AIP | The Journal of Chemical Physics

Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model

Jay L. Banks, George A. Kaminski, Ruhong Zhou, Daniel T. Mainz, B. J. Berne et al.

Citation: J. Chem. Phys. **110**, 741 (1999); doi: 10.1063/1.478043
View online: http://dx.doi.org/10.1063/1.478043
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v110/i2
Published by the American Institute of Physics.

Parametrizing a polarizable force field from *ab initio* data. I. The fluctuating point charge model

Jay L. Banks and George A. Kaminski
Department of Chemistry and Center for Biomolecular Simulation, Columbia University, New York, New York 10027

Ruhong Zhou
Schrödinger, Inc., 411 Hackensack Avenue, 10th Floor, Hackensack, New Jersey 07601

Daniel T. Mainz,[a] B. J. Berne, and Richard A. Friesner[b]
Department of Chemistry and Center for Biomolecular Simulation, Columbia University, New York, New York 10027

(Received 4 June 1998; accepted 8 October 1998)

We have developed a polarizable force field for peptides, using all-atom OPLS (OPLS-AA) nonelectrostatic terms and electrostatics based on a fluctuating charge model and fit to *ab initio* calculations of polarization responses. We discuss the fitting procedure, and specific techniques we have developed that are necessary in order to obtain an accurate, stable model. Our model is comparable to the best existing molecular mechanics force fields in reproducing quantum-chemical peptide energetics. It also accurately reproduces many-body effects in many cases. We believe that straightforward extensions of our linear-response electrostatic model will significantly improve the accuracy for those cases that the present model does not adequately address. © *1999 American Institute of Physics.* [S0021-9606(99)52402-1]

I. INTRODUCTION

The development of accurate and reliable force fields is a central objective of molecular modeling. For relatively simple systems consisting of small molecules of uniform composition, such as liquid water, it is possible to fit an empirical pair potential to experimental data (typically thermodynamic quantities). Such potentials can give a reasonable description of microscopic properties such as the liquid state radial distribution function. However, for more complicated molecules in heterogeneous environments, it is not feasible to construct an accurate set of potential functions from experimental information alone. Hence, *ab initio* quantum-chemical calculations have become an increasingly important source of fitting data. Most of the current generation of protein molecular modeling potential functions use such data to a greater or lesser extent.[1–15]

Even with the use of quantum-chemical methods, however, there are fundamental limitations on the accuracy attainable in describing a complex chemical system with a pairwise additive functional form. For liquid water, it is reasonable to represent many-body energetic effects in the system via an averaged enhancement of the two-body interaction energy, because each molecule can be thought of as existing in a substantially equivalent environment on a reasonably short time scale. In a protein, however, this is not the case; different amino acids are surrounded by very different molecular environments, and averaged interaction parameters cannot hope to represent all of these reliably. It will be necessary to include many-body effects explicitly in order to obtain truly quantitative accuracy in the next generation of force fields.

There is another reason why the use of a pair potential creates major problems in the development of molecular modeling force fields. If one models polarization effects via empirical adjustment of terms in a pair potential, one cannot then fit these terms to *ab initio* calculations on gas-phase molecular pairs. This makes it difficult to reliably parametrize interactions like hydrogen bonding. Most force fields represent hydrogen bonding primarily by electrostatic interactions between the donor and acceptor groups. Our group has recently shown, however, that electrostatics often provides a quantitatively inaccurate evaluation of hydrogen bonding; in fact, hydrogen bonding correlates more strongly with the acidity or alkalinity of the participating groups than with their partial charges.[16] A quantitatively accurate description of hydrogen bonding requires development of a force field explicitly representing many-body effects. Such a force field can be required to reproduce both gas-phase and (via the many-body terms) condensed-phase molecular properties; thus it can be fit to quantum-chemical pair data without compromising its applicability to condensed-phase calculations.

In order to construct a force field explicitly modeling many-body effects, we must first determine what functional form is required to properly represent many-body energetics as determined from accurate quantum chemistry. Our group, and others, have concluded that the great majority of many-body interactions can be well represented via a simple classical electrostatic model that includes local polarizability.[17,18] Furthermore, linear response is a very good approximation

[a]Present address: Materials and Process Simulation Center, Beckman Institute (139-74), California Institute of Technology, Pasadena, CA 91125.
[b]Electronic mail: rich@chem.columbia.edu

when electric fields on the order of typical strong intermolecular interactions are applied to a molecule. Thus, linear response models, such as those involving fluctuating charges[19,20] or polarizable dipoles,[21–23] are in principle capable of providing a quantitatively accurate description of many-body energetics.

If this analysis is correct, the major obstacle to widespread use of polarizable force fields for complex chemical systems is that they involve many more parameters than corresponding potentials that do not include polarizability. It is nontrivial to determine these added parameters from experimental data. Furthermore, using polarizable force fields increases the computational cost of carrying out simulations, although the increase can be relatively modest for certain functional forms. If we are to go to the considerable additional trouble and expense of incorporating polarizability, we require methods for assessing the errors in the model. For broad applicability, we also seek a systematic approach to parameter development. In contrast, the great majority of polarizable models published in the past decade have relied upon *ad hoc* parametrization methods that may yield reasonable results for a small number of examples, but would likely fail or become intractable in attempting to deal with a large set of molecules.

We present here a systematic, largely automated approach to determining the parameters for a polarizable force field from *ab initio* quantum-chemical calculations, and a rigorous method for monitoring the accuracy with which the force field reproduces many-body energetics. We have chosen in our initial work to describe the polarizable electrostatics using a fluctuating charge (FQ) model such as that described by Berne and co-workers,[20] based on electronegativity equalization principles.[24–29,19] Such a model has the advantage of having a relatively low computational cost when applied to large-scale molecular simulations. FQ and other point-charge models may not be adequate, however, for several important classes of functional groups such as aromatic rings (where a point charge model certainly cannot describe the out-of-plane polarization response) or bifurcated hydrogen bonds to carbonyl oxygens, an important motif in some types of drug–protein interactions. Thus we believe that we will need to augment the current FQ model in order to achieve a level of quantitative accuracy suitable for the next generation of force fields; the purpose of our exploration of this model is principally to demonstrate the feasibility of automatic fitting of parameters to *ab initio* data. It is formally and computationally straightforward to incorporate polarizable dipoles into the model, and we expect to report results in the near future for such an expanded model.[30]

In Sec. II, we present a theoretical formulation of the model, and describe the computer implementation required to fit parameters efficiently for large systems. In Sec. III, we construct a polarizable force field for the alanine dipeptide based on the OPLS-AA force field,[1] by replacing the fixed charge electrostatics with the FQ model and making a slight modification to the Lennard-Jones parameters. We test the transferability of this force field by applying it to the serine dipeptide, and by calculating energies for ten conformations of the alanine tetrapeptide, structures that our group has previously studied with quantum chemistry.[31] We compare the performance of this ''OPLS-FQ'' force field with conventional fixed charge force fields and accurate quantum-chemical data. Finally, in Sec. IV, we discuss future directions of research.

II. THEORY

A. Fluctuating charge model

The potential energy in an FQ model is given by[19]

$$E=\frac{1}{2}\sum_{ij} J_{ij}q_iq_j+\sum_i q_i(\chi_i+\phi_i), \tag{1}$$

where J_{ij} is a ''Coulomb'' matrix element, q_i is a partial charge on site i, ϕ_i is the value of the external electrostatic potential at site i, and χ_i is the ''electronegativity'' of atom i. Others have justified this equation in terms of fundamental quantum-chemical arguments, for example via density functional theory.[26–29] Our use of the FQ model focuses more on the mathematical form of Eq. (1) and its solutions, and less on the precise physical interpretation of the quantities J_{ij} and χ_i.

Application of the variational principle to Eq. (1), by minimizing the energy with respect to the charges q_i subject to the constraint of constant total charge, yields the set of linear equations

$$\mathbf{Jq}=-(\boldsymbol{\phi}+\boldsymbol{\chi}). \tag{2}$$

Our interpretation of the components of Eq. (2) is as follows. At sufficiently large separation of sites i and j, the matrix element J_{ij} must revert to the bare Coulomb interaction $1/r_{ij}$, as other effects fall off rapidly with distance. Empirically, the use of the bare Coulomb term appears to be accurate unless the atoms in question are connected by one or two bonds (i.e., ''1–2'' or ''1–3'' interactions). The self-interaction terms J_{ii} represent the penalty for adding or removing charge from site i, which presumably depends on very complex quantum-chemical effects as well as Coulombic energies.

These diagonal and near-neighbor J_{ij} interactions can qualitatively be thought of as ''screened Coulomb'' terms, but it is not clear that we could easily derive quantitatively accurate formulas for them from *ab initio* quantum chemistry along these lines. Rather, it seems more useful to simply consider **J** as a linear response matrix that converts an input external field and/or electronegativity vector into a set of partial atomic charges. If we were to replace the partial charges q_i on atomic sites by a dense grid of charge densities $\rho(\mathbf{r}_i)$ where the $\mathbf{r}_i$ are the grid point locations, we could rigorously identify the inverse of the **J** matrix as the Green's function $G(\mathbf{r},\mathbf{r}')$ and derive the equivalent of Eq. (2) from standard Rayleigh–Schrödinger perturbation theory. In practice, we must use a more heuristic definition of the charge distribution [in the present work, we use electrostatic-potential-fitted (ESP) partial charges[32]], which renders any rigorous derivation problematic. This means that we may as well consider the ''near'' J_{ij}s to be adjustable parameters, which we hope will be specific to local functional groups and

hence transferable between molecules. The problem of developing an FQ model from *ab initio* data is thus reduced to determining the J_{ij}s not specified by Coulomb's law. We subsequently determine the χ_is using charges computed in the absence of any external field.

B. General linear response analysis of polarizable models

The above analysis is not restricted to an FQ model with point charges on atom centers only. A simple generalization of the model is to allow sites to be defined anywhere in physical space—for example, on the ''*M* site'' of the TIP4P water model,[33,20] on bonds between atoms, at lone pair positions, or above and below the plane of an aromatic ring. All of these models are linear response models and the only difference is the number of charge sites *N* and the location of these sites. Furthermore, a polarizable dipole model is also a linear response model, although its equations involve electric fields (gradients of potentials) as well as electrostatic potentials. The formalism and computational procedures we present here can readily be applied, with minimal modification, to incorporate off-atom charges, polarizable dipoles, or both.

The theoretical accuracy achievable with the model depends on the choice of the charge representation—the number, type, and location of polarizable sites. Assuming that a linear response model is an accurate picture—and our numerical tests indicate that it is, for field strengths relevant to intermolecular interactions—the only important question is whether a particular linear response model is capable of reproducing the properties of the actual quantum-mechanical system. On the other hand, we have found that attempting to reproduce the quantum-chemical response spectrum to arbitrary precision can lead to an unstable model. We investigate these issues in more detail in Secs. II C 1 and III. Here, we discuss the technology that we use to collect the quantum-chemical data and derive a linear response model from it.

Suppose we want to build a linear response model of a single conformation of a particular molecule. Our basic procedure is to perturb the molecule with a series of applied electric fields, which we design to span the space of fields relevant to intermolecular interactions. We have developed an automated code that places point-charge or dipole probes at various locations surrounding the molecule. The first sites we fill are potential hydrogen bonding positions, including points above and below the centers of aromatic rings (taking the π electron clouds to be potential hydrogen-bond acceptors). In general, there are not enough such sites to generate a sufficiently large set of fields to fit the parameters of our model, and we add probes either at random locations (beyond some minimum distance from the molecule) or uniformly spaced on the molecular van der Waals surface. Because we exclude an infinite set of fields from our fit, and indeed because we restrict our model itself to a finite number of sites comparable to the number of atoms, we may end up with a model that cannot fit the response to very strong fields, or those containing large, rapid oscillations on an atomic scale. But such a model may still be perfectly sufficient to describe the interactions of actual molecules. Moreover, our tests on several small molecules indicate that the polarization energies that the model predicts are insensitive to the precise choice of nonhydrogen-bonding probe sites for the fit, suggesting that additional probes would not add significant information.

Initially, we determine ''zero-field'' charges by performing an ESP fit[32] on the gas-phase wave function of the molecule. Then we perform such a fit for the wave function computed from each applied field calculation. The differences in the partial charges on sites *i*, due to applied field *k* as compared to the gas phase, are denoted δq_{ik}, or in vector notation $\delta \mathbf{q}_k$. The linearity of the model then allows us to subtract Eq. (2) for the gas phase from the same equation for each applied field, implying that the FQ Coulomb matrix $\mathbf{J}$ should satisfy

$$\mathbf{J}\delta\mathbf{q}_k = -\boldsymbol{\phi}_k, \tag{3}$$

where $\boldsymbol{\phi}_k$ is a vector specifying the value of the *k*th applied potential at the various sites of the FQ model. Note that the site electronegativities $\boldsymbol{\chi}$ do not appear in this equation. They affect the gas-phase charges on the molecule, but the linear response matrix depends solely on the charge shifts induced by applied fields.

In general, if there are *N* charge-carrying sites in the FQ model, we require data from *M* applied fields, where $M > N$. Given sufficient data, we use a least-squares formalism, which adds stability and reliability to the fitting procedure. In a straightforward implementation of this formalism, the residual error $\boldsymbol{\epsilon}$ in the least-squares sense is defined as:

$$\boldsymbol{\epsilon} = \sum_{k=1}^{M} w_k \sum_{i=1}^{N} \left(\sum_{j=1}^{N} J_{ij}\delta q_{jk} + \phi_{ik} \right)^2, \tag{4}$$

where w_k is a weight assigned to dataset *k*. (In fitting experimental data, the appropriate weight for a data point is the reciprocal of its variance; but the weights we use have no such interpretation in terms of uncertainties in our *ab initio* ''data.'' Unless otherwise specified, we use equal weights for all data.)

If we ignore for the moment the issue of the appropriate functional form for the J_{ij}, we can find the set of values J_{ij} that minimizes $\boldsymbol{\epsilon}$ by solving the normal equations:

$$\sum_{j,k} J_{ij}\delta q_{jk} w_k \delta q_{kl}^T = -\sum_k \phi_{ik} w_k \delta q_{kl}^T. \tag{5}$$

We can solve this set of equations numerically, using standard techniques for linear systems on the square matrix $\delta\mathbf{q}\mathbf{w}\delta\mathbf{q}^T$. Since we do not know in advance that all the $\boldsymbol{\phi}_k$s are linearly independent (and indeed our calculations indicate that they often are not), we prefer to use the singular value decomposition formalism (SVD) to solve Eq. (5). Equivalently, we can apply SVD directly to the overdetermined system of Eq. (3), ''inverting'' the $N \times M$ matrix $\delta\mathbf{q}$ to obtain the values of J_{ij} that come closest, in the least-squares sense, to satisfying this equation. We set the SVD cutoff sufficiently low that the eigenvalues we omit (typically ranging from 10^{-5} to 10^{-30} times the largest eigenvalue) clearly represent degeneracies in the fields rather than

actual physics. Although SVD is generally slower than other linear system methods, we have found the computation times acceptable for the systems we have studied.

In the calculations described here, we impose the constraint $J_{ij}=J_{ji}$ on the matrix elements, in keeping with the physical interpretation of the J_{ij}s as Coulomb matrix elements. Allowing an asymmetric **J** matrix would in general produce a fit with a smaller residual, but we have not found this generalization necessary.

It is important to note that solving Eq. (5) for J_{ij} actually solves the wrong optimization problem. We want to find the FQ model that produces polarization charges $\delta\mathbf{q}$ that agree with those from quantum mechanics, given input potentials $\boldsymbol{\phi}$. As written, the least-squares equations instead minimize the error in $\boldsymbol{\phi}$ given an input $\delta\mathbf{q}$. If the solution for **J** results in a residual exactly equal to zero, then the two conditions are equivalent. In general, however, there will be significant differences in the nature of the errors that appear.

One way to understand this problem is through an eigenvalue analysis of the linear response matrix **J**. Solving Eq. (5) weights the contributions of the various eigenvectors of **J** in direct proportion to their eigenvalues. In the polarization response to an external field, however, the contribution of a given eigenvector of **J** to $\delta\mathbf{q}$ is proportional to the reciprocal of the corresponding eigenvalue, so it is more important to represent the small-eigenvalue modes accurately.

To minimize the errors in $\delta\mathbf{q}$, it would thus be better to fit the inverse of the **J** matrix, $\mathbf{J}^{-1}$. In this formulation, the residual is

$$\epsilon=\sum_{k=1}^{M} w_k \sum_{i=1}^{N}\left(\sum_{j=1}^{N} J_{ij}^{-1}\phi_{jk}+\delta q_{ik}\right)^2. \tag{6}$$

We show in Secs. II C 3 and III that minimizing this residual leads to a highly accurate reproduction of the quantum-mechanical response modes.

Solving for $\mathbf{J}^{-1}$ is likely to give the best and most stable fit to a single conformer for any given version of the FQ model, and as such can be a useful validation tool for methods such as the ''direct'' solution for **J**. Unfortunately, a given element of $\mathbf{J}^{-1}$ has contributions from all atoms in the molecule, and its analytical form at long distances is unclear, unlike the simple Coulomb form of the elements of **J**. Therefore, this approach is not useful for developing a model that is transferable among different molecules, or even different conformations of one molecule.

To fit a transferable functional form, then, we must use Eq. (5), and face the numerical instabilities resulting from a fitting procedure that suppresses precisely the modes corresponding to large polarization responses. A naïve implementation of this approach is in fact numerically unstable for all but the smallest target molecules. We have made significant technical improvements, however, which render the fitting procedure stable. We have automated most parts of these techniques, which is essential to our objective of developing the model with a minimum of human intervention. We describe these improvements in Sec. II C.

C. Construction of a numerically stable fitting procedure for the linear response matrix

We describe below three techniques for enhancing the stability of the linear response model when fitting the **J** matrix directly [Eq. (3)]. The combined use of all three techniques appears to be necessary to generate stable and reliable results, as we discuss in Sec. III. Careful examination of factors affecting stability is in its infancy, however, and we expect that new approaches to filtering noise will provide further improvements in the results.

1. Removal of small-eigenvalue modes from the electrostatic potential fit results

It is well known that when a molecule has ''buried'' atoms (that is, atoms with no exposed surface area), the ESP fitting procedure becomes numerically unstable.[34] In such cases, the matrix that appears in the normal equations of the ESP fit has small eigenvalues. This matrix is derived from the Coulomb operator $R_{ij}=1/r_{ij}$, where r_{ij} is the distance between a charge site *i* and a grid point *j* where the electrostatic potential is to be fit. Its small eigenvalues correspond to charge distributions that have small effects on the electrostatic potential outside the surface on which the grid points *j* lie, typically the molecular van der Waals surface. Such charge distributions have very small dipole (and other low multipole) moments, and are typically delocalized over the entire molecule.

Because the external field associated with an unstable ESP eigenmode is small, obtaining an accurate measure of its amplitude in a least-squares procedure is problematic. Small changes in the molecular geometry can have a significant influence on the optimized eigenvector coefficients, particularly if the modes are relatively delocalized. When the ''signal'' is small to begin with, as it inevitably will be for an unstable mode, the ''noise'' associated with these geometry alterations leads to an erratic representation of the geometric dependence of the mode. It is very difficult to reproduce the behavior of such modes with a simple functional form such as we use in our FQ model, particularly since there are problems with the inverted weighting of the small-eigenvalue modes in any case. In practice, we find that inclusion of unstable ESP-fit modes leads to chronic and apparently irremediable instabilities in the resulting FQ model.

We take a straightforward approach to the unstable mode problem, and project these modes out of the fit using singular value decomposition. Note that here, in contrast to the use of SVD to remove degeneracies in fitting the **J** matrix, we may be omitting physically plausible charge distributions. This raises the practical question of where to set the eigenvalue cutoff. If the cutoff is too small, some small-mode instabilities remain; if it is too large, the electrostatic representation of the molecule may be inaccurate. We desire a systematic method for finding a compromise that ensures stability while only marginally diminishing accuracy.

Fortunately, the same characteristic that renders a mode unstable—it produces a very small electrostatic potential outside of the molecular van der Waals surface—also implies that its contribution to the polarization energy is minimal. Indeed, the interaction of such a charge distribution with

TABLE I. Eigenvalues (atomic units) of the ESP fit for the alanine dipeptide $C7_{eq}$ conformation, and interaction energies (kcal/mol) of the corresponding charge distributions with each of three external probes placed in hydrogen-bonding positions. The first and last eigenmodes arise from the charge-conservation constraint.

Eigenvalue	*E*(probe 1)	*E*(probe 2)	*E*(probe 3)
−0.0150	−0.0544	−0.0569	0.0045
0.0034	−0.0635	0.1548	−0.0485
0.0045	0.2370	0.1739	−0.0017
0.0051	−0.3277	0.0395	−0.0747
0.0067	−0.0148	−0.1989	0.1844
0.0115	0.9978	0.4152	0.6276
0.0259	−1.1860	1.2246	−1.4713
0.0969	−3.1833	1.9274	0.3588
0.1698	2.3057	2.3277	−5.6939
0.5429	−3.5148	−6.9225	−7.2461
0.6478	−0.7221	6.0713	−4.5768
0.7707	−4.9499	−2.3820	−4.1117
0.8823	8.3657	−5.4829	−3.6771
1.3756	2.9930	−1.1686	−0.8254
2.3583	2.4858	1.7092	−1.8747
2.5891	6.7303	2.2138	1.0462
3.1523	−3.5317	2.0916	−4.7158
5.0301	6.9582	0.8175	−15.5601
5.4441	7.2251	15.7573	3.3441
15.0500	−16.7209	8.4220	−7.6560
54.3794	6.9615	−6.9842	6.1629
75.2700	−9.0679	11.6403	12.1804
1598.6489	−30.5530	−30.3299	33.5109

other molecules is fundamentally limited, as a test charge at a realistic location for any interacting molecule will be outside of the van der Waals surface. This suggests a practical approach to assessing the energetic importance of any given mode. We place probe charges at hydrogen-bonding distances (the closest intermolecular interaction that is important in a real system) at various points surrounding the molecule and calculate the electrostatic interaction energy of the charges with the various ESP eigenmodes of the molecule. Table I presents typical results from one conformation of the alanine dipeptide. We see here that the small-eigenvalue modes indeed interact only weakly with the probes. We can therefore use this criterion to choose an eigenvalue cutoff for the ESP fit. In practice, we are still experimenting with the choice of this cutoff, which is therefore not yet an automated feature of our procedure.

It is also helpful that the many-body polarization energy is typically one order of magnitude smaller than the two-body Coulomb interaction. This suggests that we can cut off the ESP fit at a larger eigenvalue for fitting the linear response matrix, and use a smaller cutoff to generate the electronegativities. This approach increases the stability of the linear response matrix while allowing a more accurate description of the permanent (zero-field) charges. If we instead calculate the electronegativities using the larger cutoff required for stability, we produce inaccuracies in many-body energies on the order of 0.2–0.5 kcal/mol. With the two-tiered scheme proposed above, errors attributable to the ESP fitting protocol are only about 0.1 kcal/mol. We consider this to be acceptable for a prototype force field. For the alanine dipeptide, all six conformers we use in the FQ fit have similar patterns of ESP eigenvalues, and similar interaction energies of the corresponding charge distributions with external probes. In Sec. III B, we present results for an FQ model for this molecule that involves cutting the four smallest positive eigenvalues (as well as the negative eigenvalue associated with the charge-conservation constraint) out of the data for the electronegativity calculation (zero-field charges), and cutting the eight lowest positive modes out of the polarization charges to which we fit the linear response matrix.

To implement the ESP eigenvalue cutoffs in fitting the linear response model, we simply use the polarization charges computed with these cutoffs as the fitting data $\delta\mathbf{q}$ in Eq. (3). In other words, we ignore polarization responses proportional to the omitted ESP modes. As we have noted above, attempting to fit the linear response matrix **J** to responses including these modes leads to serious instabilities, so this effective damping of the unstable modes is crucial to achieving a robust and reliable linear response model that can properly represent the system across a wide range of geometries.

2. Use of bond space rather than site space in the fitting protocol

Equations (1)–(3) represent the charge distribution of the molecule in terms of atom-centered point charges, whose sum is constrained to equal the net charge on the molecule. This constraint does not directly appear in the least-squares formalism for solving Eq. (3). Instead, the ESP fitting procedure incorporates a Lagrange multiplier formalism, which guarantees that the *ab initio* polarization charges resulting from any applied field sum to zero, preserving the net charge. The $\delta\mathbf{q}$ vectors that we use to fit the **J** matrix thus span only an $N-1$-dimensional subspace of the N-dimensional space of all possible charge distributions for an N-atom molecule, and the resulting **J**, if it were an exact solution of Eq. (3), would be of rank $N-1$ and have a zero eigenvalue.

We have empirically determined that we can avoid some of the instability in the **J** matrix by explicitly changing variables to an $N-1$-dimensional representation of the charge distribution, which conserves charge automatically. We define bond-charge increments (BCIs) h_{ij}, on bonds connecting atom sites i and j, which are related to the site charges q_i by

$$q_i = \sum_j h_{ij}, \tag{7}$$

where the sum is over all atoms j bonded to atom i. If $h_{ji} = -h_{ij}$, the charges q_i automatically sum to zero; for an ionic molecule, we can add fixed charges $q_i^{(0)}$ to sum to the desired net charge, without affecting the hs or the linear response matrix. An acyclic N-atom molecule has exactly $N-1$ bonds, and Eq. (7) uniquely determines the h_{ij}s. A cyclic molecule has additional bonds to close rings, leading to ambiguities in the BCIs. We resolve this by fixing $h_{ij} = 0$ for one bond in each ring. (The arbitrary choice of which bond to ''cut'' in this manner can significantly affect the resulting values of the response matrix elements, and de-

tailed work with ring molecules will require an appropriate method for ''averaging out'' such effects. In preliminary investigations, we have achieved reasonable results by simply replicating datasets in the least-squares fit, choosing $h=0$ for a different bond in each replica.) Once we have chosen an unambiguous set of BCIs, it is straightforward to express them in terms of the charges:

$$h_k=\sum_i a_{ki}q_i, \tag{8}$$

where the coefficients a_{ki} are all 0, 1, or -1, and depend solely on the topology of the molecule, not on the values of the qs in a given conformation or environment. Thus we can compute the coefficients once for a given molecule, and store them for repeated use as needed.

The formal transformation of the fitting equations from ''site space'' to ''bond space'' is straightforward. We first rewrite Eq. (7) as a sum over bonds l rather than neighbor atoms j:

$$q_i=\sum_l h_{il}, \tag{9}$$

where we can take the sum to run over all bonds in the molecule, if we define $h_{il}=0$ for bonds l that do not connect to atom i. Clearly if bond l connects atoms i and j, then $h_{il}=-h_{jl}$. If we adopt the convention to list the atoms in bond l as (p,q) with p always less than q, then we can write

$$h_{il}=(\delta_{ip}-\delta_{iq})h_l, \tag{10}$$

where δ_{ip} and δ_{iq} are Kronecker deltas, and h_l is the bond-charge increment that bond l contributes to its ''first'' atom, p.

We insert the defining equations (9) and (10) into Eq. (1) to obtain the energy expression in bond space:

$$E=\frac{1}{2}\sum_{lm}(J_{il,im}-J_{il,jm}-J_{jl,im}+J_{jl,jm})h_lh_m + \sum_l[(\chi_{il}+\phi_{il})-(\chi_{jl}+\phi_{jl})]h_l, \tag{11}$$

where for each term in the sums, bond l connects atoms (il,jl) and bond m connects atoms (im,jm), with $il<jl$ and $im<jm$. Applying the variational principle to this equation yields

$$\sum_m M_{lm}h_m=-(\chi'_l+\phi'_l), \tag{12}$$

where we have defined the bond-space equivalents of $\mathbf{J}$, $\boldsymbol{\chi}$, and ϕ by

$$M_{lm}=J_{il,im}-J_{il,jm}-J_{jl,im}+J_{jl,jm}, \tag{13}$$

$$\chi'_l=\chi_{il}-\chi_{jl}, \tag{14}$$

$$\phi'_l=\phi_{il}-\phi_{jl}. \tag{15}$$

As we did above to obtain Eq. (3) from Eq. (2), we subtract Eq. (12) for zero field from the same equation for a given electrostatic potential distribution $\boldsymbol{\phi}'$, to obtain:

$$\mathbf{M}\delta\mathbf{h}=-\boldsymbol{\phi}', \tag{16}$$

where $\delta\mathbf{h}$ is related to $\delta\mathbf{q}$ by exactly the same linear transformation as between $\mathbf{h}$ and $\mathbf{q}$.

The definition, in Eq. (13), of the bond-space matrix element M_{lm} expresses it in terms of the site-space matrix elements J_{ij} that connect one atom in bond l to one in bond m. As we have explained in Sec. II A, we take the functional form of a given J_{ij} to depend on the ''topological separation'' of atoms i and j—the number of bonds connecting them. In particular, we take the J_{ij}s for ''1–4'' (three intervening bonds) and more distant interactions to have the pure Coulomb form, and those for closer interactions to be independent of the molecular geometry (see Sec. II D). A given M_{lm}, however, combines J_{ij}s of several different topological separations, and a given J_{ij} contributes to several different M_{lm}s. For M_{lm}s that contain no Coulomb contributions, we can fit M_{lm} directly rather than the individual J_{ij}s. For those that involve both Coulomb and non-Coulomb J_{ij}s, we must fit the non-Coulomb contributions separately, and ensure that the same non-Coulomb contribution to two different M_{lm}s is not fit by two separate parameters. We enumerate here a minimal set of independent parameters that determine all the non-Coulomb parts of various topological classes of M_{lm}s. The notation $J^{(1-n)}$ here indicates a J_{ij} for separation $1-n$, where $n=1$ (diagonal), 2, 3, etc.

Diagonal elements: For $l=m$, M_{lm} has contributions from two $J^{(1-1)}$s and one $J^{(1-2)}$, but we fit the combination as a single parameter.

Touching bonds: If l and m have one atom in common, there are again no Coulomb contributions, so we again fit a single parameter.

Torsional terms: If l and m are separated by one additional bond (that is, they form the outer bonds of a torsional angle), then M_{lm} contains one $J^{(1-2)}$, two $J^{(1-3)}$s, and one Coulomb term. We fit separate parameters for the $J^{(1-3)}$s, and a ''torsional'' parameter representing the $J^{(1-2)}$. Note that we must constrain each $J^{(1-3)}$ contribution to be equal to the contribution of the same atoms to other M_{lm}s. But we need not relate the torsional parameter to the $J^{(1-2)}$ contributions in the previous classes of M_{lm}s, since these are combined with other contributions in a single overall parameter in those classes. (If we call the overall parameter a and the $J^{(1-2)}$ contribution b, and let $c=a-b$, then we are free to treat either a and b or c and b as the independent parameters.)

Two intervening bonds: Such M_{lm}s are composed of one $J^{(1-3)}$ and three Coulomb terms. Since $J^{(1-3)}$s are treated as independent parameters in the previous class, we must again constrain the $J^{(1-3)}$ contribution to one of these M_{lm}s to be equal to the contribution of the same atoms to other M_{lm}s.

More than two intervening bonds: For such distant separations, M_{lm} contains only Coulomb contributions.

3. Nonlinear refinement of the linear response matrix elements

The procedures we have described have the advantage of employing a linear least-squares methodology to determine the linear response matrix elements. Solution of a linear system of equations is rapid even if there are thousands of vari-

ables, and issues such as multiple minima or slow convergence, which are important effects when solving nonlinear equations, do not arise for a linear system.

As we have noted in Sec. II B, however, our formalism really solves the *wrong* linear least-squares problem. We minimize the error in $\boldsymbol{\phi}$, whereas for stability and accuracy we really want to minimize the error in $\delta\mathbf{q}$, which is a *nonlinear* function of the response matrix elements J_{ij}. (It is of course a linear function of the inverse matrix elements J_{ij}^{-1}, but as we have pointed out, these matrix elements are not suitable for constructing a transferable potential function.) This suggests that we can improve on the results of the least-squares fit to Eq. (3) [or Eq. (16) in bond space] by using them as an initial guess in a standard (e.g., conjugate gradient) nonlinear optimization of the true residual, Eq. (6). Using a very good initial guess defuses the multiple minimum problem—presumably, we are starting quite close to the optimal solution—and significantly reduces the computational effort necessary to carry out the optimization, as compared to a random starting point.

The major computational expense in the nonlinear optimization procedure is the determination of the derivatives of the objective function with respect to the fitting parameters. Finite difference computations of these derivatives would involve evaluating the objective function, and thus inverting the linear response matrix, at least one extra time *per parameter* at each iteration of the minimization algorithm. Instead, we derive analytic expressions for the derivatives that involve far fewer inversions of the matrix, and allow an evaluation procedure that scales favorably with system size.

We want to minimize the objective function ϵ as given by Eq. (6). We rewrite this equation as

$$\epsilon = \sum_k w_k \sum_i (y_{ik}^{(\mathrm{fit})} - y_{ik}^{(0)})^2, \tag{17}$$

where $y_{ik}^{(0)}$ is the ''correct'' (*ab initio*) value of δq (or δh) for site (or bond) i in dataset k, and

$$y_{jk}^{(\mathrm{fit})} = -\sum_i J_{ji}^{-1} \phi_{ik} \tag{18}$$

is the fit value of the same quantity at a given stage of the fitting procedure. We assume that the elements of the **J** (or **M**) matrix, and thus those of $y^{(\mathrm{fit})}$, depend on some parameters A_p, which we vary in the nonlinear fit. As we explain in Sec. II D, we typically take the parameters to be the non-Coulomb J_{ij}s themselves, with no spatial dependence. This makes the dependence of the matrix elements on the parameters completely trivial in site space, and almost as trivial in bond space. The dependence of the objective function on the parameters is given by

$$\frac{\partial \epsilon}{\partial A_p} = 2 \sum_{k,j} w_k (y_{jk}^{(\mathrm{fit})} - y_{jk}^{(0)}) \frac{\partial y_{jk}^{(\mathrm{fit})}}{\partial A_p}. \tag{19}$$

We note that as the A_ps change, Eq. (18), defining the relationship between the J_{ij}s and the $y_i^{(\mathrm{fit})}$s, remains valid. In other words, the quantities

$$\sum_j J_{ij}(A) y_{jk}^{(\mathrm{fit})}(A) = -\phi_{ik} \tag{20}$$

are independent of the A_ps. We use this equation to solve for the $\partial y_j / \partial A_p$s in terms of the $\partial J_{ij} / \partial A_p$s:

$$\sum_j \frac{\partial}{\partial A_p} (J_{ij} y_{jk}^{(\mathrm{fit})}) = \sum_j \left[\left(\frac{\partial J_{ij}}{\partial A_p} \right) y_{jk}^{(\mathrm{fit})} + J_{ij} \left(\frac{\partial y_{jk}^{(\mathrm{fit})}}{\partial A_p} \right) \right] = 0, \tag{21}$$

$$\frac{\partial y_{jk}^{(\mathrm{fit})}}{\partial A_p} = -\sum_{i,r} J_{ji}^{-1} \left(\frac{\partial J_{ir}}{\partial A_p} \right) y_{rk}^{(\mathrm{fit})}. \tag{22}$$

Plugging Eq. (22) into Eq. (19), we get an analytic expression for the derivatives of the objective function ϵ with respect to the parameters A_p:

$$\frac{\partial \epsilon}{\partial A_p} = -2 \sum_{k,j} w_k (y_{jk}^{(\mathrm{fit})} - y_{jk}^{(0)}) \sum_{i,r} J_{ji}^{-1} \left(\frac{\partial J_{ir}}{\partial A_p} \right) y_{rk}^{(\mathrm{fit})}, \tag{23}$$

where the $(\partial J_{ir} / \partial A_p)$s are all constants, many of them zero. At each optimization step, we need to invert the **J** matrix *once* rather than once per parameter, and obtain the new y_js from Eq. (18). Furthermore, we can arrange the order of calculating the sums so as to minimize computation time.

D. Specification of a transferable FQ functional form

As stated above, we treat all interactions involving ''1–4'' and more distant connections as simple Coulomb electrostatics, with no free parameters. Thus the $J^{(1-1)}$, $J^{(1-2)}$, and $J^{(1-3)}$ parameters are sufficient to specify the linear response model. Our first ansatz is to define these parameters to be independent of interatomic distances, for the same reason that conventional molecular mechanics force fields set the equivalent Coulomb and van der Waals interactions equal to zero. As in previous work,[18] we assume that the valence part of the energy function, including such terms as stretches and bends, completely describes the energy changes as a function of displacements of these internal coordinates. In a polarizable model, however, setting the ''valence'' electrostatic terms to zero might make it more difficult to describe delocalized polarization modes involving coupled charge shifts. Instead, we achieve the same effect by eliminating the distance dependence of the interaction parameters, thus avoiding large changes in the electrostatic energy upon displacement of internal coordinates (which in quantum-chemical calculations are counterbalanced by other equally large terms). With this approach, we have found that it is possible to use existing molecular mechanics parameter sets, with minimal modification, to describe stretching and bending interactions.

In this formulation, it is then natural to associate parameters with atom types and attempt to build up a transferable database of parameters. Because we fit directly to *ab initio* quantum chemistry, with no empirical modification for ''condensed phase'' effects, in principle we can use a very large set of highly sophisticated atom types, limited only by our ability to generate the quantum-chemical data, which due to recent computational advances is actually quite rapid. In the present paper, we make no attempt to describe a com-

plete atom-typing approach; rather, we hand code specific cases. However, the generalization to an arbitrary molecule is straightforward, if tedious, and is currently in progress.

III. RESULTS

A. Overview

We present here five results we have achieved using the methods of Sec. II:

(1) We have constructed a polarizable electrostatic model for the alanine dipeptide, with parameters fitted to atom types over several conformations of the molecule.
(2) This electrostatic model accurately reproduces quantum-chemical three-body energies for test cases consisting of the dipeptide and two external probes.
(3) We have assembled a complete force field for the dipeptide by combining our electrostatic model with the OPLS-AA stretching, bending, torsional, and van der Waals terms, with minimal modifications in parameters.
(4) This force field performs well in ranking the energies of alanine tetrapeptide conformations, using as a benchmark high level quantum-chemical energies that our group has determined in previous work.[31]
(5) Using backbone parameters from the alanine dipeptide model, and in one test adding side-chain parameters from fits to small molecules, we have constructed similar electrostatic models for the serine dipeptide. The success of these models indicates that we can hope to obtain transferable electrostatic parameters by fitting a polarizable model to a tractable subset of the universe of all possible molecules.

We have selected polyalanine as a test case because it is central to the construction of protein force fields. Most such force fields in common use derive peptide backbone parameters by fitting the alanine dipeptide surface. The coupled torsional interactions in the peptide backbone, and the highly polar amide groups, present a significant challenge in force field development. It is highly nontrivial to reproduce tetrapeptide energetics: as our group has previously reported,[31] many widely used protein force fields perform this task quite poorly. If we can successfully model these energies, and also many-body effects, using a truly systematic protocol, then there is substantial reason to believe that our methods will give satisfactory results for a wide variety of complex molecules.

B. Polarizable electrostatic model for the alanine dipeptide

In previous work,[31] our group has generated the six minima on the alanine dipeptide surface at the HF/6-31G** level via quantum-chemical minimization, followed by single-point LMP2/cc-pVTZ(-f) calculations to determine accurate energy differences. We list the (ϕ,ψ) torsional angles from the HF/6-31G** structures, and the LMP2/cc-pVTZ(-f) relative energies, in Table II.

For each of these conformers, we generate a set of quantum-chemical responses to applied fields. By including data from all six minima, we aim to develop a model that reproduces the molecular polarizability over a wide range of the dipeptide geometry. We use up to 30 polarization data sets per conformer in a combined least-squares fit. For the results presented below, we solve the bond-space equation (16) for the non-Coulombic parameters M_{lm} describing 1–1, 1–2, and 1–3 interactions. We assign unique parameters according to the atom types defined in Table III. We use atom types, rather than optimizing each matrix element individually, in the hope of generating a transferable model.

TABLE II. Alanine dipeptide HF/6-31G** torsional angles ϕ and ψ (deg), and LMP2/cc-pVTZ(-f) relative energies (kcal/mol), for six conformations at local minima of the energy surface. Reproduced from Ref. 31, Tables 1 and 2.

Conf.	ϕ	ψ	Energy
$C7_{eq}$	−85.8	−78.5	0.00
C5	−157.9	160.3	0.95
$C7_{ax}$	75.8	−56.5	2.67
β_2	−128.6	23.2	2.75
α_L	66.9	29.7	4.31
α'	−166.4	−40.1	5.51

Table IV gives the root-mean-square (rms) error, and maximum absolute error, in the individual polarization charges resulting from several least-squares fits to alanine dipeptide data. In all cases shown, we used bond-space charges and fit the matrix elements M_{lm} by atom types rather than individually. The cases differ in whether, and to what extent, we discarded small-eigenvalue modes in the ESP fit (see Sec. II C 1), and whether we used a nonlinear optimization step to improve the model (see Sec. II C 3). If all data sets have unit weight in the fit, the rms error is related to the objective function ϵ of Eq. (17) by

$$\mathrm{rms} = \sqrt{\epsilon / N_{\mathrm{pts}}}, \tag{24}$$

TABLE III. Atom type codes for the FQ model. (In all cases, code=100×atomic number+subcode.)

Code	Description	Where found
104	Amide hydrogen	Peptide backbone
108	Hydrogen on α-carbon	
608	Carbonyl carbon	
609	α-carbon	
702	Amide nitrogen	
802	Carbonyl oxygen	
102	Acetyl (methyl) hydrogen	Peptide end groups
107	*N*-methylamide (methyl) hydrogen	
606	*N*-methylamide (methyl) carbon	
607	Acetyl methyl carbon	
618	Acetyl carbonyl carbon	
110	(Methyl) hydrogen on β-carbon	Alanine side chain
610	(Methyl) β-carbon	
132	CH_2 β-hydrogen	Serine side chain
133	Hydroxyl hydrogen	
637	CH_2 β-carbon	
833	Hydroxyl oxygen	
101	Water hydrogen	Water
801	Water oxygen	

TABLE IV. rms and maximum absolute errors, in electronic charge units (ECU), in alanine dipeptide polarization charges, from bond-space FQ models with various ESP cutoffs and with or without nonlinear refinement of parameters. (Perm. cut=number of modes omitted in calculation of zero-field charges. Pol. cut=number of modes omitted in calculation of polarization charges.)

Perm. cut	Pol. cut	Refinement?	rms error	Max |error|
0	0	no	0.0062	0.0728
4	7	no	0.0046	0.0527
4	8	no	0.0028	0.0294
4	9	no	0.0036	0.0356
4	8	yes	0.0015	0.0097

where N_{pts} is the total number of charge sites in the fit: with six conformers, 30 polarization datasets per conformer, and 22 charge sites (atoms) in the molecule, $N_{\text{pts}}=6\times30\times22=3960$. Both the rms error and the maximum absolute error are useful predictors of the stability of the model. In particular, if the model severely overestimates polarization charges, these overestimates will increase during the course of a geometry optimization or dynamics simulation, as the larger fields from inaccurately large charges produce inaccurately large polarizations in nearby atoms. It is clear from Table IV that cutting ESP modes, and nonlinear refinement, both lead to significant improvements in the quality of the fit. We expect that accuracy on the order of 1×10^{-2} charge units, as in the last line of the table, will be sufficient for most purposes.

We present in Tables V–VIII the parameters of the FQ model resulting from the bond-space fit to six alanine dipeptide conformations with four ESP eigenmodes cut from the zero-field charges (SVD cutoff=0.01 charge units) and eight modes cut from the polarization charges (cutoff=0.2), including nonlinear refinement. These parameters are as defined in Sec. II C 2: in particular, the torsional parameter in Table VIII corresponds to the $J^{(1-2)}$ contribution to the interaction between a pair of bonds that, with the one intervening bond, form a torsional angle. In addition to testing the quality of the fit as shown in Table IV, we have used these parameters to calculate polarization charges for two conformations not included in the fit. The (ϕ,ψ) values for these conformations (in degrees) are $(50, -130)$, the location of a seventh minimum of the LMP2/cc-pVTZ(-f)//HF/6-31G** energy surface that Beachy of our group has generated;[35] and $(-60, -60)$, in the α-helical region. The maximum absolute deviation of the polarization charges from the *ab initio* values for these two conformations is less than 0.02 charge

TABLE V. Diagonal bond-space FQ parameters for the alanine dipeptide. The bond l connects atoms of types i and j. Here and in the following three tables, the elements of the **M** matrix are in units of kcal/(mol×ECU^2), and the left-hand columns indicate the atom types to which each parameter on the right applies.

i	j	M_{ll}
607	618	568.9222
607	102	608.6219
618	702	467.8894
618	802	727.3375
702	609	688.3098
702	104	953.3389
609	610	768.1983
609	608	750.4136
609	108	869.9510
610	110	588.4623
608	802	658.3106
608	702	502.9180
702	606	493.7056
606	107	555.3920

TABLE VI. FQ parameters for pairs of touching bonds in the alanine dipeptide. The bonds l and m have the atom of type j in common.

i	j	k	M_{lm}
618	607	102	124.5923
607	618	702	−49.1500
607	618	802	−242.4931
102	607	102	288.0543
702	618	802	195.8489
618	702	609	−71.5046
618	702	104	−113.0463
609	702	104	235.8284
702	609	610	−269.8045
702	609	608	−230.9390
702	609	108	−341.8228
610	609	608	308.4848
610	609	108	423.1181
609	610	110	−242.2106
608	609	108	360.0264
609	608	802	−275.0995
609	608	702	−123.0615
110	610	110	291.9243
802	608	702	221.5866
608	702	606	−39.2249
608	702	104	−151.3333
606	702	104	190.9834
702	606	107	−110.7396
107	606	107	240.3135

TABLE VII. FQ $J^{(1-3)}$ parameters from the bond-space fit to the alanine dipeptide.

i	j	k	J_{ik}
607	618	702	124.7523
618	702	609	140.9889
618	702	104	−109.4704
102	607	618	−37.8984
607	618	802	21.9929
702	609	610	113.3700
702	609	608	126.6616
702	609	108	−10.3136
802	618	702	−98.9855
609	610	110	39.9734
609	608	802	−3.5818
609	608	702	88.0342
104	702	609	−74.3920
610	609	608	122.6962
608	702	606	90.6942
608	702	104	−88.2562
108	609	610	54.0549
108	609	608	−6.5267
802	608	702	−53.5582
702	606	107	−23.3979
104	702	606	18.3017

TABLE VIII. FQ $J^{(1-2)}$ parameters for bond pairs that define torsional angles in the alanine dipeptide. Atoms of types j and k form the central bond of the torsional angle.

i	j	k	l	J_{jk}
607	618	702	609	−256.5389
607	618	702	104	−126.4916
102	607	618	702	59.8913
102	607	618	802	−21.9929
618	702	609	610	−199.2578
618	702	609	608	−193.4015
618	702	609	108	−63.4588
802	618	702	609	116.0067
802	618	702	104	−17.0212
702	609	610	110	−2.6303
702	609	608	802	−24.4432
702	609	608	702	−181.9878
104	702	609	610	−50.0440
104	702	609	608	77.4932
104	702	609	108	−53.1452
610	609	608	802	38.7759
610	609	608	702	161.9010
608	609	610	110	11.4512
609	608	702	606	−166.5722
609	608	702	104	−103.7836
108	609	610	110	−54.0549
108	609	608	802	−59.6373
108	609	608	702	66.1640
802	608	702	606	69.0856
802	608	702	104	15.5274
608	702	606	107	−41.6996
104	702	606	107	−18.3017

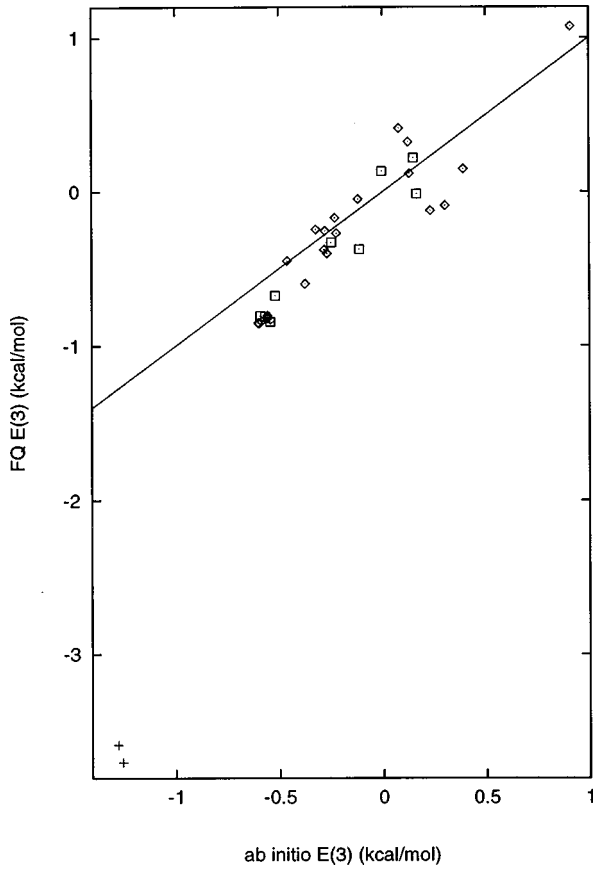

FIG. 1. Comparison of HF/6-31G** and FQ three-body energies for the alanine dipeptide with two fixed-charge dipole probes. The + symbols correspond to trimers in which both probes interact with the same dipeptide atom. Among the other data points, (◇) corresponds to the six dipeptide conformations used in fitting the FQ model, and (□) corresponds to two additional conformations.

units, indicating that M_{lm} matrix elements fit to the initial six conformations are accurate for conformations elsewhere in (ϕ,ψ) space.

We have also tested the alanine dipeptide FQ model by calculating three-body energies, which in the FQ model result from the polarization charges induced in one molecule by the field of others. For the six conformations used in the fit and the two additional ones mentioned above, we compared these energies with quantum-chemical calculations of the same quantities, which we performed using the JAGUAR package.[36] To compute three-body energies, we place two dipolar probes at various hydrogen-bonding locations around the dipeptide, in each of the six conformations. If we designate the dipeptide as molecule 0 and the probes as molecules 1 and 2, and write $E_{ab\ldots}$ for the total energy of molecules *a*, *b*,..., in the same configuration in which they appear in the trimer, then the three-body energy $E(3)$ is:

$$E(3)=E_{012}-E_{01}-E_{02}-E_{12}+E_0+E_1+E_2. \tag{25}$$

In order to test only the polarization response of the dipeptide rather than its zero-field charge distribution, we use fixed charges for the probes rather than full molecules with their own FQ parameters. In this case, the two probes do not affect each other's charges or energies, so E_{12} is always equal to E_1+E_2, and we need calculate only four of the seven terms in Eq. (25). We discuss the specification and effects of the zero-field charges of the dipeptide in Sec. III D 1.

Figure 1 compares the three-body energies in the FQ model to the corresponding *ab initio* values. Except for the two points represented by plus signs, the agreement is excellent, with an rms error (in the absence of these points) about 0.21 kcal/mol over all eight conformations studied. Examining the trimer configurations, we find that both ''bad'' structures have both probes in position to interact with the same dipeptide atom, as shown in Fig. 2 for the ''$C7_{ax}$'' conformation. In previous work,[17] our group has found similar discrepancies in three-body energies for such ''bifurcated hydrogen bond'' configurations. Physically, we would expect the minimum energy for such a configuration to occur when the dipeptide atom's dipole vector points in a direction ''in between'' the two probes. A model such as the current one, with point charges on atomic centers, clearly cannot reproduce such behavior no matter how much the charges fluctuate. Preliminary results indicate that a model that also incorporates point dipoles will produce significantly better results for these trimers, as well as for those involving probes above or below the plane of aromatic rings.[30]

C. Polarizable electrostatic model for the serine dipeptide

We have also studied seven conformers of the serine dipeptide. Jorgensen provided the starting coordinates from

OPLS calculations, and Beachy in our laboratory optimized them at the HF/6-31G** level of theory.[37,35] Following Jorgensen, we label these structures ''ser_1''–''ser_5,'' ''ser_7,'' and ''ser_8.'' (Jorgensen's original ''ser_6'' conformation apparently merged with another structure in his calculations.) For this molecule, the appropriate number of ESP modes to cut appears to vary over the different conformations. The results we report here involved cutting between two and four positive-eigenvalue modes in the zero-field charge distribution, and six or seven for the polarization charges.

In addition to fitting FQ parameters to the serine data directly, we have used this dipeptide to explore the transferability of the alanine FQ model. In this test, we constrained the FQ parameters for interactions involving only backbone atoms to be equal to the parameters for the same atom types that we obtained from fitting the alanine dipeptide. In another fit, we used parameters for the serine sidechain atoms that we obtained from an FQ fit to ethanol, in addition to the alanine parameters for the backbone. In this case, the only free parameters in the model were those describing interactions between backbone and sidechain atoms. In both of these transferability checks, the fixed parameters we used were from fits including nonlinear refinement. Table IX shows the rms and maximum absolute errors for each of these three fits, both before and after nonlinear refinement of the remaining variable parameters (in the first case, all the parameters) to fit the serine data. Interestingly, nonlinear refinement of the fit to the alanine data results in backbone parameters that fit the serine data *better* than those resulting from fitting the serine data directly but omitting the nonlinear refinement step. Thus nonlinear refinement may improve not only the fit to a given molecule, but also transferability to other molecules. On the other hand, nonlinear refinement of the remaining parameters does not significantly improve the fit, and in particular does not give as good results as nonlinear refinement of all the parameters from the initial linear fit directly to serine.

TABLE IX. rms and maximum absolute errors (ECU) in serine dipeptide polarization charges, from bond-space FQ models with or without some parameters fixed at values from fits to other molecules, and with or without nonlinear refinement of the variable parameters.

Fixed param?	Refinement?	rms error	Max |error|
None	no	0.0169	0.1918
	yes	0.0044	0.0394
Alanine	no	0.0061	0.0535
	yes	0.0057	0.0554
Alanine+ethanol	no	0.0070	0.0581
	yes	0.0060	0.0568

Even with all parameters fit directly to serine, and nonlinear refinement, the fit is not quite as good as we obtained for alanine. But Table X shows that most of the largest errors in polarization charges in this model are *underestimates*, rather than overestimates, of the quantum-chemical values. Such errors are less likely than overestimates to lead to catastrophic instabilities in simulations.

Figure 3 shows three-body energies for the model directly fit to serine data, plotted against the corresponding *ab initio* energies. As for alanine, there are several ''bad'' points, which again correspond to trimer structures in which both probes interact with the same dipeptide atom. Without these points, the rms error is about 0.14 kcal/mol.

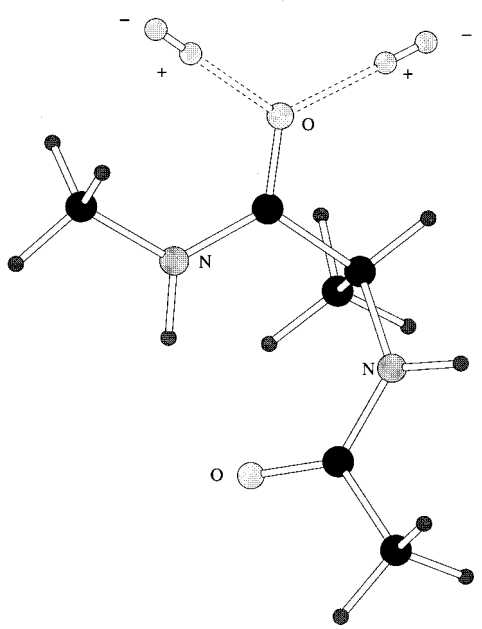

FIG. 2. Alanine dipeptide $C7_{ax}$ conformation, with two dipolar probes, both in positions to have hydrogen bond-like interactions with the same oxygen atom. The point-charge FQ model cannot accurately reproduce the three-body energy of this and similar trimers, resulting in the ''bad'' points in Figs. 1 and 3.

D. Complete polarizable force field for the alanine dipeptide

1. Zero-field charges for the FQ model

The first step in constructing a complete polarizable force field for the alanine dipeptide is to decide how to specify the zero-field charges in the FQ model i.e., how to

TABLE X. Errors>0.025 ECU (absolute value) in the FQ fit to serine dipeptide polarization charges.

Conf.	Data set	Site	$\delta q^{(\mathrm{FQ})}$	$\delta q^{(\mathrm{QM})}$	Error
ser_1	1	9	0.001 24	−0.024 63	0.025 87
	1	10	0.000 91	0.029 41	−0.028 50
ser_2	2	3	−0.000 81	−0.027 51	0.026 70
	2	9	0.001 18	0.027 96	−0.026 78
	2	17	0.005 89	−0.024 58	0.030 47
	2	18	0.000 10	0.039 50	−0.039 40
ser_3	1	13	−0.000 16	0.033 02	−0.033 18
ser_4	3	9	−0.011 29	0.017 30	−0.028 59
	3	18	−0.001 50	0.023 67	−0.025 17
ser_5	2	12	−0.118 23	−0.143 33	0.025 10
	2	18	0.003 09	0.029 57	−0.026 48
ser_7	2	12	−0.118 28	−0.088 06	−0.030 22
	2	18	0.008 42	−0.023 73	0.032 15
ser_8	2	12	−0.118 22	−0.144 44	0.026 22
	2	18	0.002 95	0.029 00	−0.026 05

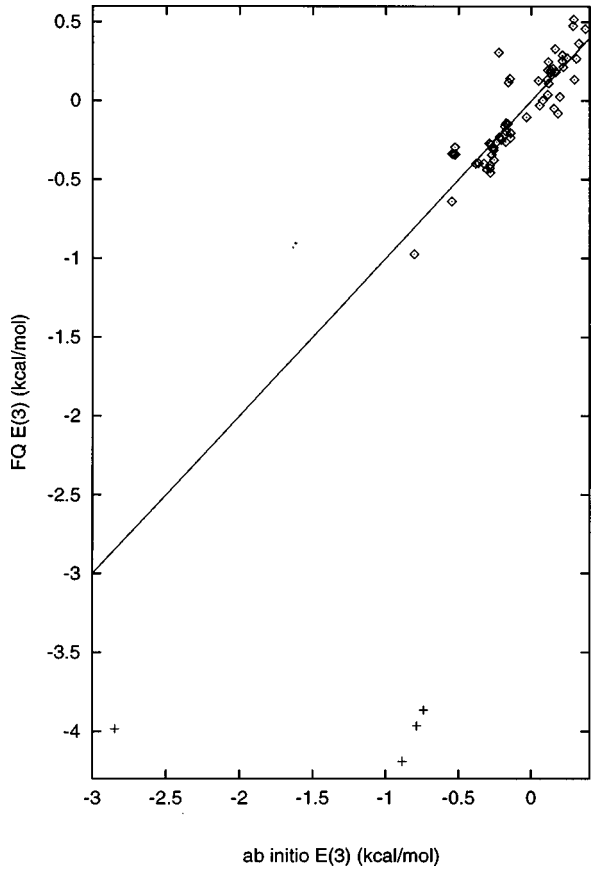

FIG. 3. Comparison of HF/6-31G** and FQ three-body energies for the serine dipeptide with two fixed-charge dipole probes. (+) corresponds to trimers in which both probes interact with the same dipeptide atom.

define the electronegativities. One possible strategy is to calculate ESP-fit charges at a high level of correlated *ab initio* quantum chemistry and, using a least-squares formulation, fit the FQ electronegativities to reproduce these charges as well as possible over the range of dipeptide minima. This would be appropriate if we were assembling a force field entirely from scratch. In this case, we would fit all electrostatic parameters to *ab initio* data, and fit van der Waals parameters to liquid-state thermodynamic data, as in the work of Jorgensen and co-workers.[38,1]

In this paper, however, our goal is proof of concept rather than immediate development of a production-level force field, and the liquid-state simulations and refitting would be a demanding task. We therefore use the van der Waals parameters from the existing OPLS-AA force field.[1] This force field successfully reproduces molecular properties including conformational equilibria, heats of vaporization for neat liquids, and free energies of hydration. It also performed near the top in our group's earlier tests of fixed-charge force fields.[31] We cannot simply use *ab initio* charges with OPLS-AA van der Waals parameters, however, primarily because the hydrogen bonding interaction between the carbonyl oxygen and N–H group (crucial for alanine polypeptide structures and energetics) requires a balance between the van der Waals and electrostatic terms. Therefore, we adopt a different strategy here, one that may even prove to be viable in development of a production-quality model: we fit the electronegativities to reproduce OPLS-AA charges in a dipeptide conformation with an internal hydrogen bond. We hope that this protocol will properly model the key hydrogen bonding interactions when the molecule is in the appropriate geometry. The test of the protocol is the ability of the resulting dipeptide force field to reproduce the tetrapeptide energetics. Table XI gives the electronegativity values that result from combining the OPLS-AA charges for the alanine dipeptide with the FQ matrix elements of Tables V–VIII. Since we obtained those parameters from a bond-space fit, the electronegativity parameters are χ_l', the difference in electronegativity between the atoms i and j joined by the bond l. Since adding a constant to all electronegativities does not affect relative FQ energies, these differences are sufficient to define the model. In order to obtain a unique parameter χ_l' for each pair of atom types, we perform a least-squares fit in cases where the molecule contains multiple atoms of the same type.

TABLE XI. FQ bond-space electronegativity differences [kcal/(mol×ECU)] for atom types i and j joined by bond l, in the alanine dipeptide. These result from applying the Coulomb matrix elements for the $C7_{ax}$ conformation to the OPLS-AA charges.

i	j	χ_l'
607	618	237.513 322
607	102	131.822 982
618	702	−256.290 958
618	802	−242.078 425
702	609	367.391 441
702	104	293.722 543
609	610	−135.837 733
609	608	−107.168 504
609	108	−77.927 204
610	110	23.123 078
608	802	−184.034 028
608	702	−195.146 530
702	606	177.875 327
606	107	24.447 998

2. Reproducing the alanine dipeptide potential surface

Once we have specified our polarizable electrostatic model as above, and chosen to use the stretching, bending, torsional, and van der Waals functional forms from OPLS-AA,[1] we proceed to make any necessary modifica-

TABLE XII. Relative energies (kcal/mol) of alanine dipeptide conformations in the OPLS-FQ model (present work), with *ab initio* and standard OPLS-AA results for comparison. For both force fields, the conformations with ''—'' in the energy columns were not local minima of the potential, but instead ''decayed'' during molecular mechanics optimization to lower minima nearby.

Conf.	*ab initio*	OPLS-FQ	OPLS-AA
$C7_{eq}$	0.00	0.00	0.00
C5	0.95	0.86	1.31
$C7_{ax}$	2.67	2.11	2.55
β_2	2.75	—	—
α_L	4.31	4.97	—
α'	5.51	4.63	6.49

TABLE XIII. rms deviations in geometry (Å) and energy (kcal/mol) for the alanine tetrapeptide, as compared to LMP2/cc-pVTZ(-f) energies at HF/6-31G** geometries. The force field energies are at local minima on each force field surface (from unrestrained optimization), and the rms energy deviation is over ten tetrapeptide conformations. The geometry rms is the average over the ten conformations of the all-atom rms deviation between the force field and HF/6-31G** coordinates. Results are reproduced from Table 5 of Ref. 31, with results of the present work added in bold face type. The null hypothesis is the case for which all conformers are equal in energy.

Force field/Method	Geom. rms	Energy rms
LMP2/cc-pVTZ(-f)	···	0.00
OPLS-FQ	**0.38**	**0.94**
HF/6-31G**	0.00	1.10
MMFF93	···	1.20
MMFF	0.32	1.24
OPLS-AA(2,2)	0.16	1.31
MMFFs	0.24	1.40
OPLS/A-UA(2,8)	0.18	1.43
MM2X($\epsilon=1.5$)	0.51	1.49
MM3*	0.48	1.53
OPLS*	0.49	1.55
MM3*($\epsilon=1.5$)	0.45	1.58
GROMOS	0.39	1.60
HF/cc-pVTZ(-f)	···	1.69
MMFF($\epsilon=1.5r$)	0.27	1.75
CFF95	0.41	1.86
MM3*($\epsilon=1.0r$)	0.48	2.00
Null hypothesis	···	2.07
OPLS-UA(2,2)	0.26	2.19
AMBER*	0.48	2.39
MSI CHARMm	0.40	2.54
MMFF($\epsilon=2.0r$)	0.29	2.56
CHARMM 22	0.86	2.56
CHARMM 19	0.76	2.73
AMBER*($\epsilon=1.0r$)	0.47	2.75
AMBER 4.1	0.55	3.35
AMBER94	0.58	3.42
CVFF	0.98	3.91
AMBER 3	0.35	4.17
MM2*($\epsilon=1.5$)	0.96	4.91
MM2*	0.80	6.09
MM2*($\epsilon=1.0r$)	0.83	6.14

tions to OPLS-AA parameters in order to reproduce the relative energies of the alanine dipeptide conformers. Although, as noted above, for compatibility with OPLS-AA van der Waals parameters we have chosen FQ electronegativities based on the OPLS-AA charges, the variability of the charges in our model may lead to ''overcorrection'' of effects that using the OPLS-AA ''permanent'' charges already corrects for. In the dipeptide $C7_{eq}$ and $C7_{ax}$ conformers, for instance, standard OPLS-AA produces internal hydrogen bond distances that are 0.1–0.2 Å longer than the *ab initio* values, but accurately reproduces the energetics of the hydrogen bonds. With fluctuating charges, the polarization of the hydrogen bonds increases, resulting in distances 0.1–0.2 Å *shorter* than the *ab initio* values. As a result, the hydrogen bond energy becomes more attractive than in the standard OPLS-AA force field, and with the rest of the force field unchanged, this leads to discrepancies in the energies of these two conformers relative to those that lack the internal hydrogen bond. To counteract this effect, we introduce nonzero Lennard-Jones parameters for the polar peptide hydrogens, which have no van der Waals interactions in standard OPLS-AA. The values of these parameters, however, are considerably smaller than the corresponding ones for aliphatic hydrogens: the radial parameter $\sigma=1.5$ Å rather than 2.5, and the well depth $\epsilon=0.02$ kcal/mol rather than 0.03. For the following results, we made no other modifications of the nonelectrostatic parameters of OPLS-AA.

TABLE XIV. Relative conformational energies (kcal/mol) for ten alanine tetrapeptide conformations. The *ab initio* column contains LMP2/cc-pVTZ(-f) energies, taken from the first line of Table 6 in Ref. 31. For the OPLS-FQ (present work) and OPLS-AA results, the zero of energy is adjusted to minimize the rms deviation from the *ab initio* results.

Conf.	*ab initio*	OPLS-FQ	OPLS-AA
1	2.71	3.03	2.78
2	2.84	3.97	2.50
3	0.00	0.26	−1.34
4	4.13	2.34	3.48
5	3.88	4.62	4.46
6	2.20	1.67	3.31
7	5.77	6.18	3.82
8	4.16	4.39	6.93
9	6.92	5.33	5.90
10	6.99	7.82	7.78

Table XII gives our results for the dipeptide minima, along with the *ab initio* values and the results obtained from the standard OPLS-AA force field. Here and in Sec. III E, we use the Fletcher–Powell optimization algorithm in the BOSS program,[39] modified to use FQ electrostatics in the force field. For the conformers that are the local minima of OPLS-AA, the relative energies in the OPLS-FQ model are approximately as close to the *ab initio* values as those in OPLS-AA (the rms deviation $\simeq$0.6 kcal/mol in both cases). Furthermore, the α_L conformer, which is not a local minimum of OPLS-AA, is a local minimum of OPLS-FQ, whose energy relative to the global minimum $C7_{eq}$ is within 0.7 kcal/mol of the *ab initio* value. We were able to achieve slightly better dipeptide energetics (rms $\simeq$0.4 kcal/mol) by refitting torsional parameters, but the resulting model was unsuccessful in reproducing tetrapeptide energetics. Preliminary results indicate that achieving a substantially better fit requires a better description of the charge distribution, such as one incorporating point dipoles.

E. Calculation of tetrapeptide energetics

We now use the polarizable electrostatic model described above, combined with OPLS-AA parameters for other interactions, to calculate relative energies of ten alanine tetrapeptide conformers. As for the dipeptide, we start from the standard OPLS-AA bond-stretch, angle, torsional, and Lennard-Jones parameters, and fit the electronegativities in the fluctuating charge model to the OPLS-AA charges. We assign the same Lennard-Jones parameters to the peptide polar hydrogens as for the dipeptide, for the same reason. For the tetrapeptide, we find that the original OPLS-AA torsional parameters give the best relative energies, so the following results do not involve any torsional refitting.

In previous work, our group has described the test set of tetrapeptide conformations, and the comparison of results for

a wide range of fixed charge force fields with accurate quantum chemistry.[31] That work involved generating ten conformations of the tetrapeptide by molecular mechanics search, then optimizing their geometries at the HF/6-31G** level and calculating their energies at the LMP2/cc-pVTZ(-f) level. Table XIII reproduces the rms structure and energy deviations from Table 5 of Ref. 31 for the force fields tested therein, along with the OPLS-FQ results. Table XIV presents the individual relative energies for each of the ten conformers, as computed with OPLS-FQ and standard OPLS-AA, compared with the LMP2/cc-pVTZ(-f) results of Ref. 31. The performance of OPLS-FQ in this test is comparable to that of the original OPLS-AA, and superior to most of the force fields examined in Ref. 31. As with the dipeptide, preliminary data indicate that improving the rms energy deviation beyond the level of ~1 kcal/mol will require a better description of the molecular charge distribution.

IV. CONCLUSION

We have developed a polarizable force field for polyalanine, based on the FQ model, that reproduces both the intramolecular energetics and the many-body polarization response of *ab initio* quantum mechanics with reasonable accuracy. We have demonstrated the transferability of the force field by applying parameters fit to the alanine dipeptide both to a larger system and to another amino acid. These results suggest that construction of a full polarizable force field for proteins, including all 20 amino acids, is a viable goal.

As we have indicated above, we believe that problems in our current technology arise from using a point-charge electrostatic model. We have found a significant number of examples in which the point-charge model is clearly inadequate, and at least dipoles will be necessary in order to improve the results.[30] Our conclusion that point charges are insufficient is analogous to that of York and Yang, who describe an electronegativity (or chemical potential) equalization model using Gaussian basis functions to describe the charge density.[40] They found that atom-centered spherically symmetric (*s*-type) functions, which have the symmetry of point charges, were inadequate to characterize the polarization response to the set of fields they studied, while adding *p*-type functions, which have the vector character of dipoles, was sufficient to model such phenomena as polarization of linear or planar molecules in directions orthogonal to their geometry. We expect that we can use the same fitting methods as in the present work for a linear response model incorporating both point charges and dipoles. We are currently working on this straightforward extension of our model.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health under Grants Nos. 5-R01-GM52018 and P41-RR06892. The authors thank Professor W. L. Jorgensen for the BOSS and CHEMEDIT programs and for serine dipeptide structures, and Dr. M. D. Beachy for *ab initio* data and for assistance in using the JAGUAR package.

[1] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
[2] W. D. Cornell *et al.*, J. Am. Chem. Soc. **117**, 5179 (1995).
[3] J. R. Maple, M.-J. Hwang, T. P. Stockfish, U. Dinur, M. Waldman, C. S. Ewig, and A. T. Hagler, J. Comput. Chem. **15**, 161 (1994).
[4] M.-J. Hwang, T. P. Stockfish, and A. T. Hagler, J. Am. Chem. Soc. **116**, 2515 (1994).
[5] A. D. MacKerell, Jr., J. Wiórkiewicz-Kuczera, and M. Karplus, J. Am. Chem. Soc. **117**, 11946 (1995).
[6] A. D. MacKerell *et al.*, J. Phys. Chem. B **102**, 3586 (1998).
[7] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Huenenberger, P. Krueger, A. E. Mark, W. R. P. Scott, and I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Vdf Hochschulverlag AG an der ETH-Zuerich, Zuerich, Switzerland, 1996).
[8] N. L. Allinger and L. Yan, J. Am. Chem. Soc. **115**, 11918 (1993).
[9] T. A. Halgren, J. Comput. Chem. **17**, 490 (1996).
[10] T. A. Halgren, J. Comput. Chem. **17**, 520 (1996).
[11] T. A. Halgren, J. Comput. Chem. **17**, 553 (1996).
[12] T. A. Halgren and R. B. Nachbar, J. Comput. Chem. **17**, 587 (1996).
[13] T. A. Halgren, J. Comput. Chem. **17**, 616 (1996).
[14] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard, III, and W. M. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992).
[15] C. J. Casewit, K. S. Colwell, and A. K. Rappé, J. Am. Chem. Soc. **114**, 10035 (1992).
[16] B. Marten, K. Kim, C. Cortis, R. A. Friesner, R. B. Murphy, M. N. Ringnalda, D. Sitkoff, and B. Honig, J. Phys. Chem. **100**, 11775 (1996).
[17] Y.-P. Liu, K. Kim, B. J. Berne, R. A. Friesner, and S. W. Rick, J. Chem. Phys. **108**, 4739 (1998).
[18] D. T. Mainz, Ph.D. thesis, Columbia University, New York, 1996.
[19] A. K. Rappé and W. A. Goddard, III, J. Phys. Chem. **95**, 3358 (1991).
[20] S. W. Rick, S. J. Stuart, and B. J. Berne, J. Chem. Phys. **101**, 6141 (1994).
[21] L. X. Dang, J. E. Rice, J. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **113**, 2481 (1991).
[22] D. N. Bernardo, Y. Ding, K. Krogh-Jespersen, and R. M. Levy, J. Phys. Chem. **98**, 4180 (1994).
[23] Y. Ding, D. N. Bernardo, K. Krogh-Jespersen, and R. M. Levy, J. Phys. Chem. **99**, 11575 (1995).
[24] R. T. Sanderson, Science **114**, 670 (1951).
[25] R. T. Sanderson, *Chemical Bonds and Bond Energy*, 2nd ed. (Academic, New York, 1976).
[26] R. G. Parr, R. A. Donnelly, M. Levy, and W. E. Palke, J. Chem. Phys. **68**, 3801 (1978).
[27] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, Oxford, 1989).
[28] W. J. Mortier, K. Van Genechten, and J. Gasteiger, J. Am. Chem. Soc. **107**, 829 (1985).
[29] W. J. Mortier, S. K. Ghosh, and S. Shankar, J. Am. Chem. Soc. **108**, 4315 (1986).
[30] H. Stern *et al.* (in preparation).
[31] M. D. Beachy, D. Chasman, R. B. Murphy, T. A. Halgren, and R. A. Friesner, J. Am. Chem. Soc. **119**, 5908 (1997).
[32] L. E. Chirlian and M. M. Francl, J. Comput. Chem. **8**, 894 (1987).
[33] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[34] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, J. Phys. Chem. **97**, 10269 (1993).
[35] M. D. Beachy (personal communication).
[36] JAGUAR V3.0 (Schrödinger, Inc., Portland, OR, 1997).
[37] W. L. Jorgensen (personal communication).
[38] W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).
[39] W. L. Jorgensen, BOSS VERSION 3.6 (Yale University, New Haven, CT, 1995).
[40] D. M. York and W. Yang, J. Chem. Phys. **104**, 159 (1996).