The Choice of Internal Coordinates in Complex Chemical Systems

KÁROLY NÉMETH,[1,2] MATT CHALLACOMBE,[3] MICHEL VAN VEENENDAAL[1,2]
[1]*Department of Physics, Northern Illinois University, DeKalb, Illinois 60115*
[2]*Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439*
[3]*Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*

Received 27 July 2009; Revised 16 September 2009; Accepted 3 December 2009
DOI 10.1002/jcc.21494
Published online 19 January 2010 in Wiley InterScience (www.interscience.wiley.com).

Abstract: This article presents several considerations for the appropriate choice of internal coordinates in various complex chemical systems. The appropriate and black box recognition of internal coordinates is of fundamental importance for the extension of internal coordinate algorithms to all fields where previously Cartesian coordinates were the preferred means of geometry manipulations. Such fields range from local and global geometry optimizations to molecular dynamics as applied to a wide variety of chemical systems. We present a robust algorithm that is capable to quickly determine the appropriate choice of internal coordinates in a wide range of atomic arrangements.

© 2010 Wiley Periodicals, Inc. J Comput Chem 31: 2078–2086, 2010

Key words: geometry optimization; curvilinear internal coordinates

Introduction

Atomic positions inside a molecule can be described in terms of Cartesian coordinates, that is, for each atom, relative to a 3D coordinate system independently from other atomic positions, or alternatively they can also be defined relative to other atoms, such as in terms of atomic distances, valence angles, or dihedral angles. These relative coordinates are also referred to as "internal coordinates" due to the fact that they can describe only the internal degrees of freedom of a molecule and not the external ones, that is not those related to the location of the center of mass of the molecule and the orientation of the whole of the molecule in space. It has been recognized in vibrational spectroscopies[1] a long time ago that most molecular vibrations are fairly well localized on internal coordinates that reflect chemical concepts, such as chemical bonds, valence angles, or dihedral torsions. These latter coordinates are also called "primitive" internal coordinates as opposed, for example to internal coordinates built as linear combinations of primitive ones. In fact, vibrational coupling between appropriately chosen internal coordinates is typically an order of magnitude smaller than in the corresponding Cartesian representation.[2–5] This observation holds not only for harmonic but more importantly for anharmonic vibrational couplings. The recognition of this reduced vibrational coupling led to the development of internal coordinates-based geometry optimization,[2,3,5–7] which is now the standard means of local optimization in most quantum chemistry software packages. Internal coordinate geometry optimization reduces the number of optimization steps typically by a factor of 2–10 for small or medium sized molecules compared with Cartesian conjugate gradient algorithms.[8] Also molecular dynamics can greatly benefit from the use of internal coordinates in the efficient simulation of long time-scale motions of large molecules.[9,10] In this latter field, the advantage of internal coordinates stands in the fact that many units of the structure, for example, the covalent bonds or the valence angles between them remain essentially rigid during the motion, and one can treat these parts of the molecular structure as rigid bodies in internal coordinate representation, thus, allowing motions only along the slowly moving complex degrees of freedoms that are mostly composed as highly delocalized linear combinations of torsions along the covalent bonds. The acoustic phonon modes of solids are highly delocalized and can have arbitrarily small frequencies associated with extremely flat potential energy surfaces. The larger the linear extent, L, of the solid, the smaller the frequency of the acoustic phonons may become as it is proportional to $1/L$ through a linear dispersion relation.[11] Optimization along these normal modes is difficult even in the framework of internal coordinates. Efficient treatment of such difficult to optimize cases is based on the separation of long and short wavelength modes in the framework of elasticity theory[12] or on the construction of simple model Hessians based on internal coordinates and simple force constant estimates.[13] This latter study also points out that given a good enough initial guess to the Hessian, in many cases even optimization in Cartesian coordinates can be very efficient. This is mostly true if starting geometries

Correspondence to: K. Németh; e-mail: nemeth@anl.gov

Contract/grant sponsor: U.S DOE; contract/grant numbers: DE-FG02-03ER46097 and W-7405-ENG-36

are close to a local minimum, otherwise, for example, in the case of transition state search, internal coordinate optimization is superior to Cartesian ones even if the exact Hessian is available at every step of the optimization,[4,14,15] because of the far more efficient treatment of anharmonic couplings in internal coordinate representation. The application of internal coordinate optimization to solids also proved to be very efficient.[8,16–18]

Every set of atomic Cartesian coordinates (molecular geometry), $\boldsymbol{X}$, can be associated with actual values of the internal coordinates, $\boldsymbol{\Phi}$, defined to describe the relative atomic positions at $\boldsymbol{X}$ and with the orientation of the whole of the molecule in 3D. Displacements of internal and Cartesian coordinates can be related to each other through Taylor series expansion:

$$\delta\phi_i = \sum_j \frac{\partial\phi_i}{\partial x_j}\delta x_j + \frac{1}{2}\sum_{j,k}\frac{\partial^2\phi_i}{\partial x_j \partial x_k}\delta x_j \delta x_k + \cdots, \tag{1}$$

where ϕ_i is the i-th internal coordinate, x_j and x_k are Cartesian coordinates and δ denotes displacement. Internal coordinates are curvilinear as they are nonlinear functions of Cartesian ones. The Jacobian of the internal coordinates, Wilson's $\boldsymbol{B}$ matrix,[1] forms an important quantity that relates internal and Cartesian displacements:

$$\boldsymbol{B}_{ij} = \frac{\partial\phi_i}{\partial x_j}. \tag{2}$$

$\boldsymbol{B}$ has the size of $N_i \times N_c$, with N_i being the number of internal coordinates (also the size of $\boldsymbol{\Phi}$), and N_c the number of Cartesian coordinates (also the size of $\boldsymbol{X}$). For isolated molecules $N_c = 3N$, with N being the number of atoms. For crystals, $N_c = 3N + 6$, with the extra six coordinates referring to the six independent lattice parameters. The degrees of freedom, N_f, is $N_f = 3N - 6$ for nonlinear molecules, $N_f = 3N - 5$ for linear ones, and $N_f = 3N + 3$ for crystals. Internal coordinates must be able to fully represent all possible internal motions of a molecule or crystal, thus, $N_i \geq N_f$ must be satisfied. The matrix $\boldsymbol{G}_c$ connects the $\boldsymbol{B}$ matrix with its pseudo inverse $\boldsymbol{A}$ via the definitions

$$\boldsymbol{G}_c = \boldsymbol{B}^t\boldsymbol{B}, \tag{3}$$

$$\boldsymbol{A} = \boldsymbol{G}_c^{-1}\boldsymbol{B}^t, \tag{4}$$

with the superscript t denoting transposition. For isolated molecules, $\boldsymbol{G}_c$ has $N_f = N_c - 6$ nonzero eigenvalues that refer to the internal degrees of motion of a nonlinear molecule, and six zero eigenvalues referring to translations and rotations of the whole of the molecule. Thus, for isolated molecules, the zero subspace of $\boldsymbol{G}_c$ is spanned by the N_c-sized vectors that represent the simultaneous translations and rotations of the atoms in the three spatial directions and around these directions. For crystals, $\boldsymbol{G}_c$ has $N_f = N_c - 3 = 3N + 3$ nonzero eigenvalues with the zero eigenvalues referring to the translations of the unit cell with respect to the crystal, that is, the unit cell can be chosen from the crystal in an infinitely large number of ways that can be translated into each other in a continuous fashion, along the lattice vectors, but not rotated. Thus, for crystals, the zero subspace of $\boldsymbol{G}_c$ is spanned by the three N_c-sized vectors that represent simultaneous translations of the fractional coordinates of the atoms in the three lattice directions, while leaving the lattice parameters unchanged. The matrix, $\boldsymbol{G}_c^{-1}$ must be computed as a generalized inverse for the zero eigenvalues in $\boldsymbol{G}_c$.

In internal coordinate geometry optimization, one transforms the Cartesian energy-gradients, $\boldsymbol{g}_c$, obtained for a given set of atomic Cartesian coordinates, $\boldsymbol{X}$, into curvilinear internal coordinate representation, $\boldsymbol{g}_i$, in an iterative process, using the $\boldsymbol{B}$ and $\boldsymbol{A}$ matrices:

$$\boldsymbol{g}_i = \lim_{k\to\infty} \boldsymbol{g}_i^{(k)} \tag{5}$$

$$\boldsymbol{g}_i^{(k+1)} = \boldsymbol{g}_i^{(k)} + \boldsymbol{A}^t\left[\boldsymbol{g}_c - \boldsymbol{B}^t\boldsymbol{g}_i^{(k)}\right], \tag{6}$$

with $\boldsymbol{g}_i^{(0)} = 0$ and all vector quantities here and in the followings being column vectors. The internal coordinate gradients, $\boldsymbol{g}_i$, will be used to predict (see e.g., refs. 18 and 19) new values, $\boldsymbol{\Phi}'$ of the internal coordinates that are expected to describe an energetically more stable atomic distribution. As energies and gradients can typically be calculated only over Cartesian atomic positions, a new set of Cartesian coordinates, $\boldsymbol{X}'$ is calculated, such that the corresponding set of internal coordinates, $\tilde{\boldsymbol{\Phi}}'$ is as close to the predicted $\boldsymbol{\Phi}'$ as possibly realizable by any $\boldsymbol{X}'$:

$$\boldsymbol{X}' = \lim_{k\to\infty} \boldsymbol{X}'^{(k)} \tag{7}$$

$$\tilde{\boldsymbol{\Phi}}' = \lim_{k\to\infty} \boldsymbol{\Phi}^{(k)} \tag{8}$$

$$\boldsymbol{X}'^{(k+1)} = \boldsymbol{X}'^{(k)} + \boldsymbol{A}[\boldsymbol{\Phi}' - \boldsymbol{\Phi}^{(k)}], \tag{9}$$

with $\boldsymbol{X}'^{(0)} = \boldsymbol{X}$. This process is called the "iterative back-transformation". Note that while the relationship between the internal coordinate gradients and the Cartesian ones, eq. (6), involves only the $\boldsymbol{A}$ (inverse $\boldsymbol{B}$) matrix, and knowing the exact $\boldsymbol{A}$ can lead to a single step of gradient transformation, the relationship between the Cartesian and internal coordinate displacements involves higher order derivatives as well and the knowledge of the $\boldsymbol{A}$ matrix is not sufficient in general to carry out the required transformation and an iterative process is necessary. In both eqs. (6) and (9), $\boldsymbol{A}$ plays the role of a preconditioner and thus an approximate $\boldsymbol{A}$ is sufficient for a successful transformation.

The rigidity, χ_r, of a given internal coordinate system to a given displacement of the internal coordinates can be defined as the norm of the residual, $\boldsymbol{\Delta r}$, of the internal-coordinate displacements from the iterative back-transformation, divided by the norm of the optimizer predicted internal coordinate step:

$$\chi_r = |\boldsymbol{\Delta r}|/|\boldsymbol{\Phi}' - \boldsymbol{\Phi}^{(0)}|, \tag{10}$$

$$\boldsymbol{\Delta r} = \boldsymbol{\Phi}' - \tilde{\boldsymbol{\Phi}}', \tag{11}$$

and it measures the degree of inter-dependency of the internal coordinate displacements. Note that such a measure of rigidity is a function of not only the choice of internal coordinates but also that of the minimizer that determines $\boldsymbol{\Phi}'$. Also note that χ_r may be larger than 100% in the case of extreme rigidity. Simple tests to observe the relationship between χ_r and the efficiency of the optimizer can be carried out on small molecules in Baker's test set.[14] One can for example add an extra stretching between two H atoms (attached to different carbons) of the ethane molecule. As a result, the optimization takes 11 steps instead of four with the usual[19] internals and convergence criteria. Rigidity is always zero for this molecule with the usual internal coordinates. It is however 1% with the added one in the first step and 30–50% in the subsequent ones. In general, rigidity increases the coupling of harmonic and anharmonic force constants leading to a decrease in the rate of convergence of the optimization. More specifically in the present case, the force acting along the added H–H stretching is very far from a Hooke's law type behavior that is characteristic for the other coordinates and that is favorable for the optimizer.[19] Although researchers working on the field of geometry optimization were always aware of the careful selection of internal coordinates for the above reason, to the best of our knowledge, no quantitative measure of the selection of internal coordinates has been published yet. In our opinion, χ_r is a much more selective measure of the quality of the internal coordinate set than the rate of convergence of the optimization, as the latter one measures many other properties of the optimizer and that of the energy hypersurface. For example, any displacements given in terms of the vibrational normal modes of a molecule would have no rigidity, as long as the normal modes are linear combinations of Cartesian displacements: such linear combinations of the orthogonal atomic Cartesian displacements are also perfectly orthogonal to each other. However, normal modes that are linear combinations of the displacements of primitive internal coordinates (calculated e.g., from internal coordinate force constant matrix) are orthogonal to each other only to first order of the expansion in eq. (1). The same is true for the so-called delocalized internal coordinates,[20] which are linear combinations of primitive internals that typically result in a diagonal $\boldsymbol{G}_c$ matrix. The so-called natural internal coordinates[4,21] are highly efficient (similar to delocalized internals) in reducing harmonic and anharmonic vibrational coupling and they represent a nonredundant ($N_i = N_f$) set of coordinates, however, this does not mean that the displacements along natural internal coordinates would be orthogonal to each other in the Cartesian space, as these coordinates are again linear combinations of (redundant) curvilinear primitive internal coordinates. Z-matrices[22,23] are another type of internal coordinate system that is built from stretches, valence angles, and dihedral angles, along a molecular backbone. They are typically used for quasi one-dimensional molecules, where a molecular backbone can naturally be recognized. It is very inefficient for geometry optimization of structures that are built from rings (e.g., benzene) or fused rings (e.g., diamond). Z-matrices are completely free from any rigidity and they can satisfy any prescribed displacements $\boldsymbol{\Phi}'$, however, satisfaction of these displacements may lead to unphysical structures, in which explicitly nonconnected atoms move unphysically close to each other. This problem of the Z-matrix is also known in the framework of the robotic-arm problem.[24,25] A simple example of rigidity is presented by an ammonia molecule in response to a set of prescribed $\boldsymbol{\Phi}'$ displacements, where the sum of the updated H–N–H bond-angles would be larger than 360°. Such a displacement can obviously not be realized by any spatial distribution of the atoms of ammonia.

On one hand, reducing the number of primitive internal coordinates is important, because this in general reduces the rigidity of the internal coordinate system and allows for faster optimizations. On the other hand, an internal coordinate system that represents overly large flexibility may also be dangerous for the optimization, as it may allow for unphysical proximity of some atoms. Thus a reasonable balance between rigidity and flexibility must be found for a good internal coordinate system.

To define an appropriate internal coordinate set, first the molecular connectivity has to be determined. The connectivity of atoms is usually recognized on the basis of overlapping spheres of atomic or Van der Waals radii.[8,19,26,27] Atomic radii are often not suitable for bonding recognition in, for example, ionic salts, as typical ionic radii may be 2–3 times larger or smaller then atomic ones.[28] The application of Van der Waals radii usually produces many connectivities that are neither chemically relevant nor helpful for the optimization. In fact, an overly large number of internal coordinates can substantially decrease the efficiency of an internal coordinate optimizer through increased rigidity. On the other hand, missing internal coordinates can lead to inefficient or nonconvergent optimizations.

The definition of connectivity is problematic everywhere in noncovalently bound systems. Such systems are, for example, random clusters of atoms, ionic crystals, systems with lots of intermolecular interactions (proteins, liquids, polymeric materials), or fragmented systems. Consider, for example, a system of isolated molecules that interact only through very long range forces but the intermolecular interaction cannot be described in terms of well defined bonds such as in covalently bound systems. Such a system is formed, for example, by two water molecules at a large separation and at a random orientation, where no clear hydrogen bonding exists but there is still a dipole-dipole interaction. In principle the relative position of these isolated units could be optimized in terms of Cartesian coordinates of the corresponding centers of masses and the appropriate rotations around them, however, use of relative (e.g., distance) coordinates can provide a substantial advantage, for example, by providing better control of the physically meaningful step size, avoiding the crashing of the two molecules.

Another important area where internal coordinate recognition may become troublesome is related to Van der Waals contacts. Consider, for example, two strands of the polymer, teflon. Which Van der Waals contacts should we allow for the optimization of their structures and which ones not? The same question emerges also in biopolymers, such as proteins, when one considers their interchain interactions. Or, also in many important inorganic systems, for example in sulfur crystal where Van der Waals interactions determine the size and shape of the elementary cell and also have a considerable influence on the structure of the covalently bound S_8 rings.

One important technique that reduces the rigidity in $\boldsymbol{\Phi}'$ is based on the projection method[29] that filters out so-called first order redundancies, while usually reduces the step-size. In this algorithm the projector $\boldsymbol{P} = \boldsymbol{B}\boldsymbol{G}_c^{-1}\boldsymbol{B}^t$ that projects onto the nonzero subspace of the $\boldsymbol{B}$ matrix is applied both to the approximate Hessian of the actual internal coordinate step $\boldsymbol{\Phi}'$ obtained from the optimizer.

The covariance matrix $\boldsymbol{R}$ of the rigidity vector $\boldsymbol{\Delta r}$

$$\boldsymbol{R} = \boldsymbol{\Delta r}\boldsymbol{\Delta r}^t, \tag{12}$$

can also be helpful in finding out less rigid linear combinations of internal coordinates. When summed up over several optimization steps and principal component analyzed, $\boldsymbol{R}$ can provide linear combinations of internal coordinates that minimize the rigidity, similarly to the geometric DIIS algorithm[30,31] that minimizes the distance to the expected local minimum. Note that the $\boldsymbol{R}$ (error) matrix of geometric DIIS is built as the overlap-matrix of the $\tilde{\boldsymbol{\Phi}}' - \boldsymbol{\Phi}$ displacements of several consecutive optimization steps, as opposed to the rigidity vectors, suggested here. The use of the outer product (as in eq. (12)) or the overlap matrix (as in DIIS) is identical in the sense that the smallest nonzero eigenvalue and eigenvector of these matrices will be used in both cases to construct the "minimal error" direction or the "minimal rigidity" linear combination both of which are of size N_i. A similar idea to the above is already widely used in the construction of effective normal modes,[32,33] for example, on the basis of correlated velocities (instead of $\boldsymbol{\Delta r}$-s) where velocity vectors are provided by molecular dynamics steps. The theory of rigidity of internal coordinate systems has also been discussed recently in the context of the flexibility analysis of protein regions.[34] Internal coordinate recognition algorithms also play an important role in structural biology and drug-design.[24,35]

Clearly, an important disadvantage of internal coordinates over Cartesian ones is that while Cartesian coordinates are always readily and obviously at hand, internal ones are often somewhat arbitrarily defined and thus can lead to large, definitions-based performance differences in geometry optimization. Unfortunately, a general algorithm that would overcome this inconvenience has not yet been developed. In this, we describe our algorithm to define internal coordinates for all possible chemical systems in a unified process with the aim of optimizing the rigidity and the flexibility of internal coordinate systems.

Coordinate-Selection Algorithms

Selection Based on the Condition Number of $\boldsymbol{G}_c$

Let $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum of the N_f nonzero eigenvalues of $\boldsymbol{G}_c$ and $\lambda_c = |\lambda_{\max}|/|\lambda_{\min}|$ the condition number of the nonzero subspace of $\boldsymbol{G}_c$. λ_c can always be calculated once an internal coordinate set is given and it is indicative to how homogeneously the internal coordinates span the space of internal motions. A condition number $\lambda_c \approx 1$ would indicate that all sorts of internal motions are equally well described by linear combinations of the actual choice of internal coordinates. A large condition number is indicative of an uneven behavior of the internal coordinate set in the description of all internal degrees of freedom. In this sense, $\boldsymbol{G}_c$ plays the role of a metric matrix. The first condition (Condition #1) that any internal coordinate set must satisfy is that $\boldsymbol{G}_c$ must have N_f nonzero eigenvalues. This condition refers to the requirement that the internal coordinates must span the whole space of internal motions of the molecule. A typical situation when this condition is not satisfied is when fragments of the molecule are not connected properly by internal coordinates resulting in less then N_f nonzero eigenvalues. Although this observation is in the common experience of the community of internal coordinate optimization developers, to the best of our knowledge nobody has yet called the attention to the intimate relationship between λ_c and the quality of internal coordinate sets. Once this first condition is satisfied, one can think of making selection between the actual internal coordinates and search for a reduced set of internal ones that still satisfies Condition #1 with the minimal number of internal coordinates, so that rigidity can be reduced. Say, for $N = 100$ atoms molecule $N_i = 294$ internal coordinates can be selected that satisfy Condition #1. However, these $N_i = 294$ internal coordinates can be selected in lots of different ways, that is, lots of subsets of the originally $N_i > 294$ elements internal coordinate set can be chosen. To make a further selection among these subsets, Condition #2 can be introduced: let us choose the one among equal-sized subsets, which provides the smallest λ_c condition number, that is, the most possibly even description of all internal degrees of freedom. This way both the size and the choice of internal coordinates can be chosen uniquely, unless degenerate subsets occur in the λ_c sense.

It is needless to say that the calculation of λ_c for all possible selections of internal coordinates would be very demanding computationally, even if it can be done very efficiently for each individual coordinate set using sparse matrix techniques. Thus, we consider the λ_c-based selection procedure impractical for large molecules, even though it provides a mathematically strict formulation for reducing the rigidity of a coordinate system.

Linear Combination of Primitive Internal Coordinates

The condition number λ_c for a given internal coordinate set can also be modified by the construction of appropriate linear combinations of a set of internal coordinates. Usually, the starting internal coordinate set consists of primitive internal coordinates. One well established set of such linear-combinations is the natural internal coordinates.[4,27] Unfortunately, the construction of natural internal coordinates is difficult to automatize especially for systems that contain lots of fused rings.[27,36] Alternatively, one can use the so-called delocalized internal coordinates.[20] Delocalized internal coordinates provide a N_f dimensional set of linear-combined primitive internal coordinates, where the linear combination coefficients come from a unitary matrix, $\boldsymbol{U}$. $\boldsymbol{U}$ is constructed either by diagonalizing the matrix $\boldsymbol{G}_i = \boldsymbol{BB}^t$ and taking eigenvectors of its nonzero subspace or by QR-decomposition of $\boldsymbol{B}$. Note that these two procedures result in different $\boldsymbol{U}$ unitary matrices. The purpose of constructing $\boldsymbol{U}$ is solely to use the nonzero subspace of $\boldsymbol{G}_i$ to get $\boldsymbol{N}_f$ orthogonal linear combinations. Any unitary rotation of $\boldsymbol{U}$ within the nonzero subspace of $\boldsymbol{G}_i$ can be considered a legitimate choice. In practice only the $\boldsymbol{U}$ matrices constructed by diagonalization[20,37] or by QR-decomposition (e.g., in ref. 8) are used. The effect of the additional unitary transformation of $\boldsymbol{U}$ has not been studied yet. Also note that the generation of the delocalized internal coordinates for very large molecules can not be carried out in a robust fashion, as both diagonalization and QR-decomposition are inefficient for this purpose. Although diagonalization scales cubically with system size, sparse QR-decomposition can best achieve quadratic scaling. Scaling of the rows of the $\boldsymbol{B}$ matrix while setting the scale factors such that λ_c be minimal could also provide a modified $\boldsymbol{B}$ matrix from which a suitable $\boldsymbol{U}$ linear-combination matrix can be generated by diagonalization or QR decomposition. Such a scaling may be carried out, for example, by using empirical force constants of internal coordinates.[38] Similar means of scaling have specifically been developed in the context of intermolecular interactions.[39–42] We do not discuss further the applicability of delocalized internal coordinates for very large systems as our primary goal here is to focus on how to select primitive internals for a given molecule.

Selection Based on Cholesky Factorization

Paizs et al. describe a technique, based on Cholesky factorization, to determine the N_f most independent coordinates of a redundant internal coordinate set.[36] The technique can be viewed as a sparse Cholesky factorization with column pivoting (see e.g., ref. 43). Such a technique is typically used to determine trapezoidal (incomplete) Cholesky factors of positive semidefinite matrices, such as the $\boldsymbol{G}_i$ matrix in the above discussions. The idea is based on the use of pivoting when factorizing the $\boldsymbol{G}_i$ matrix. The most independent N_f internal coordinates are supposed to have large pivots, which makes it possible to select a unique N_f dimensional set of coordinates, unless substantial degeneracies occur.

As is well known, the application of pivoting is a major computational bottleneck when applied in large sparse Cholesky factorization[44] as it results in a huge amount of work in reorganizing matrices in the memory. For this reason pivoting is most often avoided in sparse Cholesky factorizations. Instead, efficient symmetric permutations are used, such as the reverse Cuthill–McKee ordering,[44] to reduce filling of the Cholesky factors. Because of the technical difficulties, pivoting-based coordinate selection does not seem efficient for large molecules.

Connectivity Analysis

The von Arnim–Ahlrichs Approach

In our opinion the best algorithm for practical coordinate recognition so far described in the literature is the one by von Arnim and Ahlrichs.[27] Here we only focus on their algorithm of recognizing the molecular connectivity. In this approach, atomic radii are gradually increased while overlapping atomic spheres are checked until the whole molecule becomes interconnected. In the first phase somewhat scaled atomic radii are used to recognize the strongest bonds. In the second phase, when atomic radii are increased, first hydrogen bonds are checked and if fragments are still not connected, bonds between any other atoms are also allowed. The great advantage of the von Arnim–Ahlrichs approach of connectivity recognition against all previously mentioned analysis is that it can be carried out in an extremely robust fashion and does not require much use of large linear algebraic manipulations, as opposed to, for example, pivoting-based Cholesky factorization or λ_c-based coordinate selection methods. The rest of the von Arnim–Ahlrichs approach that deals with the construction of linear combinations of primitive internals may however fail in the recognition of the N_f natural internal coordinates, especially in topologically complex systems, for example, in systems that contain multiply fused rings.

Length-scale Scanning with Topology Analysis

Our algorithm for the recognition of connectivity is similar to the one of von Arnim and Ahlrichs in the sense that it also gradually scans ranges of atomic radii and checks the connectivity, until all fragments are connected. Although we carry out the same kind of length-scale scanning, the way we do it is efficiently adapted for use in very large molecules. Also, the intermediate steps of connectivity recognition differ from the ones used by von Arnim and Ahlrichs. Our algorithm consists of two loops: loop #1 and loop #2.

The goal of loop #1 is to find the shortest bonds that build a molecular topology (atomic connectivity) in which the molecule is not fragmented. These bonds provide a molecular skeleton, which together with the corresponding angles (bendings, torsions, out-of-planes, and linear-bendings) is sufficient to satisfy Condition #1, that is, these coordinates span the whole space of internal motions.

The goal of loop #2 is to recognize those bonds, which have not been recognized in the first loop, because they have not been dominant bonds in their local environment, but still have relevance for the optimization. These bonds are usually the weaker bonds, such as hydrogen-bonds, Van der Waals contacts, or ionic contacts.

In both loops, the molecule is divided into cubic boxes and the serial numbers of atoms falling in these boxes are sorted in a sparse matrix fashion, so that atoms of a specific box can be identified easily. The linear size of the box starts at 3.0 Å and will be multiplied by a factor of 1.05 in each new cycle of the loops with the purpose of allowing longer bonds to appear within boxes and between neighboring boxes. Then all possible atomic connectivities are checked within each box and its neighbors in both loop #1 and loop #2. Thus, all short-range connectivities can be looked up efficiently.

Loop #1

Before loop #1 starts, all atomic connectivities are checked for overlapping spheres of atoms with a radius of 1.3 times the atomic Slater radii.[28] A bond-list is formed and a sparse topology matrix is set up. The reverse Cuthill–McKee ordering[44] is applied to the sparse topology matrix to bring it into a block-structured form. In this ordering, unconnected fragments of the molecule appear as isolated blocks in the topology matrix. Based on this block structure, atoms are associated with serial numbers of fragments they are part of.

Then loop #1 starts. In this loop, atoms will be associated with a uniform radius to find closest contacts between fragments. This uniform radius starts at 0.75 a.u. and increases by a factor of 1.05 in each new cycle. In each cycle, overlap will be checked between atoms that either belong to different fragments or are at a shorter distance than 1.33 times a typical bondlength already attached to the atoms of the pair being investigated. This typical bondlength is the maximum of the shortest bondlengths of each of the atoms in the pair. This latter criterion is important to avoid the formation of Z-matrix like internal coordinate sets in several crystals, for example in distorted NaCl. Based on the updated bond-list, the fragmentation of the system is determined again and if no fragmentation occurs, loop #1 terminates, otherwise repeats the above. For most chemical systems ranging from ionic crystals through small and medium sized organic compounds to hydrogen- or Van der Waals bonded crystals, loop #1 is enough to determine an efficient bonding scheme for geometry optimization.

Loop #2

For some chemical systems in which weak bonds connect topologically far, but spatially close units another loop is necessary as these connectivities can be discovered neither on the basis of fragmentation nor by comparison to covalent bondlengths. For example, in a folded protein the first loop may recognize only the covalent skeleton. All weak interchain bonds must be added afterward.

In loop #2 topologically far but spatially close units of the molecule will be connected. The definition of "topologically far" is somewhat ambiguous. In our current recognition scheme "topologically far" means at least the 7th topological neighbor. This means that smaller than seven membered rings are not allowed to be closed by Van der Waals or ionic contacts in loop #2. We make an exception from this topology-based rule for "strong hydrogen bonds". Our criterion for "strong hydrogen bond" is that in the bond X–H· · ·Y atoms X and Y must be of the atoms N,O,F,Cl and the XHY bond angle must be between 180° and 120° and the X–Y distance smaller than 3.5 Å.

At the beginning of each cycle of loop #2 an additional topology matrix, T_{excl} is built:

$$T_{\text{excl}} = (I + T_{12})^{n_{\text{excl}}}, \qquad (13)$$

where I denotes the $N \times N$ identity, T_{12} is the topology matrix based on all bonds from loop #1 and previous cycles of loop #2, and n_{excl} refers to "topologically far", which is actually $n_{\text{excl}} = 6$. T_{excl} will thus list for each atom what other atoms are in its topological proximity of less than seven neighbors. Only the symbolic part of these sparse matrices is stored in ordered sparse row-wise representation,[45] and only symbolic multiplications are carried out. All atomic contacts between atoms i and j will be excluded from the bond-list for which the corresponding element of T_{excl} is not zero.

Loop #2 starts with the same box-size as loop #1, and atomic radii start at 0.5 times the tabulated values of Van der Waals radii[46] and are increased by a factor of 1.05 in each cycle while the box-size is increased by the same factor.

This two-loops-based recognition algorithm is now the default in our geometry optimization code and it is capable to reproduce our previously published results on Baker's test set and enzyme fragments optimizations[19] whereas it can be used with great efficiency in several challenging cases such as those discussed later in this article. We also discuss an alternative connectivity recognition scheme in Appendix A, parts of which are used in the recognition of bending (angle) coordinates as well.

Angles Related Internal Coordinates

Our current internal coordinate generator adds all possible bond angles (bending coordinates) to the internal coordinate list that is determined by nearest neighbor bonds. The planarity of the vicinity of a certain atom i is checked by calculating eigenvalues of the $\boldsymbol{S}^i$ matrix as described in Appendix A. If the $\boldsymbol{S}^i$ matrix has only two nonzero eigenvalues (after the application of the constraint on the condition number of $\boldsymbol{S}^i$) all possible out-of-plane coordinates around atom i are generated with i in the central position. Note that we use the spectroscopic out-of-plane coordinates[1] instead of the improper torsions of molecular mechanics.

Every bond will be associated with a single torsional coordinate, if applicable. The torsion is chosen so that the two valence angles at the ends of the torsions should be as close to 90° as possible, should contain the heaviest possible atoms for the terminal positions and the terminal atoms should be those having the largest number of topological neighbors. In fact, a weight function of the form

$$W_j = \frac{(N_j + 1)Z_j}{|\pi/2 - \alpha_j|} \qquad (14)$$

is used, where j is a candidate atom for a terminal position of the torsion, N_j is the number of its nearest neighbor atoms, Z_j is the atomic number of j and α_j is the angle j closes with the central atoms of the torsion.

If an angle is linear, that is, between 175° and 180°, a linear bending coordinate is set up, independent from whether the central atom has two or more neighbors. If the terminal atoms of a linear bending have more than one neighbors, one of these neighbors will be selected to reference the two perpendicular planes in whose intersection the linear bending is placed. If there is no such reference, the Cartesian framework will serve as a reference as described in ref. 26. Note that to each linear bending coordinate there will be a long-range-torsion associated that is constructed similarly to ordinary torsions but its central atoms will be the terminal atoms of the linear bending. In case of collinear chains linear bendings are generated for each atomic triplet of the chain with an appropriate long-range torsion that bridges the two ends of the collinear chain.

Case Studies

Here, we describe three difficult cases in which finding the appropriate bonds turned out troublesome in earlier versions of our default coordinate recognition algorithm. All calculations have been carried out using the MondoSCF[48] suit of quantum chemistry codes, where the above described coordinate recognition algorithms have been implemented. The optimizations were based on the QUICCA algorithm[19] as modified for crystals. Energies and forces have been computed on the PBE/STO-3G level of theory, in the Γ-point approximation. Optimization was considered converged after all atomic and lattice gradients decreased below 0.0005 a.u. All coordinates including the lattice parameters were allowed to relax. In this article, we present optimization characteristics for teflon, only. A more detailed description of the performance of our recently developed crystal structure optimizer can be found in ref. 18 also with performance data on further difficult cases of internal coordinate definitions, such as the sulfur crystal. Here, we intend to present only a few pitfalls of the selection of internal coordinates without the aim of demonstrating any ultimate way of constructing internal coordinate sets, as the aim of this article is to provide novel principles and analysis tools to the quality of internal coordinate sets.

Pentaerythritol tetranitrate – A High Explosive With Hydrogen Bonds

Pentaerythritol tetranitrate (PETN) is a high explosive.[47] Its crystal structure is bound together by C–H· · ·O–N type hydrogen bonds (see Fig. 1). The elementary cell of PETN contains 58 atoms. Our recognition algorithm founds 91 stretches, 174 bendings, 24 linear bendings, 101 torsions, 16 long-range torsions, and 36 out-of-planes. Note that internal coordinates that have no atom in common with the central cell are not listed here. Among the stretches, H-bond related coordinates can be divided into two classes: one of approximate length of 2.48 Å, there are 16 such stretches; and the ones of length about 2.61 Å (12 pieces). The difficulty in recognizing the appropriate connectivity in this case originated from the fact that if nonuniform atomic radii are used, loop #1 found some longer bonds for Van der Waals contacts, with length of about 3.14 Å, when it

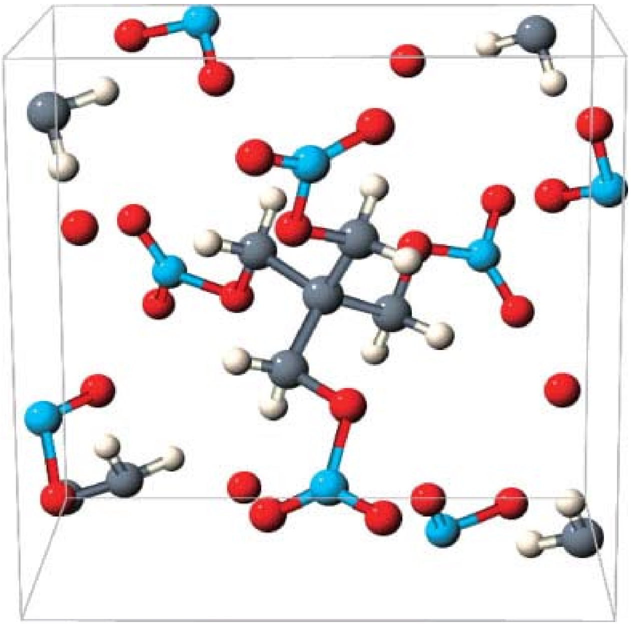

Figure 1. Crystal structure of PETN as of Ref. 47.

connected fragments that are isolated in the primary covalent bonding scheme. With the additional Van der Waals contacts and the assotiated angle-coordinates the coordinate system would become much more rigid and thus optimization could proceed less quickly.

Distorted Sodium Chloride

Distorted sodium chloride occurred during optimization with a previous version of our internal coordinates recognition scheme. It also models well the case of molten ionic crystals and, in general the case, when atomic radii are very poor approximations to discover bonding. The elementary cell consist of two atoms. All 27 stretches have been recognized that connect nearest sodium–sodium, cloride–chloride, and sodium–chloride ions. Note that changing the factor of 1.33 (multiplier for "typical" bond-length, see Section Length-scale Scanning with Topology Analysis) to a smaller value, say 1.2 results in the recognition of sodium–chloride interactions only. In the default algorithm, 276 bendings, 80 linear bendings, 63 ordinary torsions, and 51 long-range torsions have been detected. Clearly, this represents a quite redundant coordinate system for sodium chloride, as the stretches alone should be sufficient for the optimization.

Teflon Crystal

The crystal of teflon (poly-tetrafluoro-ethylene)[49,50] is held together through Van der Waals contacts (see Fig. 2). In our model a 156 atoms elementary cell was used. Our recognition scheme found 244 stretches, 204 of them ordinary covalent bonds, and 40 Van der Waals contacts of approximate length of 2.60–2.85 Å. These bonds are associated with 607 bendings and 281 torsions. The number of intermolecular bonds also depends on the topological exclusion principle presented in Section Length-scale Scanning with Topology Analysis. Note that for the Van der Waals contacts we have not used the so-called "5/R" coordinates, as suggested elsewhere.[8,40]

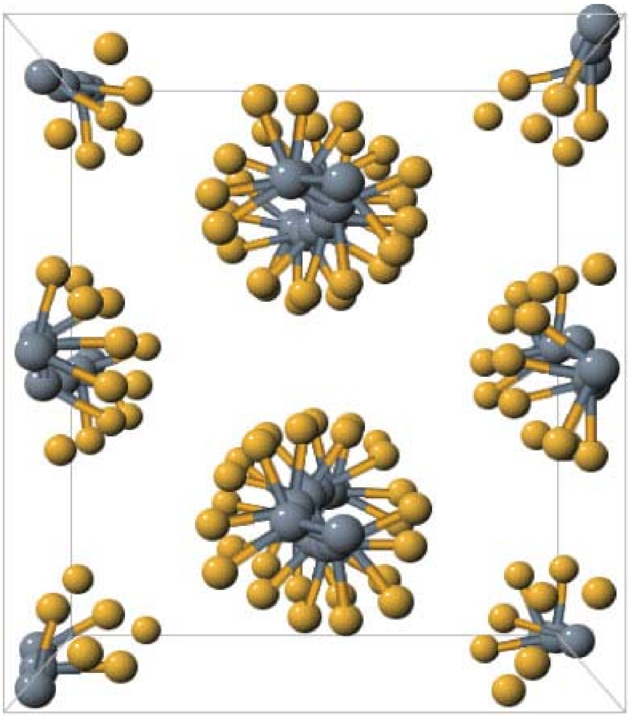

Figure 2. Optimum crystal structure of teflon on the PBE/STO-3G level of theory.

The teflon system has been optimized in 54 steps from a geometry with initial gradients as large as 0.601103 a.u. in magnitude. The energy curve (see Fig. 3) has a smooth decline, while gradient curves (Fig. 4) show slow decline for the intermolecular forces. One possible source of improvement in this decline could be the application of "5/R" coordinates. The rationale behind the use of "1/R" type coordinates is that they can correctly describe the asymptotic behavior of stretching potentials with a low order polynomial, such as for example in the case of Lennard-Jones pair-potential.[21,39]

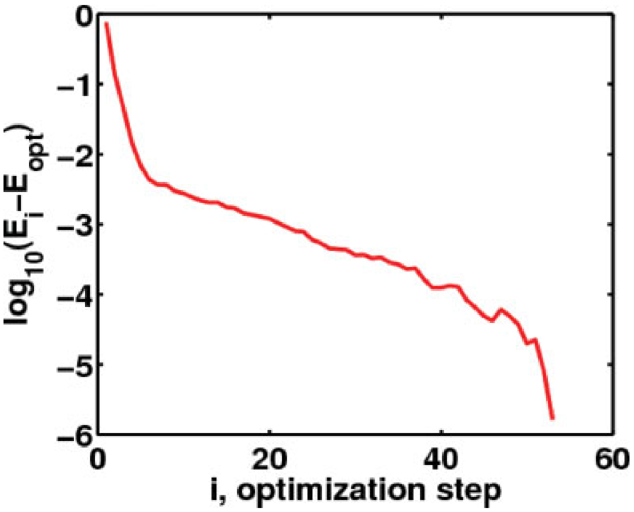

Figure 3. Convergence of the energy during the optimization of teflon.

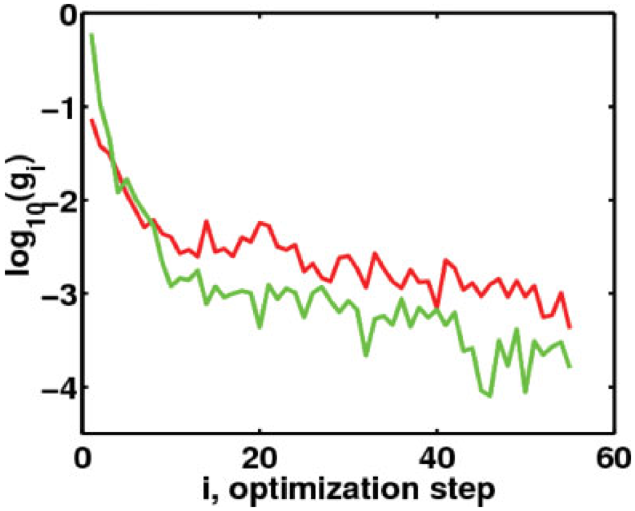

Figure 4. Convergence of the maximum atomic (lower curve) and lattice (upper curve) forces during the optimization of teflon.

Conclusions

We have discussed strengths and weaknesses of existing algorithms for the selection of internal coordinates for the purpose of molecular geometry optimization. We have recommended new, robust algorithms that are capable to recognize efficient coordinate systems for topologically complex atomic arrangements that span the range from ionic melts to Van der Waals bonded crystals. With the algorithms described in this article we have contributed to the development of internal coordinate recognition that aims on generalizing the use of internal coordinates for all situations where Cartesian coordinates have been the preferred means of geometry manipulation. We have enumerated difficult cases in coordinate recognition and provided robust solutions that proved to work well in complex, practical recognition tasks.

Acknowledgement

N.J. Henson and T.D. Sewell (Los Alamos National Lab.) are acknowledged for providing starting structures for teflon and PETN and for discussions.

Appendix A: Connectivity Recognition on the Basis of Shell-Structures

To simplify the connectivity recognition described in Section Connectivity Analysis, we have tested another algorithm that is based on the investigation of shell-structures around atoms.

For each atom i, an array of local bond-vectors $\boldsymbol{v}_j$ that point from atom i to atom j is set up, and stored in the matrix $\boldsymbol{L}^i$:

$$\boldsymbol{L}^i_j = \boldsymbol{v}^t_j, \tag{A1}$$

with

$$\boldsymbol{v}_j = w_{ij}(\boldsymbol{r}_j - \boldsymbol{r}_i)/|\boldsymbol{r}_j - \boldsymbol{r}_i|, \tag{A2}$$

where $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$ are the position vectors of the atoms i and j, and w_{ij} is a weight constructed by the overlap of two hypothetical spherical Gaussian fuctions with exponents $\xi_i = 1/R_i^2$ and $\xi_i = 1/R_j^2$, where R_i and R_j are the corresponding atomic radii:

$$w_{ij} = \left(\frac{\pi}{\xi_i + \xi_j}\right)^{3/2} \exp\left(-\frac{\xi_i\xi_j|\boldsymbol{r}_j - \boldsymbol{r}_i|^2}{\xi_i + \xi_j}\right). \tag{A3}$$

The Gaussian weights refer to a simple model of atomic orbital overlap. For close atoms this overlap is big, for farther atoms it is small, thus the weighting serves as a selection criteria for the importance of bonds.

The eigenvectors $\boldsymbol{U}$ of the 3 × 3 matrix $\boldsymbol{S}^i$ defined as

$$\boldsymbol{S}^i = \boldsymbol{L}^{i^t}\boldsymbol{L}^i, \tag{A4}$$

$$\boldsymbol{S}^i = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t, \tag{A5}$$

provide principal bonding directions around atom i. These principal bonding directions can be expressed in terms of the bond vectors $\boldsymbol{v}_j$ via the matrix $\boldsymbol{K}^i$:

$$\boldsymbol{K}^i = \boldsymbol{L}^i\boldsymbol{U}\boldsymbol{D}^{-1/2}, \tag{A6}$$

where $\boldsymbol{D}^{-1/2}$ is a diagonal matrix and contains the generalized inverse square-roots of the eigenvalues of $\boldsymbol{S}^i$. Singularities of $\boldsymbol{S}^i$ are treated by the constraint that all eigenvalues smaller then 0.01 times the largest eigenvalue of $\boldsymbol{S}^i$ are set to zero and are not inverted. Note that the columns of $\boldsymbol{K}^i$ are orthonormal to each other. The total weight W_{ij} of bond ij among the bonds of i is calculated as

$$W_{ij} = \sum_{l=1}^{3} \boldsymbol{K}^{i\,2}_{jl}, \tag{A7}$$

representing the contribution of each bond ij to principal bonding directions. What remains is the filtering of bonds that are important for atom i. This is based on the assumption that unimportant bonds will have a small enough weight. The weights W_{ij} are normalized such that they add up to one for each atom i, and are ordered in decreasing order. Then, the values Q_{ij} are computed:

$$Q_{ij} = \frac{\log(W_{ij}) - \log(W_{i(j-1)})}{\log(W_{ij}) - \log(W_{i1})}. \tag{A8}$$

Note that from the above equation the index j refers to decreasing ordering in W_{ij}. The values of Q_{ij} refer to the statistical Q-test to check whether an outlier of a series can be dropped. The tabulated Q-test values, Q^t_j for $j = [3, 10]$ are: 0.94, 0.76, 0.64, 0.56, 0.51, 0.47, 0.44, and 0.41 for 90% probability, respectively. If the bondlist has only two elements, we require that the second bond should be dropped if the first one accounts for 90% of the total weights. Otherwise, for a larger set of bonds, the series of allowed bonds will terminate at that specific $k - 1$-th bond, for which $Q_{ik} < Q^t_k$, supposing that $\sum_{l=1}^{k-1} W_{il} > 0.90$, where 0.90 refers to 90% of the weights. This simple algorithm is capable to correctly recognize all bonds in Baker's test set of small and medium sized organic molecules and reproduce our previously published geometry optimization results.[19] However, when some of the bonds are too long, like a long Van der Waals contact, this algorithm fails. More appropriate choice of the Gaussian-based weights may though help and the algorithm may become generally applicable.

References

1. Wilson, E. B.; Decius, J. C.; Cross, P. C. Molecular Vibrations, McGraw-Hill, New York, 1955.
2. Pulay, P. Mol Phys 1969, 17, 197.
3. Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. J Am Chem Soc 1979, 101, 2550.
4. Fogarasi, G.; Zhou, X.; Taylor, P. W.; Pulay, P. J Am Chem Soc 1992, 114, 8192.
5. Pulay, P. In Schäfer, H. F. III. Ed.; Modern Theoretical Chemistry, Vol. 4; Plenum: New York, 1977, pp. 153–185.
6. Schlegel, H. B. J Comp Chem 1982, 3, 214.
7. Schlegel, H. B. J Comput Chem 2003, 24, 1514.
8. Bučko, T.; Hafner, J.; Ángyán, J. J Chem Phys 2005, 122, 124508.
9. Pulay, P.; Paizs, B. Chem Phys Lett 2002, 353, 400.
10. Lee, S.-H.; Palmo, K.; Krimm, S. J Comp Chem 2007, 28, 1107.
11. Ashcroft, N. W.; Mermin, N. D. Solid State Physics, Brooks/Cole, Thomson Learning: Australia, Canada, Mexico, Singapore, Spain, UK, US, 1976.
12. Goedecker, S.; Lancon, F.; Deutsch, T. Phys Rev B 2001, 64, 161102.
13. Fernandez-Serra, M. V.; Artacho, E.; Soler, J. M. Phys Rev B 2003, 67, 100101.
14. Baker, J. J Comp Chem 1993, 14, 1085.
15. Baker, J.; Chan, F. J Comput Chem 1996, 17, 888.
16. Kudin, K. N.; Scuseria, G. E.; Schlegel, H. B. J Chem Phys 2001, 114, 2919.
17. Andzelm, J.; King-Smith, R. D.; Fitzgerald, G. Chem Phys Lett 2001, 335, 321.
18. Németh, K.; Challacombe, M. J Chem Phys 2005, 123, 194112.
19. Németh, K.; Challacombe, M. J Chem Phys 2004, 121, 2877.
20. Baker, J.; Kessi, A.; Delley, B. J Chem Phys 1996, 105, 192.
21. Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. J Am Chem Soc 1979, 101, 2550.
22. Hehre, W. J.; Radom, L.; Schleyer, P. R. v.; Popple, J. A. Ab initio Molecular Orbital Theory, Wiley: New York, 1986.
23. Mezey, P. G.; Potential Energy Hypersurfaces, Studies in Physical and Theoretical Chemistry, Elsevier: Amsterdam, 1987.
24. Kleywegt, G. J.; Henrick, K.; Dodson, E. J.; van Aalten, D. Structure 2003, 11, 1051.
25. Bae, D. S.; Haug, E. J. Mech Struct Mach 1987, 15, 359.
26. Bakken, V.; Helgaker, T. J Chem Phys 2002, 117, 9160.
27. von Arnim, M.; Ahlrichs, R. J Chem Phys 1999, 111, 9183.
28. Slater, J. C. J Chem Phys 1964, 41, 3199.
29. Pulay, P.; Fogarasi, G. J Chem Phys 1992, 96, 2856.
30. Császár, P.; Pulay, P. J Mol Struct 1984, 114, 31.
31. Farkas, Ö.; Schlegel, H. B. Phys Chem Chem Phys 2002, 4, 11.
32. Karplus, M.; Kushick, J. N. Macromolecules 1981, 14, 325.
33. Strachan, A. J Chem Phys 2004, 120, 1.
34. Rader, A. J.; Hespenheide, B. M.; Kuhn, L. A.; Thorpe, M. F. Proc Natl Acad Sci USA 2002, 99, 3540.
35. Schüttelkopf, A. W.; van Aalten, D. M. F. Acta Crystallogr D 2004, 60, 1355.
36. Paizs, B.; Baker, J.; Suhai, S.; Pulay, P. J Chem Phys 2000, 113, 6566.
37. Andzelm, J.; Wimmer, E. J Chem Phys 1992, 96, 1280.
38. Lindh, R.; Bernhardsson, A.; Schütz, M. Chem Phys Lett 1999, 303, 567.
39. Baker, J.; Pulay, P. J Chem Phys 1996, 105, 11100.
40. Baker, J.; Pulay, P. J Comp Chem 2000, 21, 69.
41. Maslen, P. J Chem Phys 2005, 122, 014104.
42. Swart, M.; Bickelhaupt, F. M. Int J Quantum Chem 2006, 106, 2536.
43. Golub, G.; van Loan, C. F. Matrix Computations, Johns Hopkins University Press: Baltimore, 1996.
44. George, A.; Liu, J. W.-H. Computer Solution of Large Sparse Positive Definite Systems, Prantice-Hall, Eaglewood Cliffs: New Jersey, 1981.
45. Pissanetzky, S. Sparse Matrix Technology, Academic Press: London, 1984.
46. Winter, M. Webelements tm periodic table (professional edition), The University of Sheffield and WebElements Ltd, UK, Accessed 2004. Available at: http://www.webelements.com, 2004.
47. Conant, J. W.; Cady, H. H.; Ryan, R. R.; Yarnell, J. L.; Newsam, J. M. Tech. Rep. LA-7756-MS, Los Alamos National Laboratory (1979).
48. Challacombe, M.; Schwegler, E.; Tymczak, C.; Gan, C. K.; Nemeth, K.; Niklasson, A. M. N.; Nymeyer, H.; Henkleman, G. MondoSCF v1.0α9, a program suite for massively parallel, linear scaling scf theory and ab initio molecular dynamics. Los Alamos National Laboratory (LA-CC 01-2), University of California: Los Alamos, NM, 2001. Available at: http://www.t12.lanl.gov/home/mchalla/.
49. Clark, E. S. Polymer 40, 46594665 (1999).
50. Flack, H. D.; J Polym Sci A-2 1972, 10, 1799.