Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

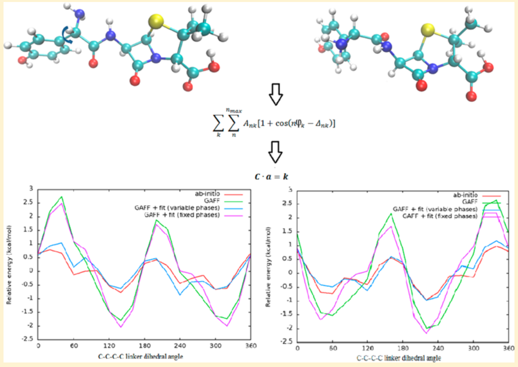
Fitting of Dihedral Terms in Classical Force Fields as an Analytic Linear Least-Squares Problem

Chad W. Hopkins[†] and Adrian E. Roitberg*[,‡]

[†]Department of Physics and [‡]Department of Chemistry, Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States

Supporting Information

ABSTRACT: The derivation and optimization of most energy terms in modern force fields are aided by automated computational tools. It is therefore important to have algorithms to rapidly and precisely train large numbers of interconnected parameters to allow investigators to make better decisions about the content of molecular models. In particular, the traditional approach to deriving dihedral parameters has been a least-squares fit to target conformational energies through variational optimization strategies. We present a computational approach for simultaneously fitting force field dihedral amplitudes and phase constants which is analytic within the scope of the data set. This approach completes the optimal molecular mechanics representation of a quantum mechanical potential energy surface in a single linear least-squares fit by recasting the dihedral potential into a linear function in the parameters. We compare the resulting method to a genetic algorithm in terms of computational time and quality of fit for two simple molecules. As suggested in previous studies, arbitrary dihedral phases are only necessary when modeling chiral molecules, which include more than half of drugs currently in use, so we also examined a dihedral parametrization case for the drug amoxicillin and one of its stereoisomers where the target dihedral includes a chiral center. Asymmetric dihedral phases are needed in these types of cases to properly represent the quantum mechanical energy surface and to differentiate between stereoisomers about the chiral center.



1. INTRODUCTION

Force field parametrization and optimization involves fitting to some target data set by varying the parameters of the force field potential energy function. Relative conformational energies are a common target, particularly when parametrizing the dihedral potential in the force field. Force fields are frequently reparametrized, or specialized for a particular class of molecules, by varying the parameters of important dihedral angles to better agree with high-level ab initio conformational energies about those angles.[1−5] Dihedral parameters mainly serve as corrections to the force field, so dihedral parametrization is generally treated in the final step of the parametrization process.

The base force field form used in many molecular mechanics programs (e.g, Amber,[6] CHARMM,[7] and OPLS[8]) uses a truncated Fourier series to represent the dihedral potential energy for a single dihedral angle θ (note that, in this manuscript, lower case bold letters represent vectors and uppercase bold letters represent matrices):

$$V(\boldsymbol{\alpha}, \boldsymbol{\delta}, \theta) = \sum_{n}^{n_{\max}} \alpha_n[1 + \cos(n\theta - \delta_n)] \tag{1}$$

The multiplicities n that are used vary from case to case, and in theory using more multiplicities allows for a closer fit. In practice, only the first few terms are used (n = 1, 2, 3, 4, 6) for symmetry considerations, decreased complexity, to avoid overfitting, and to avoid introducing additional high-frequency motions into the simulation.[9]

Equation 1 gives the 1-dimensional potential energy for a single dihedral, but the total dihedral potential energy function is a D-dimensional function, where D is the total number of dihedrals in the molecule being studied. There are differing philosophies on how to combine the single dihedral energies into the total dihedral potential energy function. The simplest approach is to decouple the individual dihedrals and treat the total energy function as a sum of the individual potential energy functions:

$$V(\mathbf{A}, \boldsymbol{\Delta}, \boldsymbol{\theta}) = \sum_{k} \sum_{n}^{n_{\max}} A_{nk}[1 + \cos(n\theta_k - \Delta_{nk})] \tag{2}$$

Here, we have written out the sum over each dihedral as the sum over k. θ_k represents the value of the kth dihedral angle, A_{nk} is the force constant for the kth dihedral angle and the nth multiplicity, and Δ_{nk} is the phase constant for the kth dihedral angle and the nth multiplicity. Alternatively, some models (e.g., CMAP[10,11]) introduce lookup tables or cross terms into the

Received: February 21, 2014
Published: June 24, 2014

dx.doi.org/10.1021/ci500112w | J. Chem. Inf. Model. 2014, 54, 1978−1986

force field to couple dihedrals. Our focus is on the simple case of independent Fourier terms for each dihedral (eq 2).

The force field parameters are fit to the target data by optimizing some merit function. One of the most common merit functions used in the literature is the square difference between the potential energy surface (PES) of the original force field about the dihedral angle or angles of interest, and the PES of ab initio energies, taken at the same values of the dihedral angles.[12,13] An additional term is often added to the difference in order to account for the shift in absolute energy levels between the force field and ab initio energies.[14] The (equivalent) goal of these approaches is to minimize (in terms of **A** and **Δ**) a function of the form:

$$\chi^2(\mathbf{A}, \mathbf{\Delta}, \{\mathbf{i}\}) = \sum_i [V(\mathbf{A}, \mathbf{\Delta}, \boldsymbol{\theta}_i) - (E_{\mathrm{QM},i} - E_{\mathrm{MM},i})]^2 \tag{3}$$

This is done for some set of data points $\{i\}$ composed of $\boldsymbol{\theta}_i$, the value for each dihedral angle in data point i, and $E_{\mathrm{QM},i}$ and $E_{\mathrm{MM},i}$, the ab initio energy and the original uncorrected force field energy, respectively, for the configuration $\boldsymbol{\theta}_i$.

The phase constants **Δ** are often fixed to the symmetry positions of 0 or 180°. Some parametrization philosophies prescribe this constraint whenever transferability of parameters is a primary goal.[15] However, asymmetric values of **Δ** can provide a closer fit to the true ab initio energy surface for chiral molecules, especially near the chiral centers. It has also been pointed out that asymmetric dihedral parameters are not necessarily transferable between stereoisomers.[14] Most of the monomers and ligands modeled in biochemical simulations are chiral, including amino acids, nucleosides, carbohydrates, and more than half of current drugs.[16] Therefore, a general dihedral parametrization scheme must allow for the elements of **Δ** to be specified as variable parameters, which leads to the dihedral function form (eq 2) being nonlinear in its parameters. Thus, the problem has traditionally been solved as a nonlinear least-squares fitting.

Authors have used both local and global optimization schemes to reach a solution for the minimum of eq 3. Local schemes, like the Levenberg–Marquardt algorithm[17] and simplex[18] or Newton/quasi-Newton minimization[8,19] find a local minimum to χ^2 relative to some preset starting point. When varying **Δ**, eq 3 is a complicated function in parameter space with multiple minima,[14] so for this case local schemes cannot always find the optimal parameter set, even after fully converging. As local schemes cannot optimally solve the dihedral parametrization problem (with variable **Δ**) in general, we will not consider them further. Global schemes such as genetic algorithms, systematic search,[20] and Monte Carlo simulated annealing[14,21−23] can find an approximate optimal parameter set after full convergence, even when varying **Δ**, though, as will be shown for genetic algorithms, the required run time for convergence of global optimizers will typically not be negligible, and convergence still requires an arbitrary stopping condition that can lead to further uncertainty in the resulting parameters. For the special case when the **Δ** values are fixed to 0 or 180°, it has been pointed out that the potential function (eq 2) is linear in the parameters and thus is subject to general linear least-squares (LLS) fitting.[14,24] This is a well-known problem and there exists an analytic solution for calculating the optimal parameter set **A** in this case. As far as we are aware, an analytic LLS scheme has not previously been applied to the simultaneous parametrization of **A** and **Δ**, though.

It should be pointed out that the largest computational cost in dihedral parametrization for all but very small model systems will obviously be the calculation of the ab initio energies ($E_{\mathrm{QM},i}$). However, it will be shown that using optimization schemes, in particular the global optimizers, for the relatively inexpensive fitting calculation adds unnecessary additional computational cost to the overall process. More importantly, it adds a level of uncertainty to the quality of the fitted parameters. We will show that it is possible to treat the full dihedral parametrization problem, along with all of the typical variants (fixed phases, weighted data points, multiple dihedrals, harmonic restraints on parameters), in a single general LLS fitting scheme, solving analytically for the optimal parameter set.

2. LINEAR LEAST-SQUARES FITTING OF DIHEDRAL PARAMETERS

The constant terms in each term of the sum in eq 2 can be combined, which gives a single constant term (independent of $\boldsymbol{\theta}$):

$$\begin{aligned} V(\mathbf{A}, \mathbf{\Delta}, \boldsymbol{\theta}) &= \sum_k \sum_n^{n_{\max}} A_{nk} + \sum_k \sum_n^{n_{\max}} A_{nk} \cos(n\theta_k - \Delta_{nk}) \\ &= C(\mathbf{A}, \mathbf{\Delta}) + V'(\mathbf{A}, \mathbf{\Delta}, \boldsymbol{\theta}) \end{aligned}$$

C can be dropped, as a constant in the energy expression does not affect the dynamics, and V' can be used for fitting. The potential energy function, for fitting purposes, then becomes (dropping the prime):

$$V(\mathbf{A}, \mathbf{\Delta}, \boldsymbol{\theta}) = \sum_k \sum_n^{n_{\max}} A_{nk} \cos(n\theta_k - \Delta_{nk}) \tag{4}$$

This change will not alter the set of parameters that best puts the force field PES in agreement with the ab initio PES. To avoid any numerical issues with the (usually large) inherent energy level difference between force field and ab initio energies, the sets $E_{\mathrm{QM},i}$ and $E_{\mathrm{MM},i}$ should be shifted so that their respective averages are zero before performing the fitting calculation.

Equation 4 can be linearized by rewriting it in terms of a new set of parameters. Through a trivial trigonometric transformation, we first expand:

$$\begin{aligned} V(\mathbf{A}, \mathbf{\Delta}, \boldsymbol{\theta}) = \sum_k \sum_n [&A_{nk} \cos(\Delta_{nk}) \cos(n\theta_k) \\ &+ A_{nk} \sin(\Delta_{nk}) \sin(n\theta_k)] \end{aligned}$$

We then combine terms to create two new parameter sets:

$$V(\mathbf{A}', \mathbf{B}', \boldsymbol{\theta}) = \sum_k \sum_n [A'_{nk} \cos(n\theta_k) + B'_{nk} \sin(n\theta_k)] \tag{5}$$

$$\begin{cases} A'_{nk} = A_{nk}\cos(\Delta_{nk}) \\ B'_{nk} = A_{nk}\sin(\Delta_{nk}) \\ A_{nk} = \sqrt{(A'_{nk})^2 + (B'_{nk})^2} \\ \Delta_{nk} = \tan^{-1}\left(\dfrac{B'_{nk}}{A'_{nk}}\right) \end{cases} \tag{6}$$

(Note that the last two lines of eq 6 are simply derived from the first two lines). Equation 5 is linear in its parameters and subject to LLS fitting. By solving for the parameters **A**′ and ***B***′, **A** and **Δ** can then be recovered via eq 6.

As previously stated, the linear least-squares method is a well-known solution to fitting a linear function form to data in a least-squares sense. A complete derivation of the method can be found elsewhere.[25] We will simply present here the final solution for the so-called "normal equations" method as applied to the dihedral parametrization problem. As discussed below, this is not the best solution to the LLS problem from a computational standpoint, but it is intuitive and easily implemented. The fitting problem can be expressed as a linear system of equations:

$$\mathbf{C}\cdot\mathbf{a} = \mathbf{k} \tag{7}$$

Here, the vector **a** is the parameter set being solved for:

$$a_n = \begin{cases} A'_{n'k} & \text{if } 1 \le n \le dn_{\max} \\ B'_{n'k} & \text{if } n > dn_{\max} \end{cases} \tag{8}$$

The matrix **C** is a geometric factor defined in terms of the configurations in the data set (the "predictor" matrix), and the vector **k** is defined in terms of the configurations and energies (the "response" vector):

$$C_{no} = C_{on} = \begin{cases} \sum_i \cos(n'\theta_{ki})\cos(o'\theta_{ji}) & \text{if } 1 \le n,\ o \le dn_{\max} \\ \sum_i \cos(n'\theta_{ki})\sin(o'\theta_{ji}) & \text{if } 1 \le n \le dn_{\max} \\ & \text{and } o > dn_{\max} \\ \sum_i \sin(n'\theta_{ki})\sin(o'\theta_{ji}) & \text{if } n,\ o > dn_{\max} \end{cases} \tag{9}$$

$$k_n = \begin{cases} \sum_i \cos(n'\theta_{ki})(E_{\mathrm{QM},i} & \text{if } 1 \le n \le dn_{\max} \\ \quad - E_{\mathrm{MM},i}) & \\ \sum_i \sin(n'\theta_{ki})(E_{\mathrm{QM},i} & \text{if } n > dn_{\max} \\ \quad - E_{\mathrm{MM},i}) & \end{cases} \tag{10}$$

C is symmetric, and **C**, **a**, and **k** have dimension $2dn_{\max}$. In eqs 8−10, d is the number of dihedrals being fit, $n_{\max}$ is the number of multiplicities being fit, and $E_{\mathrm{QM},i}$, $E_{\mathrm{MM},i}$, and θ_{ki} are the ab initio energy, original force field energy, and value for dihedral angle k, respectively, from the ith data point. The variables n' and k are defined in terms of the index n, with o' and j having analogous definitions in terms of the index o:

$$\begin{cases} n' = [(n-1)\bmod n_{\max}] + 1 \\ k = \left(\left\lfloor \dfrac{(n-1)}{n_{\max}} \right\rfloor\right)\bmod d + 1 \end{cases} \tag{11}$$

For a sufficient number of data points (more than the number of fitted parameters, $2dn_{\max}$), and nonpathological data, eq 7 can be solved for **a** using standard matrix techniques. Again, **a** contains the terms from the alternate Fourier coefficients **A**′ and **B**′, which can be converted back to **A** and **Δ** via eq 6.

It should be pointed out that this is a rather naïve solution to the general LLS problem, as it is susceptible to round off error, as well as unsolvable cases (i.e., when **C** is singular, as can happen when using an insufficient number of data points). More sophisticated standard techniques (employing, e.g., QR decomposition or SVD) for solving the general LLS problem can be found elsewhere.[25] We want to emphasize again that *at least* $2dn_{\max}$ data points (that is, points on the dihedral "scan" used) are needed in this implementation, and typically more than this number are needed to properly sample the ab initio energy surface. Using too small of a data set will return spurious results that do not represent the optimal parameter set.

A few modifications can be made to eqs 9 and 10 to allow for some oft-used features of the fitting process. Many parametrization schemes employ data set weighting, making certain data points count more toward the merit function than others (e.g., with Boltzmann weighting).[24,26] If the weight for data point i is w_i, then each occurrence of $\cos(n\theta_{ki})$ or $\sin(n\theta_{ki})$ in **k** and each product of $\cos(n\theta_{ki})$ and/or $\sin(n\theta_{ki})$ in **C** would be multiplied by w_i. Also, sometimes two or more dihedral angles are being fit that have exactly the same atom type sequence, and thus, should be assigned the same dihedral parameters. This "equivalencing" of dihedrals can also be handled by replacing each occurrence of $\cos(n\theta_{ki})$ or $\sin(n\theta_{ki})$ in **C** and **k** by a sum of cosines or sines, where the sum is over dihedrals of the same type. It is even possible to incorporate harmonic restraints on the initial parameters in order to avoid obtaining unphysical parameters; these restraints add quadratic terms to χ^2, and so lead to simple modifications of **C** and **k**. A pseudocode is provided in the Supporting Information for this paper that incorporates all of these cases, and we have a script on our Web site (http://www.clas.ufl.edu/users/roitberg/links.html, under "Software") that implements the pseudocode. The pseudocode and script also make it possible to restrain the **Δ** values exactly to 0 or 180°. The phase constants can be removed from the fitting function, giving

$$V(\mathbf{A}, \boldsymbol{\theta}) = \sum_k \sum_n^{n_{\max}} A_{nk}\cos(n\theta_k)$$

If just the set **A** is treated here, Δ_{nk} is implicitly fit to either 0 or 180°, with the result being expressed through the sign of A_{nk} (0 if positive, 180° if negative). This case is specified in our script through an input flag.

3. METHODS

All molecular mechanics (MM) calculations were performed using the Amber 12 suite.[27] We used the nonbonding, bond, and angle parameters of GAFF[28] in all calculations. Atomic charges were calculated using the AM1-BCC method in the antechamber program in Amber.[29−31] It is possible to start from any reasonable "initial guess" or starting point (i.e., values used in generating the initial MM energy surface) for the dihedral parameters being fit. Neglecting round-off differences, any starting point will lead to the same parameter set; however, the starting point must be taken into account when calculating the final parameter set. In our tests, we used GAFF dihedral parameters to generate the initial MM surfaces. Using eq 6, the GAFF dihedral force constants $\mathbf{A}_{\mathrm{GAFF}}$ and phase constants $\mathbf{\Delta}_{\mathrm{GAFF}}$ are converted to the sets $\mathbf{A}'_{\mathrm{GAFF}}$ and $\mathbf{B}'_{\mathrm{GAFF}}$. The parameters obtained from solving eq 7 will be denoted $\mathbf{A}'_{\mathrm{fit}}$ and $\mathbf{B}'_{\mathrm{fit}}$. The final presented dihedral parameter sets for the LLS method ($\mathbf{A}_{\mathrm{LLS}}$ and $\mathbf{\Delta}_{\mathrm{LLS}}$) are then obtained by converting, via eq 6, $\mathbf{A}'_{\mathrm{LLS}}$ and $\mathbf{B}'_{\mathrm{LLS}}$, which are given by

$$\mathbf{A}'_{\mathrm{LLS}} = \mathbf{A}'_{\mathrm{GAFF}} + \mathbf{A}'_{\mathrm{fit}}$$

$$\mathbf{B}'_{\mathrm{LLS}} = \mathbf{B}'_{\mathrm{GAFF}} + \mathbf{B}'_{\mathrm{fit}}$$

Optimized geometries and quantum mechanics (QM) energies were calculated using Gaussian 09.[32] The QM optimized geometries were then used to calculate the MM PES for each molecule, by performing a single point MM energy calculation for each structure. As has been pointed out, a further constrained MM optimization usually should be performed to produce the MM PES.[33] However, our main goal here is to investigate the LLS fitting method, not to produce extremely reliable parameters for the test cases.

Validation. In order to validate our implementation of the LLS fitting method and compare the method to a traditional optimization technique, we first calculate dihedral parameters for two simple, achiral molecules: butane (Figure 1a) and 1-butanol (Figure 1b). This is done with the presented LLS method and with a genetic algorithm (GA) applied in the "traditional" way. By this, we mean that the GA merit function is eq 3 with $V(\mathbf{A}, \mathbf{\Delta}, \boldsymbol{\theta}_i)$ given by the original nonlinear form (eq 2) so that $\mathbf{A}$ and $\mathbf{\Delta}$ are directly being varied. These two cases are compared in terms of the resulting parameter sets and run times for each method.

Geometry optimizations and single point calculations were performed at the MP2/6-31G* level on butane and 1-butanol. The 1-dimensional QM PES of butane was created from a 72-point relaxed scan of the C−C−C−C dihedral angle, with each point spaced by 5°. The QM PES of 1-butanol was calculated from a 2-dimensional 72-by-72 point grid (5184 data points). We note that all other dihedrals besides the ones being fit were allowed to move freely during the optimizations, including the C−C−O−H dihedral of 1-butanol. This dihedral in particular led to small "jumps" during the scan, as it moved over a large range between two consecutive steps of the scan, which can be seen as discontinuities in the resulting energy maps. A more appropriate data collection scheme would be to take the free dihedrals into better account, either restraining them or using them as a target in the fit. However, these discontinuities have a minimal effect on the fit for this simple case, and again our focus is on the fitting procedure itself, and not on deriving high quality parameters for 1-butanol.

The pyevolve[34] Python module was used for the GA fit. The normal equations solution was implemented in Perl, and all matrix calculations were done with the Math::MatrixReal module. All fitting calculations were done on a quad core laptop PC with Intel i7 processors and 8 GB of memory. One parameter set for each molecule, fitting both $\mathbf{A}$ and $\mathbf{\Delta}$, was calculated with the GA, denoted as $\mathrm{GAFF}_{\mathrm{GA}}$. For LLS, two sets of parameters for each molecule were calculated. In one set ($\mathrm{GAFF}_{\mathrm{LLS,var.}}$), both $\mathbf{A}$ and $\mathbf{\Delta}$ were fit as in $\mathrm{GAFF}_{\mathrm{GA}}$, and in the other ($\mathrm{GAFF}_{\mathrm{LLS,fix}}$), $\mathbf{\Delta}$ was fixed to be either 0 or 180° and only $\mathbf{A}$ was fit. Butane and 1-butanol are both achiral, so the latter case would be preferred for calculating physically accurate parameters, if one were actually modeling butane or 1-butanol. Four terms in the Fourier series for each dihedral ($n_{\mathrm{max}} = 4$) were fit.

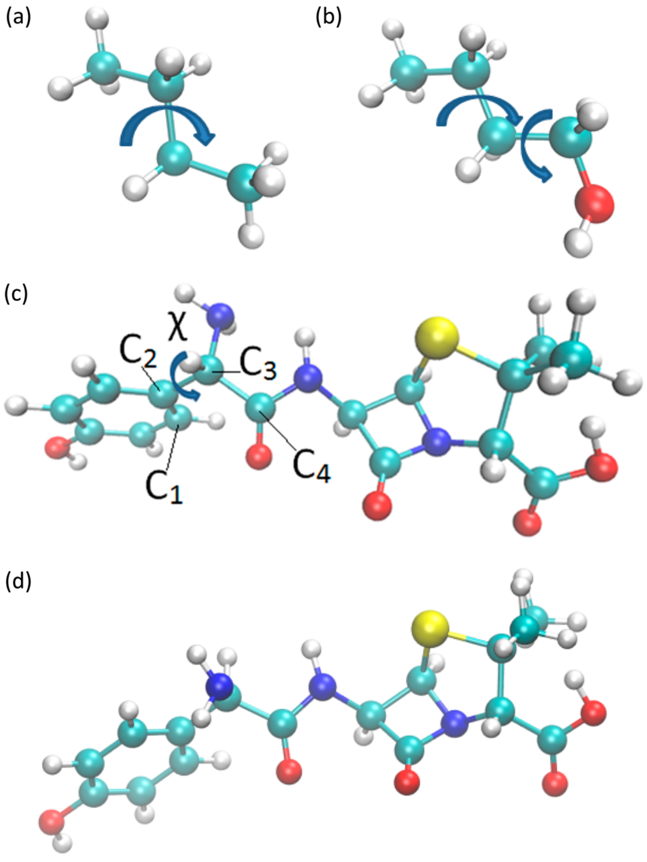


Figure 1. Test systems for the LLS method: (a) butane, (b) 1-butanol, (c) amoxicillin, and (d) stereoisomer of amoxicillin about C_3. A 1-dimensional scan was performed on butane and used to fit the C−C−C−C dihedral parameters, and for 1-butanol a 2-dimensional scan was used to simultaneously fit the C−C−C−C and C−C−C−O dihedral parameters. For the two stereoisomers, a 1-dimensional scan of the dihedral marked χ in part c was performed to fit the corresponding C−C−C−C parameters.

Chiral Molecule Parameterization. In order to investigate a realistic case where variable $\mathbf{\Delta}$ would be needed, we also derived dihedral parameters for a linker dihedral in the drug amoxicillin (DrugBank ID DB01060) (Figure 1c). The dihedral χ contains the chiral center C_3, and as previously stated, dihedral parameters are not expected to be transferable between stereoisomers, especially for dihedrals near a chiral center. Thus, a separate parametrization routine was run for χ in the stereoisomer of amoxicillin about C_3 (Figure 1d). For each case, QM calculations were done at the MP2/6-31G*//HF/6-31G* level of theory. The 1-dimensional QM PES for each molecule was created from an 18-point relaxed scan of χ, with each point spaced by 20°. For the two stereoisomers, LLS was used to parametrize χ, again with variable phases ($\mathrm{GAFF}_{\mathrm{LLS,var.}}$) and fixed phases ($\mathrm{GAFF}_{\mathrm{LLS,fix}}$). The presented parameters were applied to the two C−C−C−C (c-c3-ca-ca. in GAFF) dihedrals about the C_2−C_3 bond in order to generate the new energy profiles, though there are really six dihedrals (and three unique dihedral types) about the bond that contribute to the profile, and in a real parametrization session, the parameters should be split up over the other two types as well. Again, four terms in the Fourier series for each dihedral (n_{max} = 4) were fit.

4. RESULTS AND DISCUSSION

Validation—Butane. For butane, the GAFF$_{LLS}$ sets and GAFF$_{GA}$ yielded nearly identical results (Table 1). The difference between GAFF$_{LLS,var}$ and GAFF$_{LLS,fix}$ was negligible; there were no major degrees of freedom unaccounted for during the dihedral scan, so the inherent symmetry was maintained even when the phases were allowed to vary. The RMSE between the MM PES and the QM PES was improved identically in all of the fitting cases. Figure 2 shows a comparison of the resulting potential energy scans for both methods to the ab initio scan.

Table 1. Butane C−C−C−C (c3−c3−c3−c3 in GAFF) Dihedral Parameters in GAFF, GAFF$_{GA}$, and GAFF$_{LLS}$ Parameter Sets and RMSE between Force Field Energies and ab Initio Energies

	GAFF	GAFF$_{GA}$	GAFF$_{LLS,var}$	GAFF$_{LLS,fix}$
	force constants (kcal/mol)			
A_1	0.200	0.071	0.071	0.074
A_2	0.250	0.048	0.048	0.045
A_3	0.180	0.347	0.347	0.350
A_4	0.000	0.171	0.170	0.173
	phase constants (degrees)			
δ_1	180.000	0.548	0.511	0.000
δ_2	180.000	−179.547	−179.711	180.000
δ_3	0.000	−0.221	0.010	0.000
δ_4	0.000	0.084	0.011	0.000
RMSE (kcal/mol)	0.302	0.073	0.073	0.073

The GA started finding reasonable solutions after 50 generations, which took about a second to run. However, it was necessary to run longer for the force constants in the two sets of parameters to agree within the order of 10^{-3} kcal/mol; it was sufficient to run the GA for 1000 generations with a population of 200, which took 25.3 s to complete. The LLS calculation took 0.1 s to complete.

Validation—1-Butanol. For the 1-butanol test, a large slowdown in the GA method was apparent. The GA took about 400 generations in this case to start finding optimal solutions due to the doubling of the parameter space dimensionality, and each generation took longer to run with the increase in data points. To run the same number of generations as the previous example (1000 generations with population 200), took about 45 min, while the LLS calculation still took a negligible amount of time to complete (0.8 s). It should be pointed out that this particular scan (72 × 72 grid) is a much higher resolution than is needed to obtain numerically accurate parameters in this case, which contributed to the long run time of the GA. We reduced the grid down to 18 × 18 and obtained similar parameters, usually within 10^{-2} kcal/mol for the force constants. Still, data sets of thousands of points would start to be necessary if three or more dihedrals were being simultaneously treated.

Figure 3 shows the 2D potential energy scans for the ab initio energies, the original GAFF energies, the GAFF$_{GA}$ energies, and the GAFF$_{LLS,var}$ energies. Tables 2 and 3 show that, again, the derived parameter sets were very similar among all the fitting methods, and the RMSE for the MM PES with fitted parameters was reduced by a similar amount for each case, with GAFF$_{LLS,var}$ giving the closest fit by a small margin. However, in this case, allowing $\mathbf{\Delta}$ to take on arbitrary values in GAFF$_{LLS,var}$ introduced a slight asymmetry which is not present in the achiral structure of 1-butanol. In general, these kinds of artifacts can appear due to biases in the configurations chosen in the data set and from not taking into account other major degrees of freedom during the relaxed scan. In this case, the data set is sufficiently high resolution and uniformly distributed over dihedral space that we can conclude that data set bias is not the cause of the artifacts. The most likely reason for these asymmetric $\mathbf{\Delta}$ values is the aforementioned C−C−O−H dihedral which was not taken into account during the scan. The discontinuities it introduced into the energy surfaces from "jumping" during the scan obviously introduced a bias toward one region of the energy surface which led to an asymmetric potential giving a numerically better fit. We predict that if this dihedral were also treated in a new 3-dimensional fit, these artifacts would vanish. This was a minimal effect; compared to GAFF$_{LLS,fix}$, the calculated values of $\mathbf{A}$ are nearly identical, and the $\mathbf{\Delta}$ values in GAFF$_{LLS,var}$ do not differ very much from 0 or 180°. The biggest offender was the δ_2 term for the C−C−C−C dihedral; however, the corresponding force constant is effectively zero, and so any phase constants calculated along with it should be considered spurious anyway. Every other phase constant was within 15° of 0 or 180°. Still, we repeat our earlier warning: these parameters (including the more reasonable GAFF$_{LLS,fix}$ set) are not meant to be taken as physically realistic parameters for modeling butane or 1-butanol. Besides the presence of artifacts, isolated dihedral

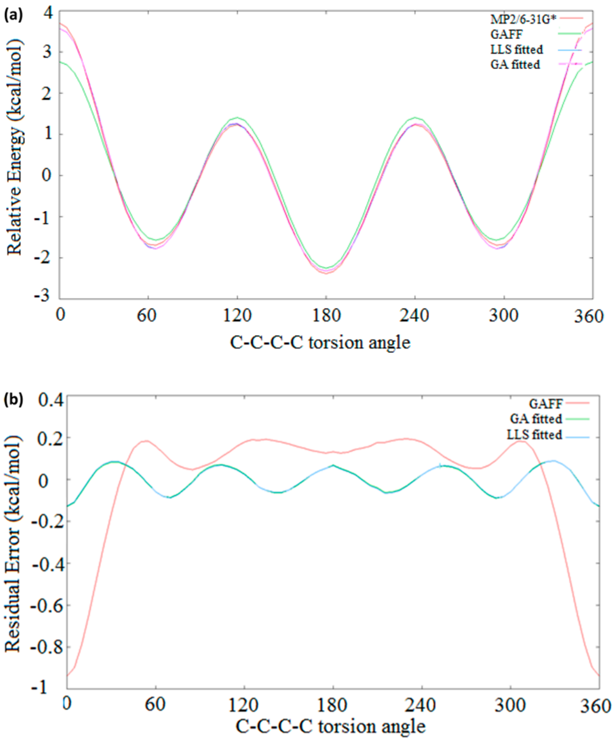


Figure 2. (a) C−C−C−C dihedral potential energy scans in butane. Comparison of ab initio (MP2/6-31G*) energies and force field energies with GAFF, GAFF$_{GA}$, and GAFF$_{LLS,var}$ parameter sets. The averages for each energy curve were set to zero before fitting was done. (b) Residual errors between the scans with the three parameter sets and the ab initio scan.

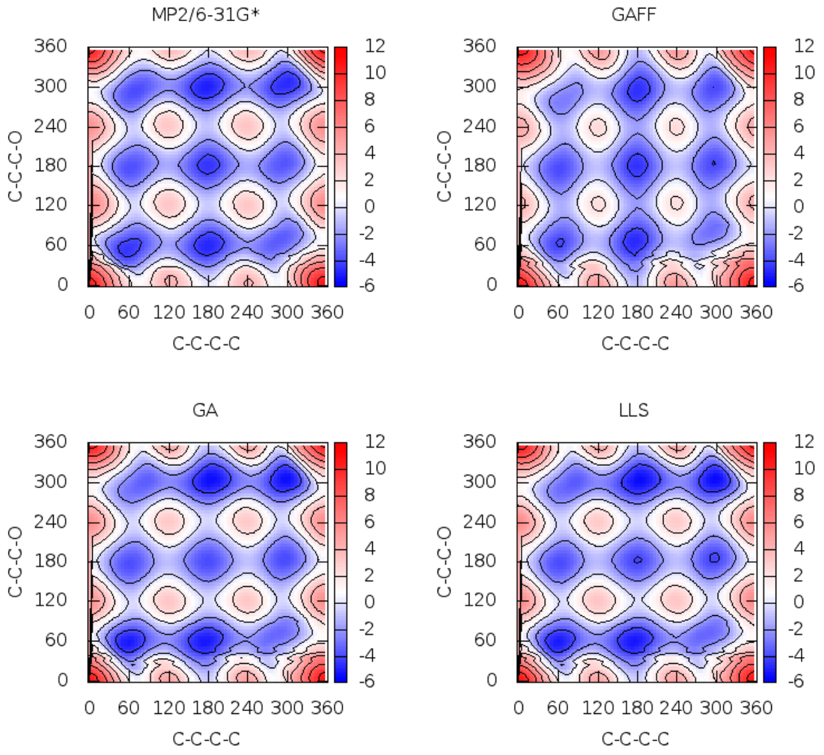


Figure 3. 2D potential energy scans of C−C−C−C and C−C−C−O dihedrals in 1-butanol. Energies are in kilocalories per mole. The scan was performed over a 72 × 72 grid, with a spacing of 5°. Energies were evaluated for MP2/6-31G* optimized geometries, and the individual plots are labeled with the type of single point energy evaluation (MP2/6-31G*, GAFF, $GAFF_{GA}$, or $GAFF_{LLS,var}$ parameters).

Table 2. 1-Butanol C−C−C−C (c3−c3−c3−c3 in GAFF) Dihedral Parameters in GAFF, $GAFF_{GA}$, and $GAFF_{LLS}$ Parameter Sets and RMSE between ab Initio Energies and Force Field Energies Calculated with the Corresponding Dihedral Parameter Set

	GAFF	$GAFF_{GA}$	$GAFF_{LLS,var}$	$GAFF_{LLS,fix}$
	force constants (kcal/mol)			
A_1	0.200	0.201	0.214	0.214
A_2	0.250	0.071	0.002	0.001
A_3	0.180	0.103	0.112	0.109
A_4	0.000	0.142	0.132	0.130
	phase constants (degrees)			
δ_1	180.000	−176.857	179.209	180.000
δ_2	180.000	116.177	−75.661	0.000
δ_3	0.000	−11.176	12.913	0.000
δ_4	0.000	−7.356	−11.629	0.000
RMSE (kcal/mol)	0.981	0.491	0.450	0.500

Table 3. 1-Butanol C−C−C−O (c3−c3−c3−oh in GAFF) Dihedral Parameters in GAFF, $GAFF_{GA}$, and $GAFF_{LLS}$ Parameter Sets and RMSE between ab Initio Energies and Force Field Energies Calculated with the Corresponding Dihedral Parameter Set

	GAFF	$GAFF_{GA}$	$GAFF_{LLS,var}$	$GAFF_{LLS,fix}$
	force constants (kcal/mol)			
A_1	0.000	0.856	0.843	0.825
A_2	0.000	0.224	0.272	0.272
A_3	0.156	0.880	0.908	0.880
A_4	0.000	0.234	0.261	0.251
	phase constants (degrees)			
δ_1	0.000	169.312	168.122	180.000
δ_2	0.000	177.000	178.418	180.000
δ_3	0.000	13.267	14.270	0.000
δ_4	0.000	24.185	15.702	0.000
RMSE (kcal/mol)	0.981	0.491	0.450	0.500

scans, without taking into further account environment or solvation effects, can be of limited usefulness as the optimal dihedral potential is potentially sensitive to the surrounding environment.[33] The most significant result for our purposes here is the fact that $GAFF_{LLS,var}$ and $GAFF_{GA}$ are the same asymmetric parameter set (neglecting the spurious δ_2 values), with $GAFF_{LLS,var}$ being a slightly better numerical fit calculated in a negligible run time.

Chiral Molecule Parameterization. Table 4 shows the resulting parameters and residual RMSEs when fitting the C−C−C−C parameters in the χ dihedral in amoxicillin with the LLS method, and Figure 4a shows the resulting energy profiles of χ for the various parameter sets and ab initio energies. In this case, there is a significant difference between $GAFF_{LLS,var}$ and $GAFF_{LLS,fix}$. Clearly, $GAFF_{LLS,fix}$ was not able to improve significantly on GAFF, which does not have any parameters for this dihedral. GAFF overestimated the rotation barriers for this dihedral by 2−3 kcal/mol, and by fixing $\boldsymbol{\Delta}$ to 0 or 180°, no improvement was made as the profile minima lie close to 120° and 300°. The values of $\boldsymbol{\Delta}$ in $GAFF_{LLS,var}$, with the exception of the smallest $n = 3$ term, vary significantly from 0 or 180°, which brings the MM energy profile into much closer agreement with the QM profile, reducing the RMSE by about a factor of 5.

Table 5 and Figure 4b show the results of parametrizing the same dihedral in one of amoxicillin's stereoisomers. Exactly the same disparity between $GAFF_{LLS,var}$ and $GAFF_{LLS,fix}$ as for amoxicillin is seen here. As predicted, the parameters derived for amoxicillin are not directly transferable to its stereoisomer.

Table 4. c−c3−ca−ca Dihedral Parameters Fitted via χ in Amoxicillin for GAFF, $GAFF_{LLS,var}$, and $GAFF_{LLS,fix}$ Parameter Sets and RMSE between ab Initio Energies and Force Field Energies Calculated with the Corresponding Dihedral Parameter Set

	GAFF	$GAFF_{LLS,var}$	$GAFF_{LLS,fix}$
	force constants (kcal/mol)		
A_1	0.000	0.128	0.019
A_2	0.000	0.691	0.120
A_3	0.000	0.085	0.100
A_4	0.000	0.173	0.110
	phase constants (degrees)		
δ_1	0.000	−103.747	180.000
δ_2	0.000	−101.034	180.000
δ_3	0.000	−3.038	0.000
δ_4	0.000	−55.354	0.000
RMSE (kcal/mol)	1.008	0.220	0.986

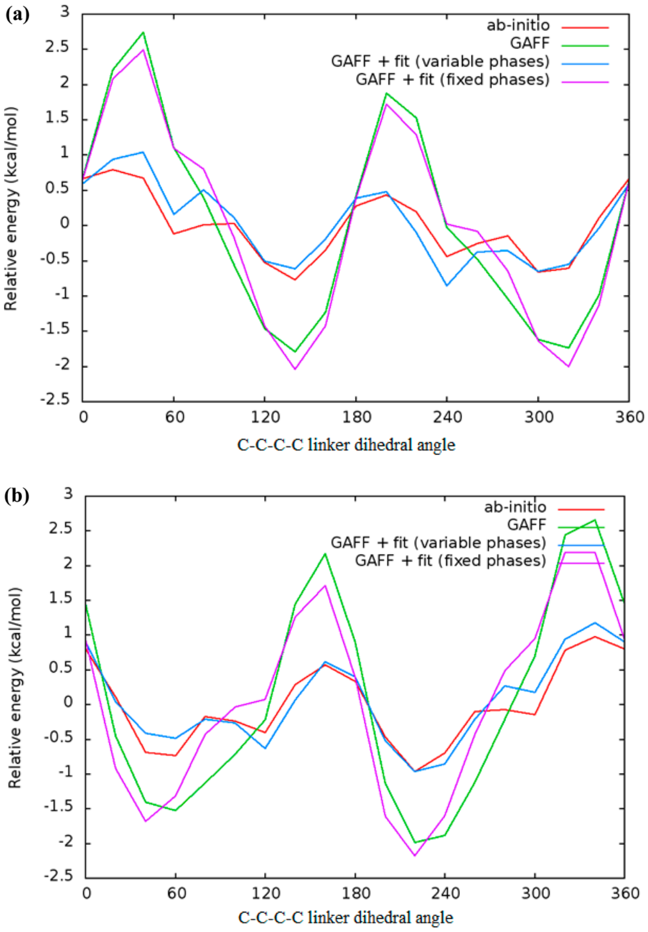


Figure 4. Dihedral potential energy scans for (a) amoxicillin and (b) an amoxicillin stereoisomer about the dihedral χ. Energies were calculated for HF/6-31G* optimized structures, with MP2/6-31G*, GAFF, $GAFF_{LLS,var}$, and $GAFF_{LLS,fix}$ energies.

The **A** values are approximately identical between the two cases, but the optimal values of **Δ**, apart from the small $n = 3$ term, are shifted by about 120 ± 20° between the stereoisomers. Some of the neglected parametrization considerations brought up for the butane and 1-butanol tests apply here as well. In particular, the other dihedrals along the linking chain between the ring systems were not accounted for. Movements in these dihedrals during the relaxed scan probably contributed to some small features in the QM surface; it would be preferable to include all of these dihedrals into a single multidimensional fit.

Table 5. c−c3−ca−ca Dihedral Parameters Fitted via χ in an Amoxicillin Stereoisomer for GAFF, $GAFF_{LLS,var}$, and $GAFF_{LLS,fix}$ Parameter Sets and RMSE between ab Initio Energies and Force Field Energies Calculated with the Corresponding Dihedral Parameter Set

	GAFF	$GAFF_{LLS,var}$	$GAFF_{LLS,fix}$
	force constants (kcal/mol)		
A_1	0.000	0.090	0.073
A_2	0.000	0.667	0.318
A_3	0.000	0.083	0.092
A_4	0.000	0.150	0.062
	phase constants (degrees)		
δ_1	0.000	157.136	180.000
δ_2	0.000	119.390	180.000
δ_3	0.000	−3.712	0.000
δ_4	0.000	69.699	0.000
RMSE (kcal/mol)	0.973	0.177	0.848

5. CONCLUSION

Modern force fields are constantly being optimized to more closely represent physical reality. In order for the coarse MM model to have any validity, high quality parameters derived through a rigorous process and validated against high level QM and experimental data are necessary. However, validation stretches far beyond the initial release; as new classes of systems are studied, and as hardware and algorithmic improvements allow the treatment of larger models with QM calculations, force fields need to frequently be updated to best reflect the latest knowledge in the field. Many of the challenges faced in force field parametrization do not have a clear, exact solution, and so there exists a multitude of techniques to approximate one. Often, a targeted physical property does not have a closed functional form in terms of the force field parameters, or the functional form is truly nonlinear (e.g., vibrational spectra, geometries, bonds, and angles). In these cases, nonlinear fitting techniques or heuristic optimizations are needed in order to arrive at a parameter set that best fits the target data. We have shown that the usual method of deriving the dihedral parameters, through a least-squares fit to conformational energies with a truncated Fourier series, is somewhat of a special case in that it does have an "easy", exact solution. The LLS method presented herein provides an analytic solution for simultaneously fitting dihedral force constants and phases, along with all of the common variants of dihedral parametrization.

Using LLS to solve the dihedral problem is clearly superior to any optimization technique, in general. However, the real improvement in run time and parameter quality over the traditional techniques will be seen when simultaneously treating multiple dihedrals while varying the phases. As shown, variable phases are clearly needed when modeling dihedrals near a chiral center. As the majority of molecules of biochemical interest are chiral, variable phases are an important consideration for the dihedral parametrization process. The remaining question here is then the extent of the utility of multidimensional fits for dihedral parametrization. The literature mostly employs 2-dimensional fits (e.g., φ/ψ fits in amino acids), and these are mostly used in conjunction with coupling methods like CMAP. Parameters will not be transferable when treating more dihedrals simultaneously, but a much better representation of the true QM PES for a particular model could outweigh other considerations. The main barrier to simultaneously treating many dihedrals is the need for more ab initio calculations in order to properly sample the QM PES. If the usual "dihedral scan" notion is directly translated to higher dimensions, then the number of ab initio calculations grows *exponentially* with the number of dihedrals being fit. However, the number of data points that the fitting problem needs to be overdetermined is *linear* in the number of dihedrals being fit, though it is still not clear how many additional points will be needed to appropriately sample the QM PES. Using alternative methods of dihedral parametrization apart from constrained scans has been the focus of recent work.[35] This investigation of methods of data collection for dihedral parametrization other than a complete multidimensional grid, and of how many ab initio calculations are really needed to find the best representation of the QM PES in terms of the dihedral potential, is a direction for further study.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Pseudocode for the linear least-squares solution of the dihedral parametrization fitting problem. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: roitberg@ufl.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge financial support from the NSF (No. CHE-0822935). The authors thank the High-Performance Computing Center at the University of Florida for providing computational resources.

■ REFERENCES

(1) Wu, J.; Chattree, G.; Ren, P. Automation of AMOEBA polarizable force field parameterization for small molecules. *Theor. Chem. Acc.* **2012**, *131*, 1−11.

(2) Yildirim, I.; Stern, H.; Kennedy, S.; Tubbs, J.; Turner, D. Reparameterization of RNA χ torsion parameters for the AMBER force field and comparison to NMR spectra for cytidine and uridine. *J. Chem. Theory Comput.* **2010**, *6*, 1520−1531.

(3) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712−725.

(4) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins.* **2010**, *78*, 1950−1958.

(5) Meagher, K. L.; Redman, L. T.; Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J. Comput. Chem.* **2003**, *24*, 1016−1025.

(6) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(7) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(8) 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*, 11225−11236.

(9) Hopfinger, A.; Pearlstein, R. Molecular mechanics force-field parameterization procedures. *J. Comput. Chem.* **1984**, *5*, 486−499.

(10) MacKerell, A.; Feig, M.; Brooks, C. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400−1415.

(11) MacKerell, A.; Feig, M.; Brooks, C. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698−699.

(12) Ermer, O. Calculation of molecular properties using force fields. Applications in organic chemistry. *Struct. Bonding (Berlin)* **1976**, *27*, 161−211.

(13) Liang, G.; Fox, P. C.; Bowen, J. P. Parameter analysis and refinement toolkit system and its application in MM3 parameterization for phosphine and its derivatives. *J. Comput. Chem.* **1996**, *17*, 940−953.

(14) Guvench, O.; MacKerell, A. Automated conformational energy fitting for force-field development. *J. Mol. Model.* **2008**, *14*, 667−679.

(15) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671−690.

(16) Nguyen, L. A.; He, H.; Pham-Huy, C. Chiral drugs: an overview. *Int. J. Biomed. Sci.* **2006**, *2*, 85−100.

(17) Reith, D.; Kirschner, K. A modern workflow for force-field development−Bridging quantum mechanics and atomistic computational models. *Comput. Phys. Commun.* **2011**, *182*, 2184−2191.

(18) Norrby, P. O.; Liljefors, T. Automated molecular mechanics parameterization with simultaneous utilization of experimental and quantum mechanical data. *J. Comput. Chem.* **1998**, *19*, 1146−1166.

(19) Waldher, B.; Kuta, J.; Chen, S.; Henson, N.; Clark, A. ForceFit: a code to fit classical force fields to quantum mechanical potential energy surfaces. *J. Comput. Chem.* **2010**, *31*, 2307−2316.

(20) Wang, J.; Kollman, P. Automatic parameterization of force field by systematic search and genetic algorithms. *J. Comput. Chem.* **2001**, *22*, 1219−1228.

(21) Zhu, X.; Lopes, P.; MacKerell, A. Recent developments and applications of the CHARMM force fields. *Wiley Interdis. Rev.: Comput. Molec. Sci.* **2012**, *2*, 167−185.

(22) Mayne, C. G.; Saam, J.; Schulten, K.; Tajkhorshid, E.; Gumbart, J. C. Rapid parameterization of small molecules using the force field toolkit. *J. Comput. Chem.* **2013**, *34*, 2757−2770.

(23) Lopes, P. E.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D., Jr Polarizable Force Field for Peptides and Proteins based on the Classical Drude Oscillator. *J. Chem. Theory Comput.* **2013**, *9*, 5430−5449.

(24) Kaminski, G.; Friesner, R.; Tirado-Rives, J.; Jorgensen, W. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.

(25) Press, W.; Flannery, B.; Teukolsky, S.; Vetterling, W. *Numerical Recipes in FORTRAN 77: Vol. 1, Vol. 1 of Fortran Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press, 1992; p 992.

(26) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *J. Comput. Chem.* **2006**, *27*, 781−790.

(27) Case, D. A.; Darden, T. A.; Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.;

Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A. W.; Kolossvari, I.; Wong, K. F.; Paesani, F.; Vanice *AMBER 12*; University of California: San Francisco, CA, 2012.

(28) Wang, J.; Wolf, R.; Caldwell, J.; Kollman, P.; Case, D. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.

(29) Jakalian, A.; Bush, B.; Jack, D.; Bayly, C. Fast, efficient generation of high-quality atomic Charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.

(30) Jakalian, A.; Jack, D.; Bayly, C. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.

(31) Wang, J.; Wang, W.; Kollman, P.; Case, D. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247−260.

(32) Frisch, M. J.; Trucks, G.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A. J.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*; Gaussian, Inc.: Wallingford, CT, 2009.

(33) Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T.; Jurečka, P. Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **2011**, *7*, 2886−2902.

(34) Pyevolve Home Page. http://pyevolve.sourceforge.net (accessed May 2014).

(35) Burger, S. K.; Ayers, P. W.; Schofield, J. Efficient parameterization of torsional terms for force fields. *J. Comput. Chem.* **2014**, *35*, 1438−1445.