a minimum in the force field generated by the ATB. Of the remainder, for 26 molecules, the RMSD was between 0.1 and 0.2 nm, and for eight molecules, the RMSD was above 0.2 nm. All of these molecules, however, contained either an alkane chain or a similar flexible group. Ketoprofen (Table S2), which contains a slightly flexible and large benzophenone ring with a branched chain, exhibited the highest RMSD of 0.27 nm.

The free energies of hydration were calculated using the GROMOS96[58] simulation package. A given solute molecule was placed in a cubic periodic box filled with SPC[59] water molecules. The size of the box was chosen such that no solvent molecule interacted with more than one periodic image of the solute. After energy minimization, the initial velocities were assigned from a Maxwell−Boltzmann distribution corresponding to a temperature of 298.15 K. All bond lengths were constrained using the SHAKE[60] algorithm with a relative geometry accuracy of 10^{-4}. The equations of motion were integrated using the leapfrog algorithm and a time step of 2 fs. All simulations were performed at a constant temperature (298.15 K) and pressure (1 atm) using a Berendsen thermostat (coupling time of 0.1 ps) and barostat (coupling time of 0.5 ps and isothermal compressibility of 4.575×10^{-4} (kJ mol^{-1} nm^{-3})$^{-1}$).[61] Nonbonded interactions were calculated using a triple-range scheme. Interactions within a shorter-range cutoff of 0.8 nm were calculated every time step from a pair list that was generated every five steps. At these time points, interactions between 0.8 and 1.4 nm were calculated as well and kept constant between updates. A reaction field contribution was added to the electrostatic interactions and forces, to account for a homogeneous medium outside the long-range cutoff. The relative permittivity for the reaction field was set to a value of 61 for SPC water.

The free energy of solvation was calculated using the thermodynamic integration (TI)[62] approach. To determine the free energy of hydration, all nonbonded interactions involving solute atoms were scaled to zero in a stepwise manner as a function of a coupling parameter λ. The change in free energy corresponding

Table 1. Comparison between the Experimental (exptl) and Calculated (calcd) Free Energies of Hydration (ΔG_{hyd}) for Analogs of the Side Chains of Common Amino Acids Calculated Using the GROMOS96 Force Field and Parameters Assigned by the Automated Topology Builder (ATB)[a]

amino acid	side chain analog	$\Delta G_{hyd;exptl}$[1]	$\Delta G_{hyd;calcd}$ GROMOS 53A6	$\|\Delta G_{hyd;calcd} - \Delta G_{hyd;exptl}\|$ GROMOS 53A6	$\Delta G_{hyd;calcd}$ ATB	$\|\Delta G_{hyd;calcd} - \Delta G_{hyd;exptl}\|$ ATB
Ala	methane	8.1; 8.4	6.8	1.3; 1.5	6.8	1.3; 1.5
Arg	*n*-propryl-guanidine	−45.7	−48.6	2.9	−43.0	2.7
Asn	acetamide	−40.6	−40.9	0.3	−37.9	2.7
Asp	acetic acid	−28.0	−30.5	2.5	−29.8	1.8
Cys	methane thiol	−5.2	−6.3	1.1	−8.4	3.2
Gln	propanamide	−39.4	−40.4	1.0	−34.4	5.0
Glu	propionic acid	−27.0	−30.0	3.0	−31.6	4.6
His	methyl imidazole	−42.9	−44.5	1.6	−42.9	0.0
Ile	*n*-butane	8.7; 8.8	8.8	0.1; 0.0	7.4	1.3; 1.4
Leu	isobutane	9.4; 9.7	10.0	0.6; 0.3	8.2	1.2; 1.5
Lys	*n*-butyl-amine	−18.3	−20.0	1.7	−9.0	9.3
Met	ethyl methyl sulfide	−6.2	−7.6	1.4	1.8	8.0
Phe	toluene	−3.1	−1.0	2.1	1.9	5.0
Ser	methanol	−21.2	−23.1	0.9	−21.9	0.7
Thr	ethanol	−20.5	−19.9	0.6	−21.8	1.3
Trp	3-methyl-indole	−24.7	−24.5	0.2	−16.2	8.5
Tyr	*p*-cresol	−26.6	−25.1	1.5	−25.8	0.8
Val	propane	8.2	7.8	0.4	7.2	1.0
average				1.2		3.3

[a] The absolute value of the difference between the calculated and experimental values is also given. Values are in kJ mol^{-1}.

to the removal of all solute nonbonded interactions was then calculated by integrating the average value of the derivative of the Hamiltonian H of the system with respect to λ:

$$\Delta G = \int_0^1 \left\langle \frac{\partial H}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{3}$$

The integral above was evaluated using 21 evenly spaced λ points with 1 ns of data collection at each λ point. Standard errors in $\langle \partial H/\partial \lambda \rangle_\lambda$ were estimated at every λ point using a block-averaging procedure.[63] The individual errors were integrated from $\lambda = 0$ to $\lambda = 1$ to yield the estimate of the total error in ΔG. A soft-core[64,65] interaction was used to avoid singularities in the nonbonded interaction function at the end state ($\lambda = 1$). The free energy of solvation was calculated as the difference between the free energy change calculated from a vacuum simulation of the solute and the free energy change when the solute is in solution. In the vacuum simulation, a stochastic bath was applied at a reference temperature of 298.15 K using an atomic friction coefficient of 91 ps^{-1}.

The primary target function in the parametrization of the GROMOS 53A6 force field was the ability to reproduce the free energy of hydration of biomolecular systems. Table 1 lists the free energy of hydration for analogs of the side chains of common amino acids calculated using parameters taken from the GROMOS 53A6 force field, calculated using united atom parameters generated using the ATB and measured experimentally. Note, the ATB uses a combination of structure calculations and a rule-based approach to obtain parameters. It aims to be compatible with the GROMOS force field but does not attempt to directly match specific functional groups to those in the GROMOS 53A6 force field. Thus, while atom types and bonded parameters will match exactly, there will (potentially) be differences in the partial charges and charge groups, as in the GROMOS force field these have been manually optimized individually for each group. Nevertheless, there is a very close correspondence between the values calculated using the GROMOS 53A6 force field and those values calculated using the force field generated using the ATB. This demonstrates that the parameters generated by the ATB are compatible with those of the GROMOS 53A6 force field. Most importantly, the ability of the two sets of parameters to reproduce the experimental free energies of hydration is similar in most cases. The average deviation between the values calculated using the GROMOS 53A6 force field and the experimental values is 1.2 kJ mol^{-1}. The average deviation between the values calculated using the ATB and the experimental values is marginally larger at 3.3 kJ mol^{-1}. In several cases, the difference between the calculated and experimental free energies of hydration is greater for the GROMOS 53A6 force field than for parameters derived from the ATB. As illustrated graphically in Figure 4, which shows the calculated hydration free energies plotted against the values obtained experimentally, the increased average deviation is primarily due to just three compounds *n*-butyl-amine (Lys; dev = 9.3 kJ mol^{-1}), ethyl methyl sulfide (Met; dev = 8.0 kJ mol^{-1}), and 3-methyl-indole (Trp; dev = 8.5 kJ mol^{-1}). Primary amines, aromatic nitrogens, and sulfur-containing groups are know to be problematic within the GROMOS force field, suggesting either that specific scaling factors for the charges in these cases may need to be introduced or that alternative van der Waals parameters need to be developed for these atoms types.

To further validate the parameters generated by the ATB, the free energies of hydration of ~90 biologically relevant small organic molecules were calculated. The molecules encompass chemical classes such as alkanes, cycloalkanes, alkenes, alkynes, alkyl benzenes, amines, amides, aldehydes, carboxylic acids, esters, ketones, thios, and sulfides. The name, access code, and chemical structure of the compounds in the ATB together with the experimental

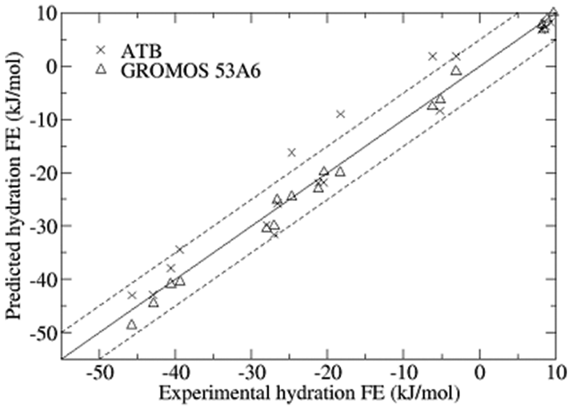

Figure 4. Free energy of hydration of amino acid side chain analogs. Comparison of experimental free energy of hydration to calculated values obtained using the GROMOS 53A6 force field (triangles) and ATB (crosses) for 18 compounds listed in Table 1. The diagonal line corresponds to perfect agreement with experimental results. Dotted lines indicate ±5.0 kJ mol^{-1} deviations from the diagonal line.

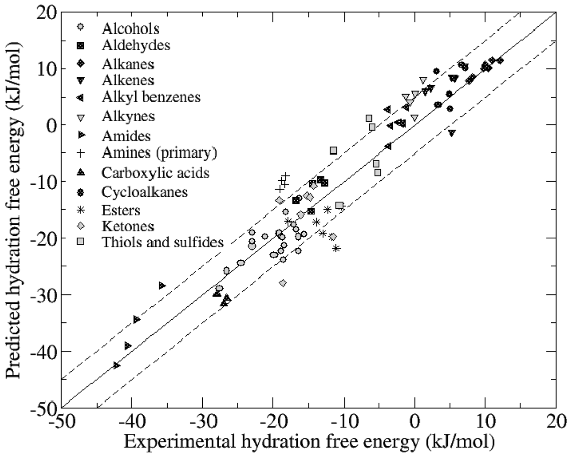

Figure 5. Free energy of hydration of ∼90 biologically relevant small organic molecules. The diagonal line corresponds to perfect agreement with experimental results. Dotted lines indicate ±5.0 kJ mol^{-1} deviations from the diagonal line.

and calculated hydration free energies as well as the difference between these values are given in Table S1 as Supporting Information. The average deviation between the values calculated using the force field obtained from the ATB and the experimental values is 3.3 kJ mol^{-1} specifically, alcohols, 2.4 kJ mol^{-1}; aldehydes, 2.0 kJ mol^{-1}; alkanes, 0.4 kJ mol^{-1}; alkenes, 3.9 kJ mol^{-1}; alkyl benzenes, 3.0 kJ mol^{-1}; alkynes, 5.0 kJ mol^{-1}; amides, 3.7 kJ mol^{-1}; amines (primary), 8.5 kJ mol^{-1}; carboxylic acids, 3.5 kJ mol^{-1}; cycloalkanes, 2.5 kJ mol^{-1}; esters, 5.1 kJ mol^{-1}; ketones, 2.1 kJ mol^{-1}; and thiols and sulfides, 5.0 kJ mol^{-1}. In Figure 5, the calculated free energies of hydration are plotted versus the experimental values. While for cases such as simple alkanes the calculated hydration free energies were distributed around the experimental values, the hydration free energies of, in particular, amines (primary), alkynes, and sulfides were systematically overestimated, while esters were systematically underestimated. This suggests that the application of scaling factors on the charges and/or the reparameterization of atoms involved in these specific groups would lead to systematic improvements in the predictive power of the force field. The magnitudes of the partial atomic charges derived from the QM calculations for sulfides, for example, are significantly smaller than those used for the side chain of methionine in the GROMOS force field. Primary amines and esters are known to be problematic in the GROMOS force field, and work to improve these groups is currently underway.

As a further validation of the ATB, the hydration free energies of ∼100 drug-like molecules have also been calculated (Supporting Information Table S2). The test molecules were taken from the CUP8(SAMPL0), SAMPL1 and SAMPL2 data sets and represent a very diverse range of functional groups, including groups that do not currently form part of the GROMOS force field and groups that are known to be nonoptimal in the GROMOS force field. Of the 100 molecules, 40 contained at least one halogen (−F, −Cl, −Br). A small number of halogen-containing compounds, notably triflouroethanol and chloroform, have been specifically parametrized as part of the GROMOS force field. However, the transferability of these parameters has not been investigated systematically. It should also be noted that the uncertainty in the experimental hydration free energies was up to 8.1 kJ mol^{-1} in many cases, much larger than the statistical uncertainty in the calculated values.

Of the 100 molecules, the calculated hydration free energy was within the combined statistical uncertainty of the calculation and experiment in 34 cases. In an additional 25 cases, the calculated value lay within 5 kJ mol^{-1} of the experimental value, again allowing for the combined statistical uncertainty. Overall, the average deviation between the values calculated using the force field obtained from the ATB and the experimental values was 9.1 kJ mol^{-1}. This was primarily due to molecules containing halogens. Of 40 molecules containing one or more halogens, the free energy of hydration was overpredicted by more than 5 kJ mol^{-1} in 28 cases. Furthermore, the extent to which the free energy of hydration was overestimated was strongly correlated with the number of halogens in the molecule unless the molecule contained additional compensating functional groups. The average deviation for the 60 molecules remaining was 7.1 kJ mol^{-1}. In general, the free energies of hydration for molecules containing $-NO_2$, ether, and/or N-alkyl groups were overestimated, while molecules containing an ester or P or S were underestimated. Again, the systematic nature of these deviations suggests that there is scope for the optimization of these groups, and work to this end is underway.

■ CONCLUSIONS

The automatic generation of molecular force fields for novel molecules compatible with biomolecular force fields such as GROMOS, AMBER, CHARMM, and OPLS is an ongoing challenge. The Automated force field Topology Builder (ATB, http://compbio.biosci.uq.edu.au/atb) and repository described here is a unique Web server that can provide topologies and parameters for a wide range of molecules appropriate for use in molecular simulations, computational drug design, and X-ray refinement. It has the advantage over other comparable servers in that the user is required to supply sufficient information (the coordinates and connectivity of all atoms in the molecule along with the formal charge) in order to define uniquely the geometry and stereochemistry as well as the protonation and tautomeric states of the molecule. The ATB combines information from QM calculations with a knowledge-based approach to derive both all-atom and united-atom force field descriptions of novel molecules compatible with the GROMOS force field in a variety of formats, including the one for the X-ray refinement program CNS. The symmetry of the molecule is analyzed to ensure that equivalent atoms are assigned identical parameters. Charge groups are assigned automatically. An important feature of the ATB is that it is recognized that it is not possible to unambiguously assign parameters in many cases, and a range of possible alternatives is provided where appropriate. The ATB also acts as a repository for the GROMOS family of force fields and a range of pre-equilibrated systems. At the time of submission, the repository contained an excess of 2100 molecules, including over 100 sugars and 60 lipids.

ASSOCIATED CONTENT

S Supporting Information. Tables containing the hydration free energy and RMSD data for 190 molecules. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: a.e.mark@uq.edu.au, a.malde@uq.edu.au.

Author Contributions

Scientific aspects and research design: A.K.M. and A.E.M. Code development: A.K.M., L.Z., M.B., M.S., and A.E.M. Manuscript writing: A.K.M. and A.E.M. Validation of the ATB: A.K.M. and A.E.M. Contribution to ATB repository: A.K.M., M.S., D.P., P.C.N., C.O., and A.E.M.

ACKNOWLEDGMENT

The authors would like to thank Wilfred van Gunsteren, Philippe Hünenberger, Volker Knecht, Alexandre Bonvin, and Bruno Horta for valuable discussions and Nikhil Biyani for his work on the Hessian. We would also like to thank the many users who have helped identify and correct errors during the development of the ATB. A.K.M. acknowledges the Australian Research Council (ARC; grant no. DP0987043) for the award of an Australian Post-Doctoral (APD) fellowship. P.C.N. acknowledges the Australian government and the University of Queensland for the award of an Endeavour International Postgraduate Research Scholarship (IPRS). Computational resources were provided by the National Computational Infrastructure (NCI, Australia) National Facility through projects m72 and n63.

REFERENCES

(1) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. 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.* **2004**, *25*, 1656.

(2) Soares, T. A.; Daura, X.; Oostenbrink, C.; Smith, L. J.; van Gunsteren, W. F. Validation of the GROMOS force-field parameter set 45A3 against nuclear magnetic resonance data of hen egg lysozyme. *J. Biomol. NMR* **2004**, *30*, 407.

(3) Lins, R. D.; Hünenberger, P. H. A new GROMOS force field for hexopyranose-based carbohydrates. *J. Comput. Chem.* **2005**, *26*, 1400.

(4) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. Validation of the 53A6 GROMOS force field. *Eur. Biophys. J. Biophys. Lett.* **2005**, *34*, 273.

(5) Soares, T. A.; Hünenberger, P. H.; Kastenholz, M. A.; Krautler, V.; Lenz, T.; Lins, R. D.; Oostenbrink, C.; van Gunsteren, W. F. An improved nucleic acid parameter set for the GROMOS force field. *J. Comput. Chem.* **2005**, *26*, 725.

(6) Jayaram, B.; Sprous, D.; Beveridge, D. L. Solvation free energy of biomacromolecules: Parameters for a modified generalized born model consistent with the AMBER force field. *J. Phys. Chem. B* **1998**, *102*, 9571.

(7) Wang, J. M.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157.

(8) Patel, S.; Brooks, C. L., III. CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* **2004**, *25*, 1.

(9) Patel, S.; Mackerell, A. D.; Brooks, C. L., III. CHARMM fluctuating charge force field for proteins: II - Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* **2004**, *25*, 1504.

(10) Jorgensen, W. L.; Tirado-Rives, J. The OPLS potential functions for proteins - Energy minimizations for crystals of cyclic-peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(11) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(12) Allinger, N. L. Conformational-analysis 0.130. MM2 - Hydrocarbon force-field utilizing V1 and V2 torsional terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127.

(13) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular mechanics - The MM3 force-field for hydrocarbons. 1. *J. Am. Chem. Soc.* **1989**, *111*, 8551.

(14) Allinger, N. L.; Chen, K. S.; Lii, J. H. An improved force field (MM4) for saturated hydrocarbons. *J. Comput. Chem.* **1996**, *17*, 642.

(15) Allinger, N. L.; Chen, K. H.; Lii, J. H.; Durkin, K. A. Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. *J. Comput. Chem.* **2003**, *24*, 1447.

(16) Halgren, T. A. Merck molecular force field. 1. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490.

(17) Halgren, T. A. Merck molecular force field. 2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17*, 520.

(18) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem.* **1999**, *20*, 730.

(19) Moriarty, N. W.; Grosse-Kunstleve, R. W.; Adams, P. D. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. *Acta Crystallogr., Sect. D* **2009**, *65*, 1074.

(20) Adams, P. D.; Afonine, P. V.; Bunkoczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr., Sect. D* **2010**, *66*, 213.

(21) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The development and use of quantum-mechanical molecular-models 0.76. AM1 - A new general-purpose quantum-mechanical molecular-model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(22) Bailey, S. The CCP4 Suite - Programs for protein crystallography. *Acta Crystallogr., Sect. D* **1994**, *50*, 760.

(23) Vagin, A. A.; Murshudov, G. N.; Strokopytov, B. V. BLANC: the program suite for protein crystallography. *J. Appl. Crystallogr.* **1998**, *31*, 98.

(24) Vagin, A. A.; Steiner, R. A.; Lebedev, A. A.; Potterton, L.; McNicholas, S.; Long, F.; Murshudov, G. N. REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr., Sect. D* **2004**, *60*, 2184.

(25) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect. D* **2010**, *66*, 486.

(26) Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr., Sect. D* **1998**, *54*, 905.

(27) Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2007**, *2*, 2728.

(28) Kleywegt, G. J. Crystallographic refinement of ligand complexes. *Acta Crystallogr., Sect. D* **2007**, *63*, 94.

(29) Nilsson, K.; Lecerof, D.; Sigfridsson, E.; Ryde, U. An automatic method to generate force-field parameters for hetero-compounds. *Acta Crystallogr., Sect. D* **2003**, *59*, 274.

(30) Malde, A. K.; Mark, A. E. Challenges in the determination of the binding modes of non-standard ligands in X-ray crystal complexes. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 1.

(31) Ribeiro, A.; Horta, B. A. C.; de Alencastro, R. B. MKTOP: a program for automatic construction of molecular topologies. *J. Braz. Chem. Soc.* **2008**, *19*, 1433.

(32) Wang, J. M.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics* **2006**, *25*, 247.

(33) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269.

(34) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623.

(35) Miller, B. T.; Singh, R. P.; Klauda, J. B.; Hodoscek, M.; Brooks, B. R.; Woodcock, H. L. CHARMMing: A new, flexible web portal for CHARMM. *J. Chem Inf. Model.* **2008**, *48*, 1920.

(36) Schuttelkopf, A. W.; van Aalten, D. M. F. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr., Sect. D* **2004**, *60*, 1355.

(37) Lemkul, J. A.; Allen, W. J.; Bevan, D. R. Practical Considerations for Building GROMOS-Compatible Small-Molecule Topologies. *J. Chem Inf. Model.* **2010**, *50*, 2221.

(38) Sigfridsson, E.; Ryde, U. Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *J. Comput. Chem.* **1998**, *19*, 377.

(39) Singh, U. C.; Kollman, P. A. An approcah to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129.

(40) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General atomic and molecular electronic-structure system. *J. Comput. Chem.* **1993**, *14*, 1347.

(41) Carnall, W. T.; Fields, P. R.; Rajnak, K. Electronic energy levels in trivalent lanthanide aquo ions. I. Pr3+ Nd3+ Pm3+ Sm3+ Dy3+ Ho3 + Er3+ and Tm3+. *J. Chem. Phys.* **1968**, *49*, 4424.

(42) Becke, A. D. Density-functional thermochemistry. 3. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.

(43) Lee, C. T.; Yang, W. T.; Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B* **1988**, *37*, 785.

(44) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation-energy. *Phys. Rev. B* **1992**, *45*, 13244.

(45) Miertus, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilization of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117.

(46) Chirgwin, B. H.; Coulson, C. A. The electronic structure of conjugates systems. 6. *Proc. R. Soc. London, Ser. A* **1950**, *201*, 196.

(47) Stewart, J. J. P. Special issue - MOPAC - a semiempirical molecular-orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

(48) Mulliken, R. S. Electronic population analysis on LCAO-MO molecular wave functions. 1. *J. Chem. Phys.* **1955**, *23*, 1833.

(49) Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **1996**, *60*, 1271.

(50) Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719.

(51) Oostenbrink, C.; Juchli, D.; van Gunsteren, W. F. Amine hydration: A united-atom force-field solution. *ChemPhysChem* **2005**, *6*, 1800.

(52) Gerber, P. R. Charge distribution from a simple molecular orbital type calculation and non-bonding interaction terms in the force field MAB. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 37.

(53) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5*, 350.

(54) Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. *Minnesota Solvation Database*, version 2009; University of Minnesota: Minneapolis, MN, 2009.

(55) Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting small-molecule solvation free energies: An informal blind test for computational chemistry. *J. Med. Chem.* **2008**, *51*, 769.

(56) Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113*, 4501.

(57) Geballe, M. T.; Skillman, A. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 blind prediction challenge: introduction and overview. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 259.

(58) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hunenberger, P. H.; Kruger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulations: The GROMOS96 Manual and User Guide*; Biomos: Zurich, Switzerland, 1996.

(59) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. Interaction models for water in relation to protein hydration. In *Intermolecular Forces*; Reidel: Dordrecht, The Netherlands, 1981; p 331.

(60) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-integration of cartesian equations of motion of a system with constraints - Molecular-dynamics of N-alkanes. *J. Comput. Phys.* **1977**, *23*, 327.

(61) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684.

(62) van Gunsteren, W. F.; Beutler, T. C.; Fraternali, F.; King, P. M.; Mark, A. E.; Smith, P. E. *Computer Simulation of Biomolecular Systems, Theoretical and Experimental Applications*; ESCOM Science Publishers: Leiden, The Netherlands, 1989; p 315.

(63) Allen, M. P.; Tidesley, D. J. *Computer Simulations of Liquids*; Oxford University Press Inc.: New York, 1989.

(64) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and numerical instabilities in free-energy calculations based on molecular simulations. *Chem. Phys. Lett.* **1994**, *222*, 529.

(65) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Seperation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* **1994**, *100*, 9025.