JCTC Journal of Chemical Theory and Computation

The Gradient Curves Method: An Improved Strategy for the Derivation of Molecular Mechanics Valence Force Fields from ab Initio Data

T. Verstraelen,*,† D. Van Neck,† P. W. Ayers,‡ V. Van Speybroeck,† and M. Waroquier†

Center for Molecular Modeling, Ghent University, 9000 Gent, Belgium, and Department of Chemistry, McMaster University, Hamilton, ON L8S 4M1, Canada

Received June 23, 2006

Abstract: A novel force-field development strategy is proposed that tackles the well-known difficulty of parameter correlations arising in a conventional least-squares optimization. In the first step of the new gradient curves method (GCM), continuity criteria are imposed to transform the raw multidimensional ab initio training data to distinct sets of one-dimensional data, each associated with an individual energy term. In the second step, the transformed data suggest suitable analytical expressions, and the parameters in these expressions are fitted to the transformed data; that is, one does not have to postulate a priori analytical expressions for the force-field energy terms. This approach facilitates the derivation of valence terms. Benchmarks have been performed on a set of small molecules. The results show that the new method yields physically acceptable energy terms exactly when a conventional parametrization would suffer from parameter correlations, that is, when an increasing number of redundant internal coordinates is used in the force-field model. The generic treatment of parameter correlations in the proposed method facilitates an intuitive physical interpretation of the individual terms in the force-field expression, which is a prerequisite for the transferability of force-field models.

1. Introduction

The development of a molecular mechanics force field based on an ab initio parametrization is a tedious task plagued by model selection and parameter correlations, especially when one wants to extend its applicability to a broad range of molecular systems. The final goal of this study lies in the construction of an accurate all-atom zeolite-guest force field that is applicable both to unconstrained bulk zeolite structures and to unconstrained interfaces between zeolite nanoparticles and their environment. It is highly ambitious to assert that such a broad domain of applications can be covered by a single force-field model. Most of the force fields proposed in the literature can be used only for a subset of the applications mentioned above.[1−7] There are two reasons for the limited applicability of existing force fields. On the one hand, molecular mechanics models are limited, in general, to a specific domain of application due to the reduction of the full ab initio description of a molecule into a set of parametrized analytical energy terms. This failure is inherent to the nature of force-field models. On the other hand, the determination of reliable and transferable parameters for the analytical expressions in a force field is a nontrivial task. The main focus of this paper is the development of a reliable parametrization technique.

Parameter correlations, which are inherent to least-squares parametrization in general, represent the major difficulty in the development of force fields based on ab initio data. In the naive approach of an accurate force-field model, a large number of parameters should be introduced to describe all possible types of interactions. The optimization then usually leads to many degenerate solutions; that is, many disparate parameter sets have nearly the same goodness of fit. Only a

* Corresponding author e-mail: Toon.Verstraelen@UGent.be.
† Ghent University.
‡ McMaster University.

10.1021/ct6002093 CCC: $37.00
Published on Web 05/01/2007

very small number of these “good” fits are physically acceptable and transferable to molecules not belonging to the training set. In an attempt to fulfill these requirements, several techniques have been proposed in literature that select a physically meaningful and potentially transferable set of parameters yielding an acceptable goodness of fit.

(i) An intuitive procedure first parametrizes a coarse force field that only contains the most important energy terms, using a traditional least-squares method. Second, the residual error is further reduced by including corrective energy terms whose parameters are optimized without modifying the original coarse force field.[10] This approach keeps the contribution of the corrective energy terms small compared to the coarse force field, but the optimal goodness of fit is not reached. Moreover, only the correlations between parameters in the coarse force field and the corrective energy terms are treated. More generally, a global optimization is divided in smaller piecewise optimizations, to make the parametrization more tractable. A piecewise optimization only considers a subset of variable parameters, for example, the parameters associated with all bond-stretch terms. The global optimum is then approximated with a limited number of iterations in which each subset of parameters is optimized or reoptimized as to give an optimal fit with respect to the training data and the other subsets of parameters that are kept fixed.[8,9]

(ii) Another procedure avoids degeneracies in the first-order energy terms (i.e., correlations between first-order force constants and reference coordinates) by imposing constraints on their coefficients,[2] but degeneracies in higher-order terms are neglected.

(iii) The most systematic approach adds quadratic penalty functions to the χ^2 cost function.[11] With each parameter, a penalty function is associated that restrains this parameter to a physically acceptable value. This regularization technique is similar to restrained electrostatic potential fitting.[12] Unfortunately, one must choose the weight for each penalty function to be small enough so that the penalty functions only make small contributions to the total cost function but large enough so that the parameters are forced to retain a physically reasonable value. This “weight determination problem” is also ill-conditioned. Essentially, one has just replaced one ill-conditioned problem (parameter fitting) with another one (weight determination).

(iv) Parameter correlations can also be avoided by reducing the number of parameters in a force-field model.[1] One has to select carefully the energy terms that can be omitted and the analytical form of the retained terms. The disadvantage of this approach is that the absence of some molecular interaction terms in the force field will be compensated by biased parameters in the retained terms. Consequently, it is a common practice to exclude the atomic charges from the optimization procedure and to assign formal charges to these atoms instead; this prevents unphysical atomic charges.

(v) The most extreme approach in this comparative overview is represented by the rule-based force fields that do not contain fitted parameters.[13−15] All parameters are directly derived from semiempirical rules or are estimated on the basis of common sense. Such force fields sacrifice accuracy to achieve transferability.

Except for the second method, all the techniques mentioned above require additional subjective choices to tackle the problem of parameter correlations: the separation of coarse- and fine-grained components, a vast amount of weight factors, and so forth. Only the third method is truly systematic since it treats all parameter correlations, but it depends on a series of manually tuned weight factors.

This work aims to present a new force-field parametrization procedure—the gradient curves method (GCM)—which is innovative in its concept and which addresses the main concerns raised above. First, the method does not rely on subjective choices, for example, predefined analytical expressions for the energy terms, manually tuned parameters, repetitive parametrizations where at each iteration some parameters are included or excluded, and so forth. Second, the new method treats the problem of parameter correlations in a rigorous way. The only input is a set of ab initio training data and a list of the internal coordinates that will be used in the force-field model.

The gradient curves method is designed to extract the maximum amount of information from the ab initio training data set. A two-step procedure is used to achieve this objective. The first step encompasses a transformation of the raw multidimensional ab initio data into distinct one-dimensional data sets, each associated with a single energy term. During this transformation, a consistent treatment of parameter correlations guarantees a unique and physically acceptable series of transformed data sets. In this context, “physically acceptable” indicates that it is possible to give an intuitive physical interpretation to the individual transformed data sets. The analytical expressions enter the procedure only in the second step, where they can be easily estimated from the transformed data sets and may be modeled with nonlinear parameters without major difficulties.

For several reasons, the present version of the gradient curves method is less appropriate to parametrize long-range interactions. These interactions (i.e., the classical electrostatic and the dispersion interactions) obey well-known physical laws. Therefore, it would be highly inefficient to derive these long-range interactions without relying on their asymptotic behavior during the first step of the new method. Specific parametrization techniques for chemically accurate electrostatic models have already been actively studied during the past decades.[12,16−18] Due to the enormous computational cost of post-Hartree−Fock ab initio calculations that describe dispersion interactions properly,[19] it is more efficient to use such calculations specifically for the parametrization of dispersion interactions.[20,21]

Most of the ingredients of the gradient curves method are new, but the idea to express a multivariate function in terms of functions depending on a smaller number of variables is frequently applied. We refer to the high dimensional model representation[25] (HDMR) which has been applied in several fields, ranging from molecular modeling[26] to global atmospheric models.[27] This technique guarantees a unique multivariate expansion; that is, it treats parameter correlations, by imposing orthogonality constraints between all the components in the expansion. HDMR is very efficient when the primary concern is only to reproduce a given set of training data. The end result is an efficient and reliable input−output model. At this point, our focus is different; that is, we would like to ensure that all the distinct energy terms are physically intuitive instead of orthogonal. Less popular black-box approaches where the expansion consists solely of one-dimensional functions[28,29] are based on Kolmogorov's solution[30] to Hilbert's 13th problem[31] and rely on nonsmooth component functions.

The applications in this paper are limited to a set of small molecules such as H_2O, NH_3, and CH_4. For the short-range aspects of interest, this is sufficient to illustrate and benchmark the new method. The aim of these examples is not to obtain transferable force-field parameters for these three molecules but rather to show how the prerequisites for transferability can be met. Additionally, it is not the intention to derive definitive force-field parameters for these three molecules that can be directly tested against experimental data, but we focus on the aspect of how well a reasonable force-field model can simulate a given set of ab initio calculations. We have intentionally generated ab initio training data for these molecules that include a significant portion of the anharmonic part of the potential energy surface, in order to test to what extent the gradient curves method is capable of parametrizing force fields that also reproduce the nonharmonic part of the potential energy surface of the three benchmark molecules. Work is in progress to extend the applicability of the gradient curves method to larger systems, taking into account long-range interactions.

The remainder of this article is organized as follows. In section 2, the new procedure is derived. The benchmark protocol that evaluates the merits of this new procedure is presented in section 3. Section 4 discusses the results obtained by the benchmarks. Finally, conclusions are given in section 5.

2. Gradient Curves Method

2.1. Outline. For the sake of simplicity, we limit ourselves to force fields of the class-I form:

$$E_{\mathrm{FF}} = \sum_{k=1}^{K} E_k(q_k) \tag{1}$$

where $K > 3N - 6$ and $N =$ the number of atoms. The force-field energy E_{FF} of a molecular geometry is expressed as a sum over functions E_k of only one internal coordinate q_k, where the q_k's are not restricted to the $(3N - 6)$ molecular degrees of freedom and may stand for a redundant set of internal coordinates. The redundancy originates from the observation that even a coarse valence force field[13,14] includes terms for all bond lengths, all bending angles, and some dihedral angles. Force fields that are accurate in the prediction of both structural and vibrational properties have to include cross terms $E_{k1,k2}(q_{k1},q_{k2})$ in the force-field expression.[22] In class-II force fields,[23] this is resolved by adding functions that depend on products of internal coordinates, that is, $q_{k1}q_{k2}$. We prefer to label products and other constructions of internal coordinates as new internal coordinates, which allows us to work with the class-I form in eq 1. This implies that for accurate force fields $K \gg 3N - 6$.

As a consequence of the redundancy, a direct fit of parametrized expressions for the E_k to a set of ab initio training data contains severe parameter correlations even when an abundant amount of training data is available. By selecting one arbitrary set of parameters that minimizes the residual errors, the resulting force field contains energy terms with an unphysical behavior and consequently lacks transferability.[2,11] Similar considerations about redundant internal coordinates in the theory of molecular vibrations have led to the canonical force-field concept, which is useful for the analysis of vibrational spectra.[24]

The detailed mathematical derivation of the gradient curves method will be presented in the next subsection. We now continue with a general outline of the method. The training data used in the gradient curves method are the ab initio calculated gradients for M different geometries of a given molecule

$$Y_i^{(m)} = \left(\frac{\partial E_{\mathrm{AI}}}{\partial x_i}\right)_{x=x^{(m)}} \tag{2}$$

where $m = 1...M$ and $x^{(m)}$ is the vector that contains all the Cartesian coordinates of the atoms in geometry m. For an energy surface of the class-I form in eq 1, one factorizes the Cartesian gradient for geometry m according to

$$G^{(m)} = J^{(m)} g^{(m)} \tag{3}$$

where the matrices in expression 3 are defined as

$$G_i^{(m)} = \left(\frac{\partial E_{\mathrm{FF}}}{\partial x_i}\right)_{x=x^{(m)}}$$
$$g_k^{(m)} = \left(\frac{\partial E_{\mathrm{FF}}}{\partial q_k}\right)_{q=q^{(m)}} = \left(\frac{dE_k}{dq_k}\right)_{q_k=q_k^{(m)}}$$
$$J_{i,k}^{(m)} = \left(\frac{\partial q_k}{\partial x_i}\right)_{x=x^{(m)}} \tag{4}$$

The convention for matrix notation in this article uses upper indexes to indicate different matrices and lower indexes to identify the matrix elements; for example, $G^{(m)}$ and $g^{(m)}$ are column matrices of dimension $3N$ and K, respectively, whereas $J^{(m)}$ is a rectangular matrix of dimensions $3N \times K$.

Since we want to find a suitable class-I representation of the true (ab initio) energy surface E_{AI} sampled in M geometries, we first identify the Cartesian gradient of the force-field energy in expression 1 with the ab initio training data

$$G_i^{(m)} \equiv Y_i^{(m)} \tag{5}$$

and try to solve the linear system

$$Y^{(m)} = J^{(m)} y^{(m)} \tag{6}$$

for the "ab initio gradient in internal coordinates", $y^{(m)}$. Due to the redundancy of the coordinates q_k, this equation has many solutions, that is, a particular solution plus an arbitrary vector from the null space of $J^{(m)}$. Step I of the gradient curves method determines which vector from the null space must be taken for each geometry by an optimization procedure. In other words, the first step defines how the ab initio training data are transformed into one-dimensional data sets of the form $D_k = \{(q_k^{(m)}, y_{k(\mathrm{opt})}^{(m)}) | m = 1...M\}$. Through the identification $y_k^{(m)} \equiv g_k^{(m)}$, or $y_k^{(m)} \equiv (dE_k/dq_k)_{q_k=q_k^{(m)}}$, step II of the gradient curves method consists of proposing a functional form for the derivative of each energy term, (dE_k/dq_k), based on its corresponding transformed data set, D_k, and the expected asymptotic behavior. Finally, each functional form can be fitted to its corresponding data set with conventional fitting procedures.

The purpose of the transformation in step I of the gradient curves method is to make step II as successful as possible. This means that—for each geometry—the vector from the null space will be taken so as to optimize the continuity conditions of the data sets D_k. In practice, this is achieved by selecting the solutions of eq 6 for all geometries that minimize a cost function, Z, which is a measure for the continuity of the data sets D_k. In this work, continuity is measured by the goodness of fit of a generic high-order polynomial to a set of data points.

Unfortunately, this continuity requirement alone will in general not result in a uniquely defined transformation. In other words, the cost function, Z, as a function of the solutions of eq 6, can have a degenerate minimum. It will be shown in the next subsection that the transformation will always be ill-defined when the number of energy terms, K, is much larger than the number of independent internal degrees of freedom, $3N - 6$. To guarantee a unique minimum, we must introduce additional but subordinate criteria that will select from all the possible transformations to continuous data sets the one solution that corresponds optimally to what we expect from physical intuition. In this work "physical intuition" is interpreted as "having minimal forces along the internal coordinates". This prescription can be implemented as a least-norm criterion on the y values of the data sets D_k, in addition to the continuity criterion. Formally, such a least-norm criterion is implemented as an extra term in the cost function $Z^* = Z + \epsilon L$, where ϵ is a very small positive number and L is the contribution from the least-norm criterion. For small values of ϵ, the minimum of the new cost function approximately also minimizes the original cost function. This least-norm criterion is also known as zeroth-order regularization, and—as shown in the next subsection—it ensures that the transformation is always uniquely defined.

In order to understand the remainder of this paper, it is not strictly required to read the next subsection which describes the detailed mathematical derivation of the gradient curves method. Nevertheless, it is highly recommended for a deeper understanding, and mandatory when one is interested in implementing or extending the method.

2.2. Detailed Procedure. Since step II is a standard fitting procedure, we now concentrate on the details of step I. The general solution of the linear system (6) is given by

$$y^{(m)} = p^{(m)} + \mathcal{N}^{(m)} s^{(m)} \tag{7}$$

where $p^{(m)}$ is a particular solution, $\mathcal{N}^{(m)}$ is a matrix with orthogonal columns spanning the null space of the Jacobian $J^{(m)}$, and the vector $s^{(m)}$ contains arbitrary coefficients that determine which vector from the null space is added to the particular solution. One can derive the particular solution and the null space of a given linear system through the singular value decomposition algorithm.[32]

The coefficients $s^{(m)}$ are fixed by imposing continuity criteria: we select the $s^{(m)}$'s that minimize the sum of squared residual errors, obtained in a linear fit of a set of generic auxiliary functions, $f_n(q_k)$ (e.g., polynomials), to the "ab initio gradient in internal coordinates", $y_k^{(m)}$,

$$y_k^{(m)} \overset{\text{fit}}{\equiv} \sum_n a_n^{(k)} f_n(q_k^{(m)}) \tag{8}$$

The sum of the squared residual errors in the fit to the data set D_k is given by the expression

$$R_k^2 = \sum_m (\sum_n a_n^{(k)} f_n(q_k^{(m)}) - y_k^{(m)})^2 \tag{9}$$

In this equation, and in the following analysis, we find it convenient to switch to a notation where the different matrix quantities are labeled by the index of the internal coordinates under scrutiny, k, for example,

$$F_{m,n}^{(k)} = f_n(q_k^{(m)}) \qquad \tilde{y}_m^{(k)} = y_k^{(m)} \tag{10}$$

In the revised notation, the sum of squared residuals (using standard manipulations) is

$$R_k^2 = (F^{(k)}a^{(k)} - \tilde{y}^{(k)})^T (F^{(k)}a^{(k)} - \tilde{y}^{(k)}) \tag{11}$$

Minimizing this expression with respect to the expansion coefficients, $a_n^{(k)}$, allows one to discern how well the gradient information can be represented by a continuous function. The least-squares expansion coefficients from eq 8 are given by the expression

$$a_{(\mathrm{opt})}^{(k)} = [F^{(k)T}F^{(k)}]^{-1} F^{(k)^T} \tilde{y}^{(k)} \tag{12}$$

and the residual error is

$$\min_{a_n^{(k)}} R_k^2 = \tilde{y}^{(k)^T}[1 - F^{(k)}(F^{(k)^T}F^{(k)})^{-1}F^{(k)^T}]\tilde{y}^{(k)} = \tilde{y}^{(k)^T} C^{(k)} \tilde{y}^{(k)} \tag{13}$$

which is indicative of the continuity of the data set D_k. Note that $C^{(k)}$ projects on the complement of the range of $F^{(k)}$. In analogy to eq 10, we can introduce relabeled matrix quantities

$$\tilde{p}_m^{(k)} = p_k^{(m)} \qquad \tilde{\mathcal{N}}^{(k)}_{m',\beta m} = \mathcal{N}^{(m)}_{k,\beta} \delta_{m',m} \qquad \tilde{s}_{\beta m} = s_\beta^{(m)} \tag{14}$$

in terms of which eq 7 can be rewritten as

$$\tilde{y}^{(k)} = \tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)} \tilde{s} \tag{15}$$

This allows a compact expression for the desired cost function, which is a weighted sum of the continuity measures of all the data sets D_k

$$Z(\tilde{s}) = \sum_k w_k^2 (\min_{a_n^{(k)}} R_k^2) = \sum_k (\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s})^T w_k^2 C^{(k)} (\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s}) \quad (16)$$

For the practical applicability of the gradient curves method, the weight factors w_k^2 which convert the R_k^2's to the dimension of an energy squared should be easy to obtain. A simple physical interpretation of w_kR_k is illustrated in Figure 1. R_k is the RMS error obtained by fitting the auxiliary functions $f_n(q_k)$ to the optimized data set D_k. One obtains a tentative energy term by integrating the fitted function $\sum_n a_n^{(k)} f_n(q_k)$ over the physically relevant interval $[q_k^{(\min)},q_k^{(\max)}]$. The error accumulated during this integration is equal to $(q_k^{(\max)} - q_k^{(\min)})R_k$. It always has the dimension of an energy, and it is a quality measure for the energy terms obtained by fitting functional forms for (dE_k/dq_k) to the data sets D_k. Therefore, it is both practical and acceptable to identify the conversion factor w_k with the width of the physically relevant interval of q_k. One can intuitively estimate w_k, or alternatively one can obtain these widths from the geometries in the training set if this training set is generated by a well-behaving and extensive sampling procedure. We have observed that the gradient curves method is insensitive to any reasonable changes in the values w_k, and that it is sufficient to estimate the correct order of magnitude.

The $\tilde{s}_{(\text{opt})}$ that minimizes expression Z can be substituted back into expression 7, after reordering this solution into vectors $s_{(\text{opt})}^{(m)}$. This yields the sets of data points D_k that are optimally continuous and thus slightly scattered around a continuous curve. The minimization of Z makes sure that this scattering is minimal. The selection of a suitable functional form for each E_k is easily accomplished by inspecting the scatter plots of the transformed data sets D_k.

Unfortunately, the solution $\tilde{s}_{(\text{opt})}$ is in general not unique. Since Z is a quadratic expression, only one global minimum exists, although that minimum can still be degenerate. In the case of a degenerate minimum, there is a subspace S that contains all the arguments of Z that yield the minimum value. The dimension of S is equal to the dimension of the null space of the matrix

$$\begin{aligned} \mathcal{H} &= \sum_k \tilde{\mathcal{N}}^{(k)T} C^{k} \tilde{\mathcal{N}}^{(k)} \\ &= \begin{pmatrix} \tilde{\mathcal{N}}^{(1)} \\ \vdots \\ \tilde{\mathcal{N}}^{(K)} \end{pmatrix}^T \begin{pmatrix} w_1^2 C^{(1)} & & 0 \\ & \ddots & \\ 0 & & w_K^2 C^{(K)} \end{pmatrix} \begin{pmatrix} \tilde{\mathcal{N}}^{(1)} \\ \vdots \\ \tilde{\mathcal{N}}^{(K)} \end{pmatrix} \\ &= \tilde{\mathcal{N}}^T C \tilde{\mathcal{N}} \end{aligned} \quad (17)$$

This matrix is a projection of the singular matrix C on a lower-dimensional space. Note that the matrix $\tilde{\mathcal{N}}^T$ is a nonsquare full-rank matrix by construction. Therefore, a unique solution $\tilde{s}_{(\text{opt})}$ will only be available if the intersection of the range of $\tilde{\mathcal{N}}^T$ and the null space of C is empty. Since

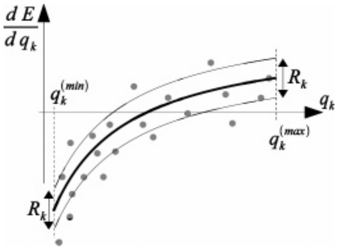

Figure 1. Schematic overview of how the weight factor w_k can be identified with the physically relevant interval of q_k. The fit of the auxiliary functions to the data set D_k is plotted, together with the RMS error on the fitted curve. The error on the integrated curve is approximated by w_kR_k.

$$\mathcal{N} \in \mathbb{R}^{KM\times[K-(3N-6)]M} \quad (18)$$

with N = the number of atoms and $K > 3N - 6$, one should expect $\mathcal{H}$ to be singular when $K \gg 3N - 6$, because then $\tilde{\mathcal{N}}$ is almost a square matrix. As stated in the introduction, an accurate force field always uses many more internal coordinates than independent coordinates. Consequently, for practical applications, a unique solution $\tilde{s}_{(\text{opt})}$ will not be available, no matter how much training data are used. This is a reformulation of the parameter correlations that occur when conventional least-squares fitting is used to parametrize force-field models.

The degeneracy of the cost function gives us the opportunity to select a solution $\tilde{s}_{(\text{opt})}$ that both minimizes Z and that will also result in a physically intuitive model. In this work, the physically intuitive character of a data set will be measured by a least-norm criterion: $\sum w_k^2||\tilde{y}^{(k)}||^2$. The lower this value, the smaller the forces along the internal coordinates in the resulting force-field model, and the more plausible the model. In general, $\mathcal{H}$ is much too large to store in any reasonable computer memory. It is therefore not feasible to perform a singular value decomposition of $\mathcal{H}$ in order to find the least-norm solution in S. Instead, a standard modification to the matrices C^k assures that Z has a unique solution that approximates the least-norm solution:

$$\begin{aligned} Z^*(\tilde{s}) &= Z(\tilde{s}) + \epsilon \sum_k w_k^2 (\tilde{y}^{(k)})^T \tilde{y}^k \\ &= \sum_k (\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s})^T w_k^2 (C^{(k)} + \epsilon I)(\tilde{p}^{(k)} + \tilde{\mathcal{N}}^{(k)}\tilde{s}) \end{aligned} \quad (19)$$

where ϵ is a positive constant that is small compared to one. This approximation (of the least-norm solution) becomes exact in the limit of ϵ toward zero, but for numerical applications, the optimal value of ϵ depends on the floating point accuracy. The minimization of Z^* can now be accomplished by a conjugate gradient method and a sparse notation for all the matrices in expression 19.

For reasons of transparency, no restrictions on the functional dependencies of the different internal coordinates have been imposed in the above derivation, and we only considered geometries of a single molecule. When creating realistic force fields, the method is complicated by two practical aspects. First, a useful force field should describe the energy

Table 1. Overview of the a Priori Information Used by the Force-Field Models[a]

benchmark model	sets of equivalent internal coordinates	number of elements
Water_default	OH bond lengths	2
	HOH bending angles	1
	HOH span	1
Water_ext1	in addition to the internal coordinates of Water_default	
	(HOH bending cosine) × (OH bond lengths)	2
	(HOH span) × (OH bond lengths)	2
Water_ext2	in addition to the internal coordinates of Water_ext1	
	(OH1 bond length) × (OH2 bond length)	1
Ammonia_default	NH bond lengths	3
	HNH bending angles	3
	HNH spans	3
Ammonia_ext1	in addition to the internal coordinates of Ammonia_default	
	N(HHH) distance	1
	(N(HHH) distance) × (NH bond lengths)	3
Ammonia_ext2	in addition to the internal coordinates of Ammonia_ext1	
	(HNH bending cosines) × (NH bond lengths)	6
	(HNH spans) × (NH bond lengths)	6
Methane_default	CH bond lengths	4
	HCH bending angles	6
	HCH spans	6
Methane_ext1	in addition to the internal coordinates of Methane_default	
	(HCH bending cosines) × (CH bond lengths)	12
	(HCH spans) × (CH bond lengths)	12
Methane_ext2	in addition to the internal coordinates of Methane_ext1	
	(CH bond lengths) × (CH bond lengths)	6

[a] All internal coordinates that belong to the same set are modeled with the same function $E_k(q_k)$ (see eq 1).

dependence of equivalent internal coordinates with the same expression E_k. Second, for a good parametrization, one would sample geometries of different molecules. Because both extensions merely introduce more indexes in the derivation, the same method applies.

3. Benchmark Protocol

The comparison of our novel procedure with conventional force-field parametrizations follows a strict protocol that will be applied on three small benchmark molecules: H_2O, NH_3, and CH_4. The protocol consists of six steps: (i) the generation of training data by a sampling procedure that performs ab initio calculations on a set of different geometries of the given molecule, in addition to the generation of test data by a similar sampling procedure that covers a larger part of the potential energy surface, (ii) the selection of the internal coordinates that are used in the force-field model and the sets of equivalent internal coordinates q_k that are modeled with the same functional dependence, (iii) the gradient curves method presented in this paper, (iv) conventional force-field constructions, using the analytical expressions generated in the former step as input, and (v) the individual validation of each force-field model based on training and test data, and the comparison of all the force-field models.

3.1. Sampling Procedure. The sampling procedure starts with a geometry optimization of the given molecule. The optimized geometry is chosen as the origin of an equidistant $(3N - 6)$-dimensional grid. The training set is then extended iteratively, by selecting the neighboring grid point of the already calculated geometries that has the lowest estimated ab initio energy. For each benchmark molecule, 200 training samples and 200 test samples have been generated. The samples in the training data set span an energy range from 0 to 60 kJ mol^{-1} with respect to the optimized geometry (the origin), while the test samples have a higher upper limit of 100 kJ mol^{-1}. This sampling procedure is only appropriate for small molecules. For larger systems, Monte Carlo sampling should be used. Since our main aim is to test the gradient curves method (while the resulting parameters are of minor importance), a rather low level of theory (DFT/B3LYP) and a small basis set (3-21G*) were used. All ab initio calculations were performed with the MPQC program.[33]

3.2. Selection of Internal Coordinates. When developing a force field, one has to select sets of equivalent internal coordinates on which the force-field energy depends. For the gradient curves method, this is the only information that must be given in advance. In this work, nine benchmark force fields are extensively studied, using the different choices of coordinates described in Table 1. The default models for the three molecules use all the interatomic distances and all the cosines of the bending angles, as illustrated in Figure 2. These internal coordinates correspond to those in the well-known Urey−Bradley-type force field, but in this work, no quadratic functional dependencies are imposed. Additionally, two extended force fields are studied for each molecule. The products of internal coordinates in the extended models only contain products of different internal coordinates, and it is always assured that only products of related internal coordinates are considered; for example, a product of two bond

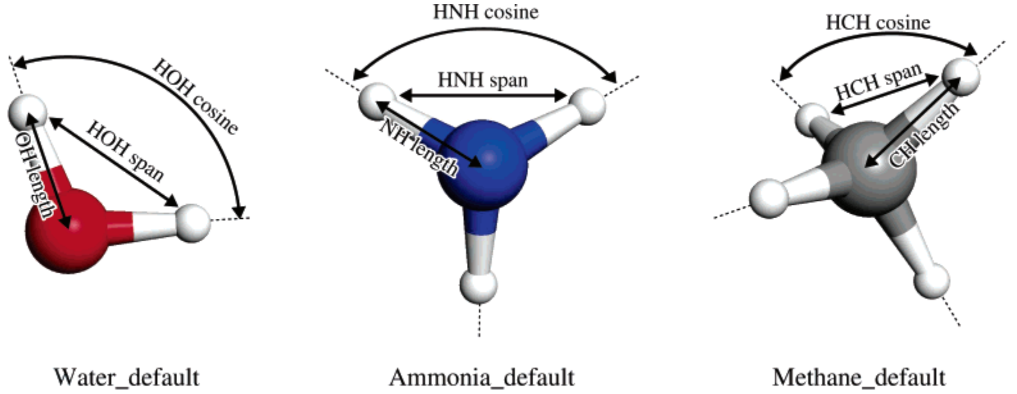

Figure 2. Schematic representation of the internal coordinates in the default models.

lengths will only be considered if the two bonds share exactly one atom. A detailed listing of which products have been used is given in the first section of the Supporting Information. Notice that the term “XYZ span” is defined as “the distance between the atoms X and Z that are both connected to the same atom Y”, and the “A(BCD) distance” is defined as “the distance between an atom A and the plane that is defined by the atoms B, C, and D”. The “XYZ span” is an internal coordinate initially introduced by Urey and Bradley[34] in their attempt to derive force fields for small molecules that show an improved reproduction of experimental vibrational frequencies. In their work, it is assumed that the corresponding energy term should be repulsive. We do not make this assumption a priori.

3.3. The Gradient Curves Method. For the auxiliary set of functions $f_n(q_k)$ in eq 8, polynomials up to the 11th order have been used. Two variants of the new gradient curves method are applied: **GCI** is the ill-conditioned variant of the new method, that is, without the least-norm criterion; **GCL** is the variant in which the least-norm correction is applied with $\epsilon = 10^{-6}$.

3.4. The Conventional Methods. In addition to the gradient curves method presented in this work, a series of conventional force-field parametrizations has been performed. They are conventional in the sense that the parameters have been obtained by directly minimizing a well-defined least-squares cost function, although in the literature, additional techniques are used to deal with parameter correlations. The different types of cost functions are listed below. Optionally, a constraint has been applied that compels the force field to reproduce the ab initio Hessian and the zero gradient for the ab initio optimized geometry.

CEU is an unconstrained minimization of the residual error on the energies[35]

$$Z_{\mathrm{CEU}} = \sum_{m=1}^{M} [(E_{\mathrm{AI}}^{(m)} - E_{\mathrm{AI}}^{(\mathrm{opt})}) - (E_{\mathrm{FF}}^{(m)} - E_{\mathrm{FF}}^{(\mathrm{opt})})]^2 \qquad (20)$$

where the sum over m contains all the molecules in the training set and corrections due to the difference in reference energies of the ab initio and the force-field model have been taken into account.

CEC is a minimization of the residual error on the energies constrained so that the ab initio Hessian and zero gradient are reproduced at the ab initio equilibrium geometry: $Z_{\mathrm{CEC}} = Z_{\mathrm{CEU.}}$

CGU is an unconstrained minimization of the residual error on the gradients[36]

$$Z_{\mathrm{CGU}} = \sum_{m=1}^{M} \sum_{i=1}^{3N} \left(\frac{\partial E_{\mathrm{AI}}^{(m)}}{\partial x_i} - \frac{\partial E_{\mathrm{FF}}^{(m)}}{\partial x_i} \right)^2 \qquad (21)$$

where i iterates over the Cartesian coordinates.

CGC is a minimization of the residual error on the gradients constrained such that the ab initio Hessian and zero gradient are reproduced at the ab initio equilibrium geometry: $Z_{\mathrm{CGC}} = Z_{\mathrm{CGU.}}$

CCU is an unconstrained minimization of the residual error on the energies and gradients of all the training geometries, as well as the Hessian of the optimized molecule where (i,j) iterates over all the pairs of the Cartesian

$$Z_{\mathrm{CCU}} = W_{\mathrm{CEU}} Z_{\mathrm{CEU}} + W_{\mathrm{CGU}} Z_{\mathrm{CGU}} + W_{\mathrm{CHU}} \sum_{i=1}^{3N} \sum_{j=i}^{3N} \left(\frac{\partial^2 E_{\mathrm{FF}}^{(\mathrm{opt})}}{\partial x_i \partial x_j} - \frac{\partial^2 E_{\mathrm{AI}}^{(\mathrm{opt})}}{\partial x_i \partial x_j} \right)^2 \qquad (22)$$

coordinates. The three contributions to the cost function have been weighted to ensure that they have a proportional influence on the obtained parameters. Alternative cost functions that combine ab initio energies, gradients, and/or Hessians have also been reported in the literature for the optimization of force-field parameters.[1,10,11]

The conventional parametrizations will serve as a reference for the results of the gradient curves method. To guarantee a fair comparison, the analytical expressions used in the conventional methods where obtained with GCL and these expressions only contain linear parameters.

3.5. Validation and Comparison. The generated force-field models are validated with three different criteria. (i) The standard deviation on $E_{\mathrm{FF}} - E_{\mathrm{AI}}$ for all geometries, defined as $\langle [(E_{\mathrm{FF}} - E_{\mathrm{AI}}) - \langle E_{\mathrm{FF}} - E_{\mathrm{AI}} \rangle]^2 \rangle^{(1/2)}$, should be small. The standard deviation is not sensitive to the reference energies of both ab initio and force-field models, in contrast to the root mean square of $E_{\mathrm{FF}} - E_{\mathrm{AI}}$, given by $\langle (E_{\mathrm{FF}} - E_{\mathrm{AI}})^2 \rangle^{(1/2)}$. (ii) The root mean square of $|\nabla E_{\mathrm{FF}} - \nabla E_{\mathrm{AI}}|$ should be small where ∇ indicates the Cartesian gradient. (iii) At the ab initio optimized geometry, the ratios of the eigenvalues for matching eigenvectors of the force field and of the ab initio Hessian should be near unity.

The third quality criterion is calculated as follows. First, the ab initio Hessian and the force-field Hessian are calculated at the ab initio optimized geometry. The eigenmodes corresponding to the external degrees of freedom are removed by projecting both Hessians on the same basis of $3N - 6$ independent internal coordinates. Then, both projected matrices are diagonalized. The overlap matrix of the corresponding sets of eigenvectors shows clearly which two eigenvectors of the ab initio Hessian and the force-field Hessian correspond with each other. Significant mismatches have not been observed. Finally, the ratios of the eigenvalues associated with the corresponding eigenvectors are calculated.

The quality of the force fields will be compared by the three criteria defined above. In order to assess the robustness of the parametrization, validations i and ii are in addition applied to the set of test data. Finally, we have examined the possibility of giving a physical interpretation to the force-field expressions E_k obtained in the different models.

4. Results and Discussions

To illustrate the usage of the new procedure, we first discuss the three gradient curves generated by GCL applied to Water_default. For each geometry m, the Jacobian, $J^{(m)}$ (see eq 3) is a $N \times K$ matrix or 9 × 4 matrix of rank $3N - 6 = 3$. The matrix $\mathcal{N}^{(m)}$ describing the null space of such a Jacobian has the dimension $N \times K - (3N - 6)$ or 9 × 1. Consequently, given the 200 geometries in the training set, 200 unknown coefficients must be obtained by minimizing the cost function, Z^*. Although there are four distinct internal coordinates in this specific force-field model, the two transformed data sets corresponding to the OH-bond length have been merged into one; that is, their continuity is measured as a whole. Consequently, the data set associated with the bond length consists of 400 data points, while the two others contain 200 data points each. The continuity of each data set is measured by the goodness of fit of an auxiliary 11th-order polynomial. We used generic high-order polynomials to prevent any assumptions about the resulting energy terms being imposed by the continuity criterion; that is, these polynomials will not enforce specific features in the final energy terms. The results are depicted in Figure 3. The data sets D_k obtained by substituting $s_{(\mathrm{opt})}$ into eq 7 are plotted as black crosses. The minimization of Z^* guarantees that these data points lie on continuous curves. The (optimized) auxiliary polynomials that are used to measure the continuity are plotted as dashed lines. Their unphysical asymptotic behavior and the oscillations at the boundaries clarify that the auxiliary polynomials can only be regarded as a measure for the continuity and that they cannot be used as functional forms for the force-field model. In a next step, the analytical form of the derivative of E_k is estimated, on the basis of the data sets. For the energy curve of the OH stretch, a sixth-order polynomial in $1/r_{\mathrm{OH}}$ gives an accurate fit, and the resulting expression has the expected asymptotic behavior. The energy curves of the cosine of the bending angle and the interatomic HH distance are estimated to be quadratic and cubic, respectively. Finally, the parameters in the functional forms are optimized using one-dimensional least-squares optimization to the data sets D_k. The resulting curves, dE_k/dq_k, are plotted as solid lines in Figure 3, and the optimized parameters are given in Table 2. The GCL parameters for all nine benchmark models are included in the second section of the Supporting Information.

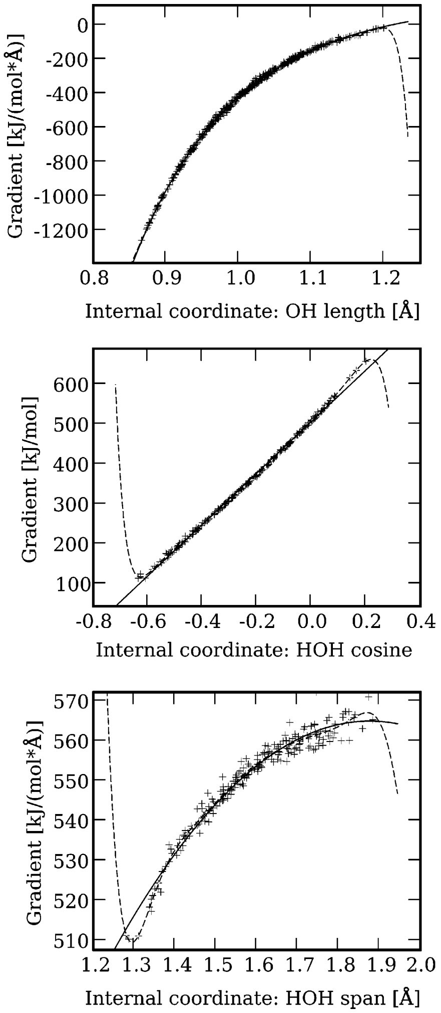

Figure 3. Gradient curves dE_k/dq_k (solid line) obtained for the Water_default model with the GCL method. The black crosses represent the transformed one-dimensional data (see text). The dashed curves are the fitted auxiliary functions for evaluating the continuity criterion.

An overview of the quality criteria for each parametrization is given in Figure 4. The x axis shows the force-field models, and for each force-field model, the different parametrization methods (GCI, GCL, CEU, and so forth) are indicated with different colors. On the y axes, the quality criteria are plotted on a logarithmic scale. Figure 4a and b display respectively the standard deviation on $(E_{\mathrm{FF}} - E_{\mathrm{AI}})$ and the root mean square of $|\nabla E_{\mathrm{FF}} - \nabla E_{\mathrm{AI}}|$ for both training geometries (filled circles) and test geometries (open circles). Figure 4c gives an overview of the validation with the third criterion, represented by the ratios of corresponding Hessian eigenvalues (force-field over ab initio estimates) at the ab initio optimized geometry. It is clear that the overall quality of

Table 2. The Parameters for the Water Default Model Obtained with GCL[a]

OH bond length r		HOH bending cosine c		HOH span d	
terms	coefficients	terms	coefficients	terms	coefficients
r^{-1}	−4.608e−01	c	1.931e−01	d	9.898e−03
r^{-2}	5.210e−02	c^2	1.228e−01	d^2	2.933e−02
r^{-3}	3.578e−01			d^3	−2.758e−03
r^{-4}	3.988e−01				
r^{-5}	3.103e−01				
r^{-6}	1.943e−01				

[a] The functional form of each energy term, $E_k(q_k) = \sum_{t=1}^{T_k} c_t \mathcal{F}_t(q_k)$, is a linear combination of terms listed in the first column of each table. The corresponding coefficients in this linear combination are given in the second column. All parameters are given in atomic units.

the force fields constructed with GCL is comparable to that obtained by the conventional methods. Nevertheless, some interesting discrepancies appear, which will be discussed below.

The ammonia molecule serves as a good example of how to obtain relevant sets of internal coordinates. Initially, the new method was applied on the Ammonia_default model, which only contains the basic internal coordinates: bond lengths, interatomic distances, and bending angles. As shown in Figure 4c, the constructed force field predicts one eigenvalue of the Hessian that deviates significantly from the ab initio value. This eigenvalue corresponds to the inversion of the ammonia molecule. At the transition state of this umbrella inversion, the NH bond length increases due to the alteration from sp^3 to sp^2 hybridization. To describe the inversion more accurately, the extended ammonia model contains two extra sets of internal coordinates: the out-of-plane distance and the products of the out-of-plane distance with the bond lengths. It is striking to observe that the parametrization of the extended ammonia model results in a seriously improved reproduction of the eigenvalues. An attempt was made to avoid the inclusion of more internal coordinates, by constraining the parameters in order to reproduce the ab initio Hessian. This failed drastically for ammonia and methane, since these constraints led to unacceptable errors on the energies and gradients for both training and test data. The corresponding quality criteria falls out of the scope of Figure 4a and b. The performance of CCU in the parametrization of the Ammonia_default model manifestly suffers from the attempt to use information of the ab initio Hessian in the optimization.

The parametrization of ammonia demonstrates that, in some cases, the inclusion of additional redundant internal coordinates in a force-field model is indispensable. This is in agreement with previous studies where it was shown that a pure Urey−Bradley force field, that is, the default model in this work, is not sufficient for an adequate description of the ammonia molecule.[37,38] Unfortunately, the parametrization of a force field with a high number of internal coordinates ($K \gg 3N - 6$) is sensitive to parameter correlations, and a good treatment of these correlations is required to obtain a useful force field.

In the remainder of this section, we discuss the effect of increasing model complexity. The main effect of the extensions to the force-field models is visible in Figure 4. An improved reproduction of the energies, gradients, and the Hessian is obtained for all methods except the GCI method. This general trend is understandable: the more parameters a model contains, the further a cost function can be optimized. The poor performance of the GCI method for the extended models needs some explanation. Both GCI and GCL yield the same transformed data sets D_k for the default models. For these models, the cost function Z (see eq 16) has a unique solution, even without applying the least-norm correction. This is no longer true for the extended models. In these cases, the minimum of the cost function, Z, becomes highly degenerate, and GCI selects from this minimum an essentially random solution, in the sense that a small change in the training data would imply a very large change in the transformed data sets D_k. On average, such a random solution consists of transformed data sets D_k with very high ranges. This is unacceptable because the absolute errors from fitting energy terms to the transformed data sets (step II of the gradient curves method) scale with the range of D_k. Consequently, the absolute errors shown in Figure 4 are much higher for GCI when applied to the most extended models. We conclude that, of the new methods, GCL is to be preferred over GCI. Both are equivalent for a small number of internal coordinates, but GCL produces superior fits for the more extended models.

The most important trend noticed by increasing the complexity of the model is the behavior of the functions E_k, which is different for GCL as compared to all other methods (i.e., the conventional methods and GCI). Figures 5−7 display all the energy terms E_k, obtained with CCU and GCL, for the water, ammonia, and methane molecules, respectively. In these figures, CCU could have been replaced by any other method except GCL without generating significant differences in the global trends. Each row in these figures contains the plots of the energy terms that belong to a specific force-field parametrization, while every column corresponds to a specific set of equivalent internal coordinates. In what follows, we will first discuss the global trends in these figures, and consequently some more specific aspects will be discussed that are not applicable to all the results.

Figures 5a, 6a, and 7a show that CCU yields energy terms E_k with increasing amplitudes, when the force-field model is extended with extra internal coordinates. The conventional methods use the extra degrees of freedom to improve the accuracy, but this improvement is the result of a nonrobust cancellation of high-energy contributions. We have tested an implementation of the conventional methods that applies a singular value decomposition to the design matrix,[32,39] but a singular value cutoff that gives a good balance between accuracy and reasonable behavior of the functions E_k is not available. The reason is that a least-norm solution in the parameter space is not meaningful since the parameters have different units. A weighted least-norm solution, where the norm of dimensionless weighted parameters is minimal, would be more correct, but then one has to determine a weight value for each parameter as in the work of Ewig et al.[11] It is highly remarkable that, as depicted in Figures 5b,

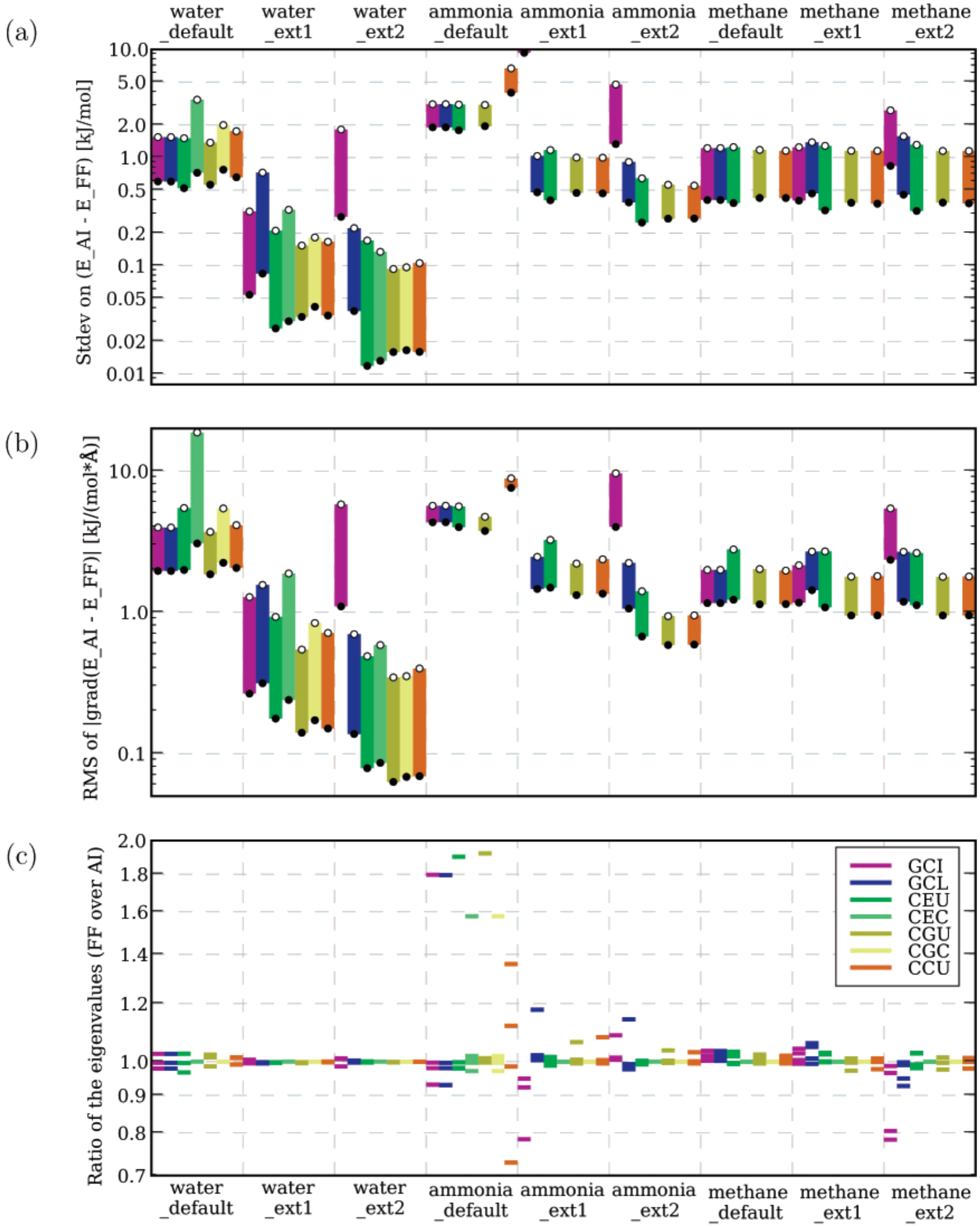

Figure 4. Overview of the force-field validations. Upper figure (a): Standard deviation of the energy differences. Middle figure (b): Root mean square of the gradient differences. Lower figure (c): Ratios of corresponding Hessian eigenvalues (force field over ab initio values), at the ab initio optimized geometry (see text). In parts a and b, the errors for the constrained methods applied on ammonia and methane are too large to fit in the scale of both plots.

6b, and 7b, GCL shows exactly the opposite trend from CCU: the ranges of the functions E_k are reduced in the extended force-field models, and for the ext2 models, it is even possible to give a physical interpretation to the important energy dependencies. For example, the minima of E_k correspond approximately to the internal coordinates of the ab initio optimized geometry. For the terms E_{OH}, E_{NH}, and E_{CH}, even a Morse-like behavior (i.e., the left side of the curve is steeper than the right side) is reproduced. It should be remarked that GCL does not depend on constraints, model selection, or ad hoc interventions to obtain physical force-field terms. When the gradient curves method will be applied on larger systems, we expect that the absence of cancellation effects will yield transferable and accurate force fields.

In addition to the global trends discussed above, some interesting specific features show up in the results. The most remarkable outcome is that the energy terms for the Ammonia_default model obtained with CCU are very reasonable, and at first instance, this appears to contradict the previous

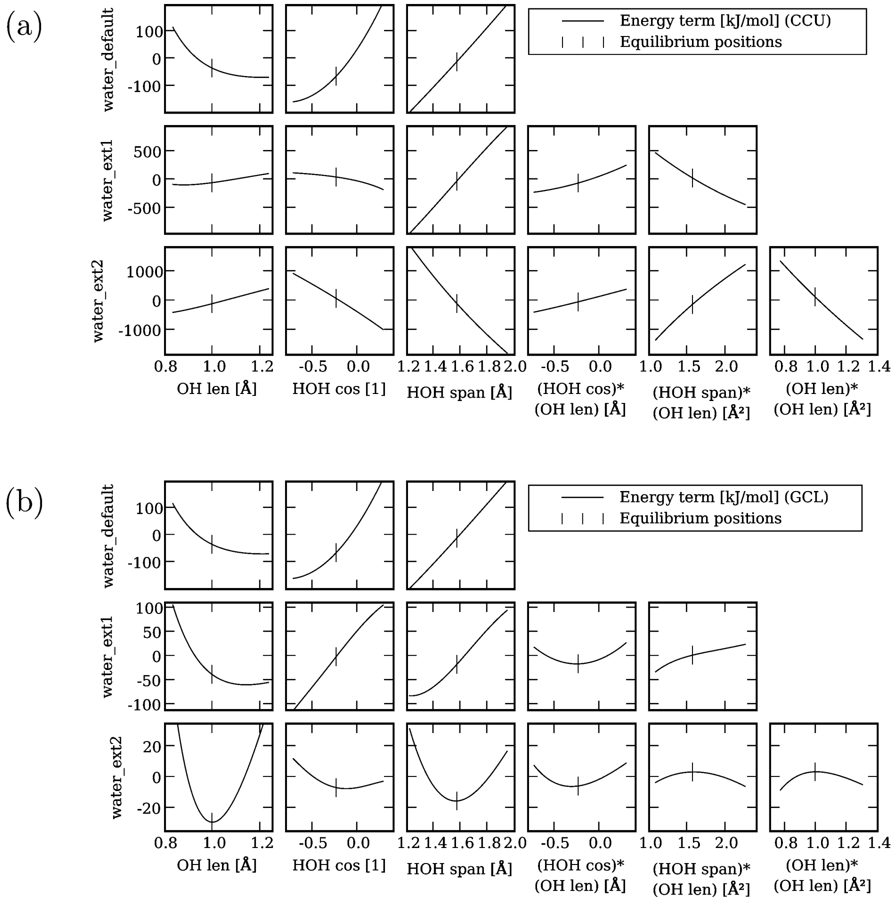

Figure 5. Energy terms E_k for the three different water models, generated (a) by CCU, a conventional parametrization method, and (b) by GCL, the gradient curves method with the least-norm correction. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

paragraph where we stated that reasonable models could only be obtained with GCL. The explanation is that the Ammonia_default model with CCU parameters is indeed reasonable but less accurate compared to other parametrizations of Ammonia_default (see Figure 4a and b). The energy terms for the Ammonia_default model obtained with CGU (see Figure 6c) reveal that the incorporation of the ab initio Hessian of the optimized ammonia geometry in the CCU cost function forces the energy terms of the Ammonia_default model to behave reasonably.

A more subtle result is that the first row of Figure 5a contains virtually the same energy terms as the first row in Figure 5b. Similarly, the first row of both parts a and b of Figure 7 are virtually equal. This situation can be summarized as follows: CCU, a method that does not handle parameter correlations, yields the same energy terms as GCL, a method that does treat parameter correlations. The reason is that none of the parametrization methods in this paper suffer from parameter correlation problems in case of the default models. In the case of the Water_default or the Methane_default model, all the uniquely defined minima of the cost functions of CCU, CEU, CGU, GCL, and GCI even result in the same energy terms. As already discussed above, the different cost functions in the case of Ammonia_default have different—but each of them uniquely defined—optimal parameters. The absence of parameter correlations does however not imply reasonable energy terms. Actually, the sets of equivalent internal coordinates in the default models are too limited for an accurate reproduction of all the training data with reasonable energy terms. The OH-stretch term represents a repulsive interaction, whereas the energy terms for the HH distance and HOH cosine are both attractive interactions. Correct behavior is obtained only when the three energy terms are combined. For reasons of clarity, we note that the GCL curves in the default models are not supposed to coincide perfectly with the quadratic energy terms in a standard Urey—Bradley parametrization, which are fitted so as to reproduce experimental frequencies.[40−42] In the present case, the curves are fitted not only to molecular configurations near equilibrium but to higher-energy configurations as well. In fact, when the curves in the first row of Figures 5b and 7b are quadratically expanded around the equilibrium values, a fair correlation with the quadratic force constants and the minima in the work of Kuchitsu and Bartell[40,41] is observed.

At this point, we have shown how the gradient curves method is able to reconcile the accuracy and the physical interpretation of a force-field model. However, one could wonder how the energy terms, as shown in Figures 5b, 6b, and 7b, evolve when the force-field model is extended with even more additional sets of equivalent internal coordinates (higher-order products, cubic terms, etc.). In the HDMR approach,[25] orthogonality criteria are introduced to assert that the addition of higher-order terms does not have any

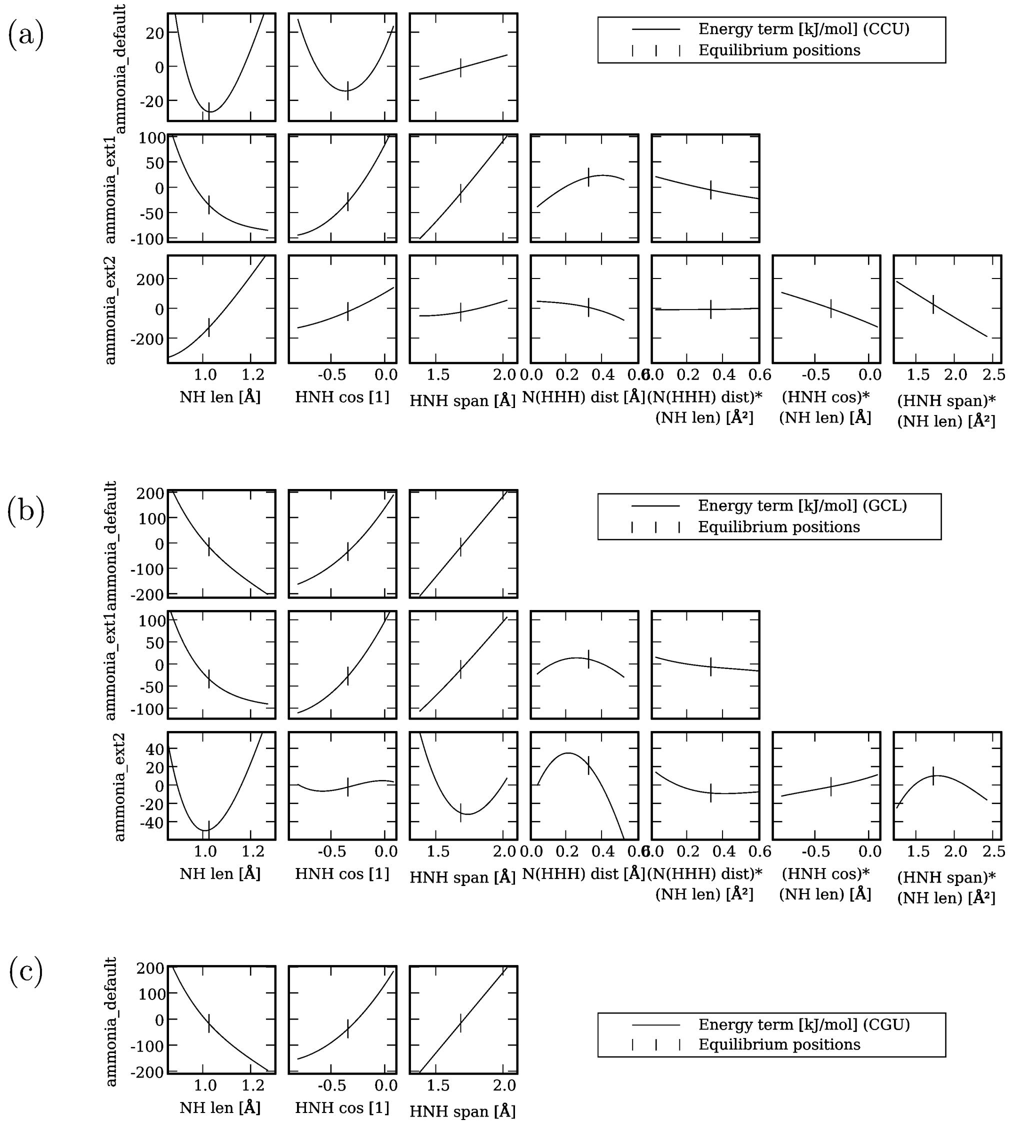

Figure 6. Energy terms E_k for the three different ammonia models, generated (a) by CCU, a conventional parametrization method, (b) by GCL, the gradient curves method with the least-norm correction, and (c) by CGU, a conventional parametrization method that only uses ab initio gradient training data. For part c, only the default model is shown. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

influence on the lower-order terms in the model. The gradient curves method never relies on such orthogonality criteria; for example, this is the reason why the energy terms for the bond length of the three models in Figure 5b are different. There is no "theoretical guarantee" that modifications will not occur when the water model is extended with even more sets of equivalent internal coordinates. Additional energy terms make the continuity criterion extremely degenerate, and in such cases, the least-norm criterion might become an overly naive representation of our physical intuition. Figure 8 demonstrates the behavior of the energy terms for a series of additional extended water models. Similar plots for ammonia and methane are included in the third section of the Supporting Information. Except for the highest-order terms in the two most extended models for the water molecule, the modifications in the energy terms seem to converge once the model is extended enough to show a physically intuitive behavior. The inclusion of second-order derivatives of the ab initio energy in the training data and more sophisticated criteria for our physical intuition are viable candidates to cure the situation for the two most extended water models and are the subject of our current active research. Nevertheless, one should realize that also these additional measures would suffer from the same defects for the very hypothetical case of even more extended models.

5. Conclusions

This work shows how the gradient curves method can surmount several difficulties that are associated with the development of force fields using least-squares parametrization. Technically, the new method is a two-step procedure: in the first step, continuity criteria and subordinate least-norm criteria are imposed to transform the multidimensional training data into a series of separate one-dimensional data sets, each associated with an energy term of the proposed force field. In this work, the training data are the gradients of the ab initio energy for different molecular geometries.

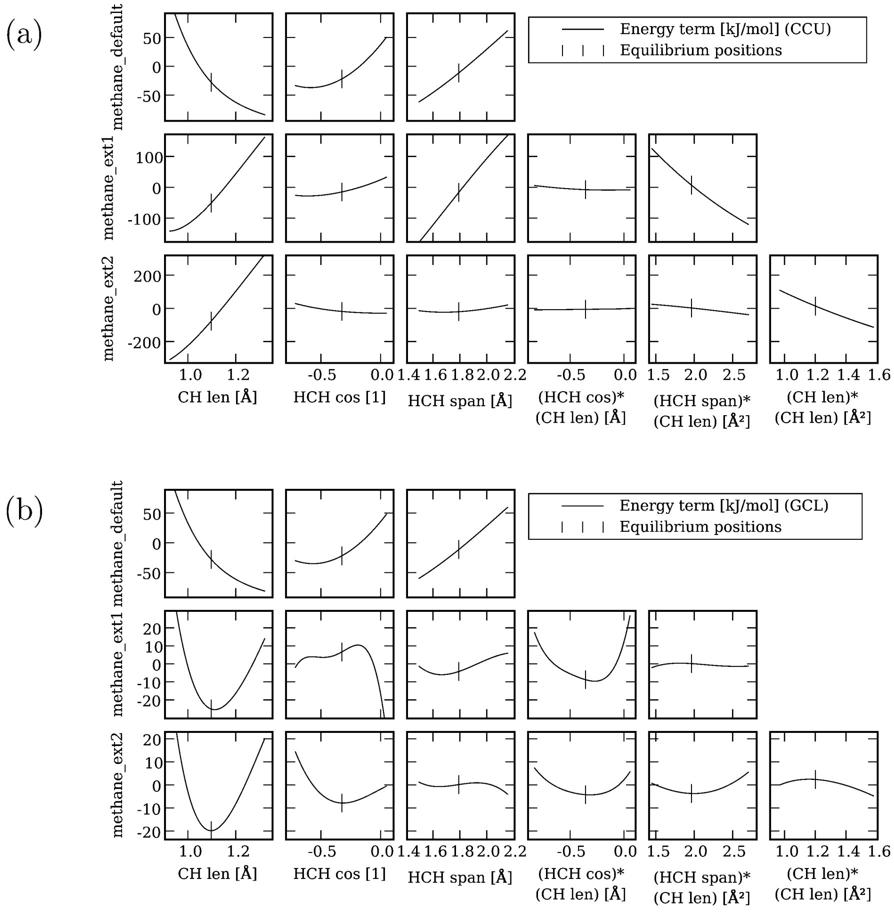

Figure 7. Energy terms E_k for the three different methane models, generated (a) by CCU, a conventional parametrization method, and (b) by GCL, the gradient curves method with the least-norm correction. The values of the internal coordinates at the ab initio equilibrium geometry are marked by vertical lines.

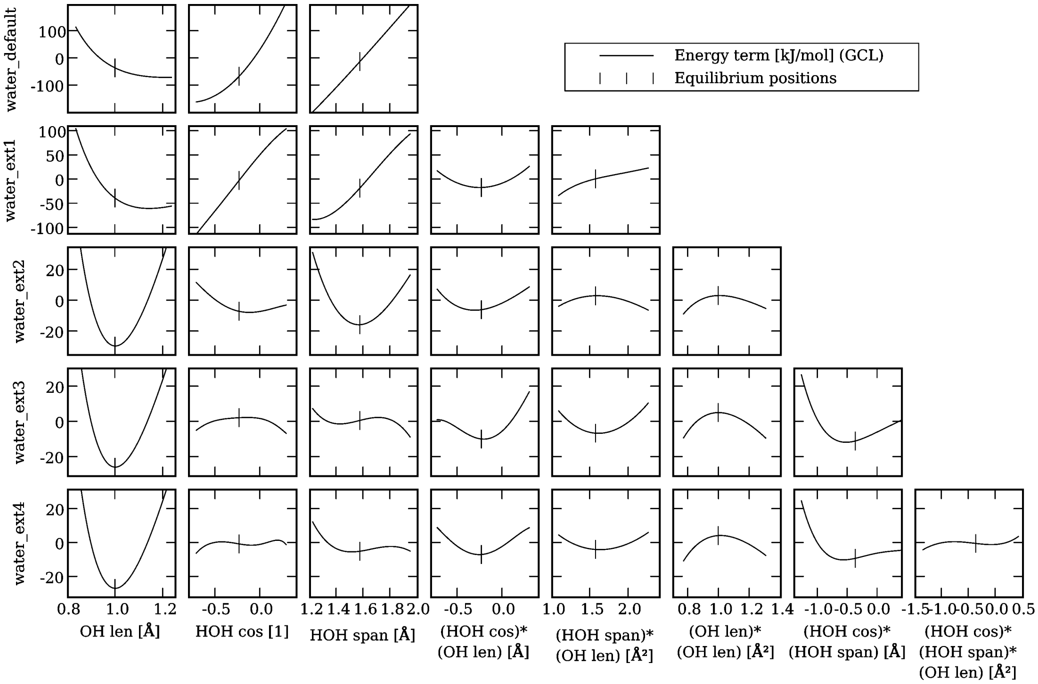

Figure 8. Overview of the energy terms for additional extended water models parametrized with GCL.

During the second step, the derivative of each energy term in the force field is fitted to the corresponding transformed data set.

The gradient curves method has several advantages. Only the internal coordinates have to be defined in advance, instead of a complete analytical ansatz of the force-field model. The problem of parameter correlations that troubles the conventional force-field development is tackled during the transformation from the multidimensional training data to separate one-dimensional data sets. The continuity and least-norm criteria that are imposed do not only guarantee that the transformed data sets are unique but they also facilitate the physical interpretation of the energy terms fitted to these data sets. In fact, the least-norm criteria express the argument that a plausible force-field model should not contain large derivatives in the energy terms to acquire a marginal increase of accuracy. This prescription fixes all the parameter correlations that originate from the redundancy of the internal coordinates in the force-field model. Once the first step is completed, suitable analytical expressions for the energy terms can be easily proposed after analysis of the transformed data sets and taking into account the expected asymptotic behavior of these energy terms. Because the ability of interpreting the individual force-field terms is known to be a prerequisite for transferable force fields,[2,11] we expect this method to be very helpful when developing accurate and robust force-field models for larger systems.

The current research mainly focuses on an extended variation of the gradient curves method which is also capable of efficiently deriving the nonbonding interactions from ab initio training data. The primary application on a large system will be the construction of an accurate all-atom zeolite-guest force field. Other active areas include the extension of the gradient curves method to include the ab initio energy and Hessian in the training data, and a more sophisticated formalism for the intuitive character of the energy terms that will eventually supersede the least-norm criterion. We also expect a generalization of the gradient curves method (beyond the scope of force fields) to be useful whenever data parametrization is complicated by parameter correlations and the absence of theoretically supported analytical models.

Acknowledgment. T.V. would like to thank the Flemish organization IWT for its financial support. P.W.A. would like to thank NSERC for funding. V.V.S and M.W. thank the Fund for Scientific Research—Flanders and the Research Board of Ghent University.

Supporting Information Available: A listing of the internal coordinates, the GCL parameters, and an overview of additional extended models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Schröder, K. P.; Sauer, J. *J. Phys. Chem.* **1996**, *100*, 11043−11049.

(2) Hill, J.; Sauer, J. *J. Phys. Chem.* **1995**, *99*, 9536−9550.

(3) Sierka, M.; Sauer, J. *Faraday Discuss.* **1997**, *106*, 41−62.

(4) Smirnov, K. S.; Bougeard, D. *Chem. Phys.* **2003**, *292*, 53−70.

(5) Ermoshin, V. A.; Engel, V. *J. Phys. Chem. A* **1999**, *103*, 5116−5122.

(6) Chandross, M.; Webb, E. B.; Grest, G. S.; Martin, M. G.; Thompson, A. P.; Roth, M. W. *J. Phys. Chem. B* **2001**, *105*, 5700−5712.

(7) Pascual, P.; Ungerer, P.; Tavitian, B.; Pernot, P.; Boutin, A. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3684−3693.

(8) Allinger, N. L.; Chen, K.; Lii, J.-H. *J. Comput. Chem.* **1996**, *17*, 642−668.

(9) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490−519.

(10) Sun, H.; Rigby, D. *Spectrochim. Acta, Part A* **1997**, *53*, 1301−1323.

(11) Ewig, C.; Berry, R.; Dinur, U.; Hill, J.; Hwang, M.; Li, H.; Liang, C.; Maple, J.; Peng, Z.; Stockfisch, T.; Thacher, T.; Yan, L.; Ni, X.; Hagler, A. *J. Comput. Chem.* **2001**, *22*, 1782−1800.

(12) Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. *J. Phys. Chem.* **1993**, *97*, 10269−10280.

(13) Mayo, S.; Olafson, B.; Goddard, W. *J. Phys. Chem.* **1990**, *94*, 8897−8909.

(14) Rappe, A.; Casewit, C.; Colwell, K.; Goddard, W.; Skiff, W. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(15) Shi, S.; Yan, L.; Yang, Y.; Fisher-Shaulsky, J.; Thacher, T. *J. Comput. Chem.* **2003**, *24*, 1059−1076.

(16) Mortier, W.; Ghosh, S.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315−4320.

(17) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 6141−6156.

(18) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621−627.

(19) Chalasinski, G.; Szczesniak, M. *Chem. Rev.* **2000**, *100*, 4227−4252.

(20) Bordner, A. J.; Cavasotto, C. N.; Abagyan, R. A. *J. Phys. Chem. B* **2003**, *107*, 9601−9609.

(21) Giese, T.; York, D. *Int. J. Quantum Chem.* **2004**, *98*, 388−408.

(22) Maple, J.; Dinur, U.; Hagler, A. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5350−5354.

(23) Maple, J. R.; Hwang, M. J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *J. Comput. Chem.* **1994**, *15*, 162−182.

(24) Martinez, E.; Lopez, J. J.; Vazquez, J. *J. Mol. Struct.* **2004**, *705*, 141−145.

(25) Rabitz, H.; Aliş, O.; Shorter, J.; Shim, K. *Comput. Phys. Commun.* **1999**, *117*, 11−20.

(26) Manzhos, S.; Carrington, T. *J. Chem. Phys.* **2006**, *125*, 084109.

(27) Shorter, J. A.; Ip, P. C.; Rabitz, H. A. *J. Phys. Chem.* A **1999**, *103*, 7192−7198.

(28) Gorban, A. N. *Appl. Math. Lett.* **1998**, *11*, 45−49.

(29) Frisch, H. L.; Borzi, C.; Ord, G.; Percus, J. K.; Williams, G. *Phys. Rev. Lett.* **1989**, *63*, 927−929.

(30) Kolmogorov, A. *Dokl. Akad. Nauk SSSR* **1957**, *114*, 679.

(31) Hilbert, D. *Bull. Am. Math. Soc.* **1902**, *8*, 461.

(32) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Singular Value Decomposition. In *Numerical Recipes in C: The Art of Scientific Computing*; Cowles, L., Harvey, A., Hahn, R., Eds.; Press Syndicate of the University of Cambridge: Cambridge, United Kingdom, 2002; Chapter 2.6, pp 59−70.

(33) Janssen, C. L.; Nielsen, I. B.; Leininger, M. L.; Valeev, E. F.; Seidl, E. Y. *The Massively Parallel Quantum Chemistry Program (MPQC)*, version 2.3.0; Sandia National Laboratories: Livermore, CA, 2004.

(34) Urey, H. C.; Bradley, C. A. *Phys. Rev.* **1931**, *38*, 1969−1978.

(35) Kramer, G.; Farragher, N.; van Beest, B.; van Santen, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43*, 5068−5080.

(36) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1994**, *26*, 583−588.

(37) Pariseau, M.; Wu, E.; Overend, J. *J. Chem. Phys.* **1962**, *37*, 217−223.

(38) King, W. T. *J. Chem. Phys.* **1961**, *36*, 165−170.

(39) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. General Linear Least Squares. In *Numerical Recipes in C: The Art of Scientific Computing*; Cowles, L., Harvey, A., Hahn, R., Eds.; Press Syndicate of the University of Cambridge: Cambridge, United Kingdom, 2002; Chapter15.4, pp 671−681.

(40) Kuchitsu, K.; Bartell, L. S. *J. Chem. Phys.* **1962**, *36*, 2460−2469.

(41) Kuchitsu, K.; Bartell, L. S. *J. Chem. Phys.* **1962**, *36*, 2470−2481.

(42) Simanouthi, T. *J. Chem. Phys.* **1949**, *17*, 245−248.

CT6002093