Energy-directed tree search: an efficient systematic algorithm for finding the lowest energy conformation of molecules†

Ekaterina I. Izgorodina,‡ Ching Yeh Lin and Michelle L. Coote*

Received 22nd January 2007, Accepted 20th March 2007
First published as an Advance Article on the web 4th April 2007
DOI: 10.1039/b700938k

We present a new systematic algorithm, energy-directed tree search (EDTS), for exploring the conformational space of molecules. The algorithm has been designed to reliably locate the global minimum (or, in the worst case, a structure within 4 kJ mol^{-1} of this species) at a fraction of the cost of a full conformational search, and in this way extend the range of chemical systems for which accurate thermochemistry can be studied. The algorithm is inspired by the build-up approach but is performed on the original molecule as a whole, and objectively determines the combinations of torsional angles to optimise using a learning process. The algorithm was tested for a set of 22 large molecules, including open- and closed-shell species, stable structures and transition structures, and neutral and charged species, incorporating a range of functional groups (such as phenyl rings, esters, thioesters and phosphines), and covering polymers, peptides, drugs, and natural products. For most of the species studied the global minimum energy structure was obtained; for the rest the EDTS algorithm found conformations whose total electronic energies are within chemical accuracy from the true global minima. When the conformational space is searched at a resolution of 120°, the cost of the EDTS algorithm (in its worst-case scenario) scales as 2^N for large N (where N is the number of rotatable bonds), compared with 3^N for the corresponding systematic search.

1. Introduction

The location of stationary points, particularly energy minima, on the potential energy surfaces of large molecules plays an important role in computational quantum chemistry. Standard geometry optimisation routines are typically based on the Newton–Raphson procedure in which derivatives are followed, so as to find the nearest local minimum energy structure. However, it is well known that any potential energy surface may contain several local energy minima, many of which can be many tens (or even hundreds) of kJ mol^{-1} higher than the lowest energy conformation. It is therefore important to take additional steps to identify the global minimum structure in order to predict accurate thermodynamic and kinetic quantities. Algorithms for efficiently locating the lowest energy conformations are therefore of great practical importance. This paper describes a new type of conformational searching algorithm, the *energy-directed tree search* method, which is designed to combine the accuracy of a full conformational search routine, such as *tree search*,[1] but at a significantly reduced computational cost. The algorithm has been developed with a view to extending the range of molecular systems for which accurate thermochemistry can be studied.

The most rigorous method for locating the global minimum is to perform a full conformational search. For most molecular systems, this would entail the calculation of the potentials for simultaneous rotations about every bond in the molecule, yielding a large multi-dimensional potential energy surface from which the lowest energy structure can be identified. In practical full search methods (such as *tree search*[1] and variants such as *SUMM*[2,3] and *complementarity*[4]), the torsional angles are varied at some specified degree of resolution to generate starting structures for Newton–Raphson geometry optimisations. The search can be further simplified by rejecting symmetry equivalent species and/or "unreasonable structures" (*i.e.* those having close contacts between non-bonded atoms and/or, in the case of cyclic structures, failing to satisfy ring closure constraints). Provided the resolution is fine enough, and the filtering is not over-zealous, such methods can almost guarantee the location of every minimum energy structure on the potential energy surface, and hence allow one to identify the global minimum with a very high degree of certainty. Although the need to specify the resolution adds a certain degree of non-objectivity to such methods, one can normally make an informed choice, based on chemical knowledge. For example, in an organic molecule a bond between two sp^3

Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia. E-mail: mcoote@rsc.anu.edu.au; Fax: 61 2 6125 0750

† Electronic supplementary information (ESI) available: The optimization results for the EC1, EC2 and NMAX parameters for three additional species (Tables S1–S3); two fully worked examples of the EDTS algorithm for two scenarios shown in Fig. 1 (Appendix S1); the details on the efficiency of the algorithm in the "worst-case scenario" (Appendix S2); optimised geometries in the form of Gaussian archive entries for all species in the test set, both in their starting conformations and final global minimum conformations (Appendix S3–S4).. See DOI: 10.1039/b700938k

‡ Current address: School of Chemistry, Monash University, Victoria 3800, Australia.

carbon centres should normally have at most 3 local minima separated at approximately (though not necessarily exactly) 120°, and this degree of resolution is usually sufficient if one commences the search from an optimised structure and performs subsequent geometry optimisations on each point. For problematic cases, a resolution of 60° virtually ensures that structures are not missed.

When implemented at an appropriate degree of resolution, full conformational search methods such as tree search are objective and offer reliable access to the global minimum energy structure in a finite number of steps. As such, these methods (either explicitly or implicitly) have been almost universally adopted in accurate thermochemical studies of small molecular systems. However, while they are practical for small systems, such methods suffer from a combinatorial explosion problem with increasing molecular size. For example, at a resolution of 120°, a full conformation search on a molecule with 2 rotatable bonds might entail the optimisation of just 9 (3^2) starting structures, however, with 7 bonds the conformational space grows to over 2000 (3^7) starting structures and with 10 bonds there would be nearly 60 000 (3^{10}) structures to consider. In such cases, a complete search is, of course, impractical. More generally, such methods scale as $(360°/R)^N$, where R is the resolution and N the number of rotatable bonds.

Until recently, the application of high-level quantum-chemical methods (*i.e.* those having "kcal accuracy") has itself been limited to relatively small systems for which the conformational space is small and manageable. However, in recent years this situation has changed: the increase in available computer power, and other developments such as parallel code, linear scaling methods and efficient algorithms, now allows us to carry out calculations at highly sophisticated levels of theory for systems having thousands of basis functions. For example, using an ONIOM-based procedure that sequentially improves large basis set RMP2 calculations to the G3(MP2)-RAD and then W1 levels of theory, we recently reported accurate thermochemical calculations on the reactions of trimeric styryl radicals with dithioester compounds such as S=C(Ph)SC$(CH_3)_2$CN.[5] However, this work entailed the laborious task of optimising thousands of starting structures for the various reactants and products, so as to identify a small number of global minimum energy structures. In essence, we are now in a situation where the computational bottleneck is the conformational searching rather than the high-level single point energy calculations, even though the former can be reliably performed at much lower levels of theory. Efforts to extend further the range of systems for which accurate thermochemistry can be studied are thus dependent upon the development of reliable methods for searching conformational space. This is the aim of the present study.

In the present work, we present a new systematic algorithm for locating global minimum energy structures, and then evaluate it for a range of organic molecules for which full conformational searches have also been performed. It should be stressed that, unlike many other studies of this kind, our primary aim is to find an algorithm that *reliably* locates the global minimum (or, in the worst case, a structure within 4 kJ mol^{-1} of this species) with a very high degree of confidence, rather than merely locating relatively low energy structures at an efficient rate. Having satisfied this condition, other desirable properties of the algorithm include computational efficiency, objectivity in its implementation, and ease of use. In what follows, we first outline the principal existing algorithms and discuss their potential suitability for accurate thermochemical studies, we then describe our new algorithm, and conclude with computational testing.

2. An overview of current conformational search methods

The problem of locating global (or even low energy) conformations has long plagued the computational study of large molecules such as proteins. In such cases, there are billions of potential conformations and full conformational searches have been (and probably always will be) impossible. For these situations, a range of stochastic searching algorithms has been developed. In addition, there are a limited number of simplified systematic algorithms, in which attempts have been made to defeat the combinatorial explosion problem by the limiting regions of the conformational space that are explored. In the following we briefly outline the main elements of these different approaches, with a view to identifying the most suitable course for obtaining low energy conformations with a sufficiently high reliability for chemically accurate studies.

Stochastic methods

In general terms, stochastic search methods explore conformational space through random or semi-random variations to the coordinates or torsional angles, usually subject to some constraints and with some element of bias. The search is stopped either when an arbitrary number of iterations is performed, or when the search fails to find new conformers over a certain number of iterations. Classic examples of stochastic search methods include the *cartesian coordinate stochastic search*[6] and the *internal coordinate Monte Carlo search*.[7] Since stochastic search methods that utilize the energies of the existing conformations to bias the sampling have been found to outperform purely random searches,[8,9] many stochastic methods have been developed to exploit this. These range from variants of the simple Monte Carlo search methods that incorporate biased sampling,[10] to methods that place random probes over a potential energy surface and then use the information to locate new probes in low energy areas,[11–14] methods that allow conformations to "walk" across the potential energy surface using adaptive grids,[15] and methods based on *genetic algorithms*.[16] In these latter methods, one starts with a random population of "chromosomes" which, in the case of a conformation search, would consist of a bit string of length N, where N is the number of rotatable bonds, and where each bit is the value of the torsional angle for the bond in question. Members of the population are then selected based on their "fitness" (*i.e.* some function of their relative energy), and various operations are then performed including random mutations and "crossover" (*i.e.* one part of one chromosome is matched with the complementary part of another). The analysis is then repeated for an arbitrary number of "generations" or until certain stopping criteria have been met (such as

the failure of the lowest energy to change over a specified number of generations).

Another important class of stochastic search method utilizes *molecular dynamics*. Starting from a random conformation, thermal energy is supplied and new geometries and velocities are derived by integration of Newton's laws over a small time step, thereby forming the starting point for the next iteration. Provided they are given sufficient thermal energy, molecules can "walk" over small energy barriers and move toward lower energy regions of the potential energy surface. At certain time intervals, the conformations of the molecules being simulated are collected and (usually) minimised; favourable trajectories are allowed to continue as long as they satisfy certain criteria. In *simulated annealing*[17,18] the conformations are accepted or rejected with a certain probability based on where they fit into a Boltzmann distribution. The analysis is initially performed at "high" temperatures (for which the high energy conformations have a reasonable probability of being accepted and energy barriers can thus be crossed) and then the system is slowly "cooled" so that only the low energy conformations are retained. The entire analysis is repeated at different initial seed conformations, to ensure greater coverage of the conformational space. A number of variants of molecular dynamics methods have been suggested, including versions in which the initial velocities are concentrated in rotational modes (rather than vibrations and bends) which are more relevant to conformational searching,[19] and *conformational space annealing*, in which a genetic algorithm is used to generate starting conformations for simulated annealing.[20] There are also a range of *quantum annealing* methods, in which quantum mechanics rather than classical mechanics is used to walk across the potential energy surface.[21–23]

Systematic methods

Strictly speaking, systematic methods are those that explore *all* conformational space at some fixed degree of resolution. Examples include methods (such as *tree search*,[1] *SUMM*[2,3] and *complementarity*[4]) that search *via* systematic variation of torsional angles as described above, and methods (such as *LMOD*[24] and *TORK*[25]) that exhaustively search conformational space using eigenvector following routines. Since full systematic searches of conformational space suffer from the combinatorial explosion problem that renders them impractical for the study of large molecules, a range of simplified systematic methods have been developed that explore only portions of the conformational space but in a deterministic manner. These methods are also referred to as "systematic" so as to distinguish themselves from stochastic methods.

The main approach to reducing the dimensionality of the conformational space is to perform systematic conformational searches on small portions of the molecule (either as isolated fragments or *in situ*), and then build the conformation of the whole molecule from the optimal parts with only limited additional searching of the relative conformations of the fragments. Methods that incorporate this principle, known as "build-up",[26–28] include chemometrics,[29] A*,[30] and sparse matrix drive.[31] It has also been said that genetic algorithms owe their success to the implicit inclusion of build-up (through the bias toward retention in the population of combinations of torsional angles that are "fit").[16] The extreme example of the build-up approach would be to perform a "linear search" of the conformational space, in which each torsional angle is optimised independently of the others resulting in a linear scaling (but potentially inaccurate) method.

Assessment

There have been numerous comparative studies in which the relative merits of various conformational search routines have been examined.[8,9,16,21,31–35] However, it is difficult to identify a definitive optimal method as the criteria by which the methods were assessed varies considerably amongst the studies, and no one study has compared all of the principal types of methods. Nonetheless, a few general observations may be made.

Amongst the stochastic methods studied, it has been found that those which use the existing results to direct the search outperform purely random methods,[8,9] methods based on genetic algorithms tend to outperform simulated annealing,[15,31] and the quantum annealing method, *quantum path minimization*, was shown in one study to outperform both simulated annealing and a genetic algorithm.[21] When stochastic methods are compared with (full) systematic methods, the stochastic search methods tend to be more efficient in identifying low energy conformations early in the search, but they rapidly lose this advantage as the search proceeds and are less efficient in finding the global minimum structure.[34] Indeed, systematic search methods, by their very nature, offer the most efficient means to cover all conformational space. In contrast, "no stochastic method has a probability of one to converge to the global minimum in a finite number of steps."[36] Moreover, it is clear that the stopping criteria needed to guarantee that a structure which lies within an acceptable level of energy of the global minimum are difficult to ascertain (without already knowing the answer) and vary considerably with the system under study. In summary, the stochastic methods would appear to be extremely useful for the applications for which they were designed—that is, identifying relatively low energy conformation(s) of large molecules with minimal computational expense. However, they do not appear to offer the degree of reliability desired for kcal-accurate thermochemical studies without sacrificing their computational efficiency.

Fortunately, the simplified systematic methods based on the build-up principle appear to be more promising. For example, a previous study compared the sparse matrix drive algorithm, a build-up method designed especially for protein side-chain optimisation, with both simulated annealing and various genetic algorithms.[31] They found that the sparse matrix drive algorithm not only yielded the most accurate results, it was also significantly more efficient. More generally, build-up based methods are attractive as a chemically intuitive approach to defeating the combinatorial explosion problem as, provided substituents are sufficiently separated from one another, it seems unlikely that their optimal conformations are affected by those of the other substituent. However, the separation point at which the conformational properties become independent of one another is often highly dependent on the chemistry of the system. It would be desirable to design a more general objective method for achieving this, and one that

avoids the cumbersome task of actually fragmenting a molecule and then later reassembling the optimised aggregates. In the present work, we present a new systematic algorithm that is inspired by the build-up approach but is performed on the original molecule as a whole, and determines the combinations of torsional angles to optimise using a learning process.

3. Algorithm description

In simple terms, the *energy-directed tree search* (EDTS) algorithm is a modified version of the *tree search* algorithm in which the combinatorial explosion problem is defeated through the examination of only certain "branches" (*i.e.* combinations of torsional angles) of the tree. It is based on the build-up principle, whereby it is assumed that the optimal conformations of certain parts of a molecule (usually remote parts) are unlikely to influence one another and can thus be searched independently, thereby reducing the dimensionality of the conformational space. However, the key aspect of our new algorithm EDTS is that the choice of which combinations of torsions are important is determined objectively, based on the energies of the existing conformations. In other words, it includes a learning process and does not rely upon the specific chemistry of the system and/or the chemical intuition of the user. It is also distinct from many of the existing build-up algorithms in that the torsional angles are optimised *in situ*—that is, the molecule is not fragmented into portions. This not only simplifies the analysis but also allows for the optimisation of combinations of physically remote torsions, should their energies appear to be coupled (due to, for example, through-space interactions).

A flowchart of the EDTS algorithm is provided in Fig. 1. In broad terms, the algorithm consists of one or more "linear" searches of the complete conformational space. In a linear search, geometry optimisations are performed for all values of the first rotation but with the others unchanged. The first

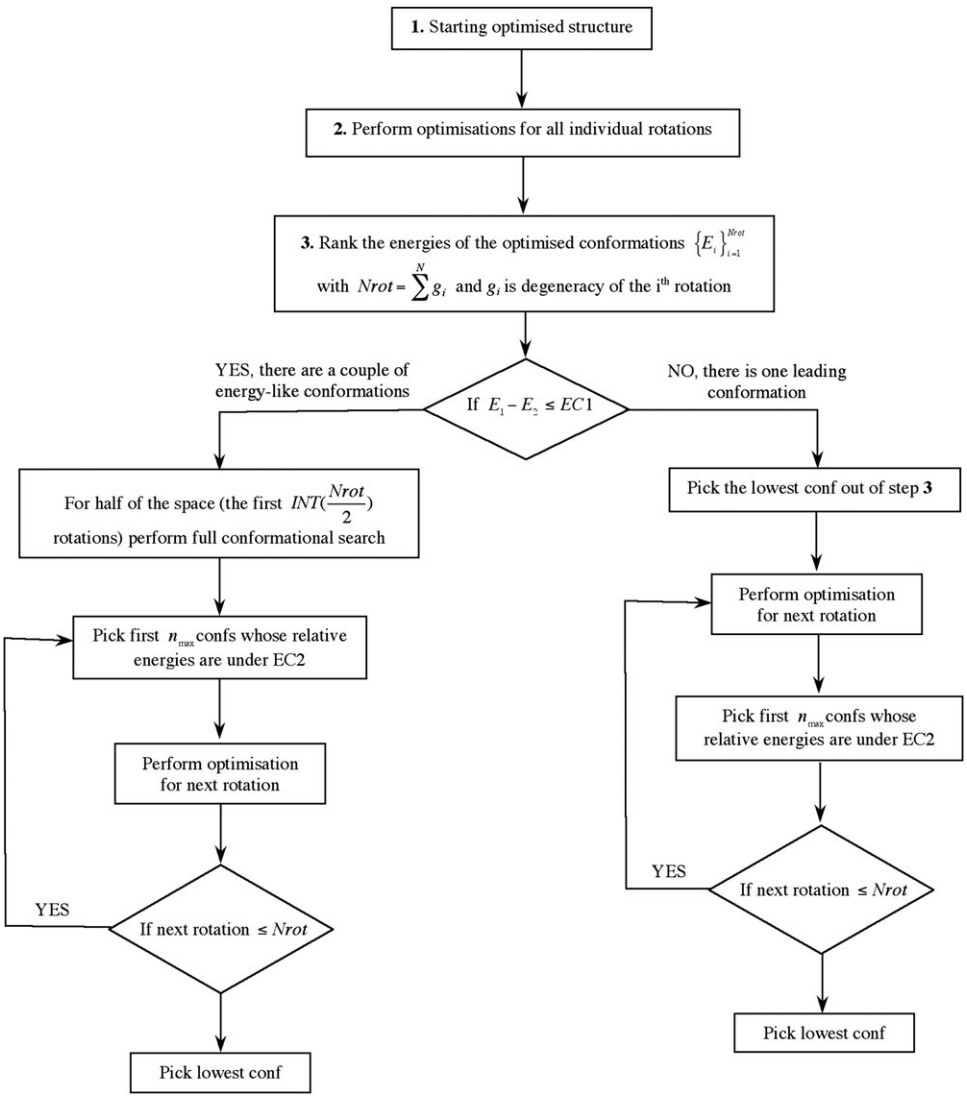


Fig. 1 The flowchart of the new algorithm. Based on our empirical studies, the values of EC1 and EC2 are set at 3 and 4 kJ mol^{-1}, respectively, and NMAX is set at 5 so as to ensure the reliability with which the global minimum can be found. In the flowchart g_i shows how many individual conformations can be generated rotating around the *i*th bond.

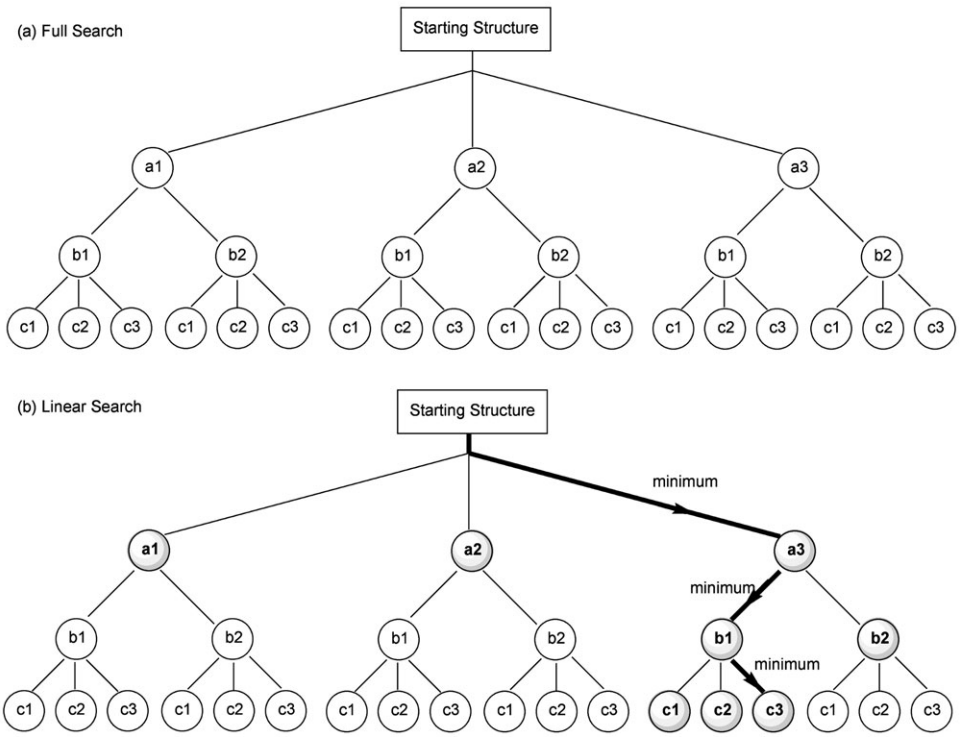


Fig. 2 Tree diagram (a) showing the combinations of rotations examined in a full conformational search for structure **1**, and (b) highlighting those that might be examined in a linear search of the same space.

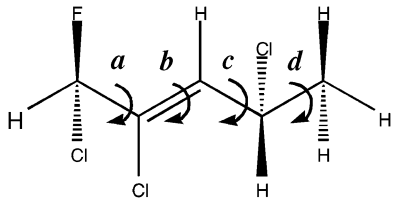
rotation is then set at its optimal value, and geometry optimisations are then performed for the second rotation, which is then also updated, and so on until all rotations are exhausted (see Fig. 2). During the course of our work, we have found that such linear searches of conformational space are sometimes capable of identifying the global minimum energy structure but their success is highly dependent upon the starting structure chosen, and the order in which the rotations are performed. This is seen quite clearly in Table 1, which shows the results of various linear searches for a number of the species in the present work. Our new algorithm addresses this problem by using an initial scan (and if necessary subsequent full searches on part of the conformational space) both to identify one or more improved starting structures for the subsequent linear search and to define the order in which the rotations are to be examined. We now explain the implementation of the algorithm *via* means of an example.

To implement the algorithm, one first chooses an arbitrary conformation of the molecule and optimises its geometry to the nearest local minimum. From this structure, one can then define the full conformational space as generated by the tree search method. For example, for structure in Scheme 1 there are 4 carbon–carbon bond rotations to consider, and at a conservative resolution of 60°, one would consider all

Table 1 B3LYP lowest energy conformation (C1), dihedral angle sequence used in the linear search algorithm, lowest-energy structure (C2) found by the algorithm with the predefined sequence and the energy difference (in kJ mol^{-1}) between the C1 and C2 structures[a]

Species	C1	Dih. angle sequence	C2	ΔE(C1–C2)
$S{=}C(CH_3)$–$SCH(COOCH_3)CH_2$–$CH(COOCH_3)$–CH_3	a1e2	f e d c b a	a1e2	0.0
1		a c f b d e	a2b3c3	3.0
		d e b c f a	a1b3d2e2	4.0
CH_3–$CH(COOCH_3)$–S–$C^{\bullet}(CH_3)$–S–$CH(COOCH_3)$–CH_3	a2b3c2	f e d c b a	a1b3c2f2	0.7
2		b a c d e f	a1b2c2f2	5.4
		c d e b a f	a1c2d2e2	7.4
CH_3–$CH(COOCH_3)$–CH_2–$C^{\bullet}H(COOCH_3)$	a1b2c2	a b c d	a2b3	1.6
4		d c b a	a1b2c2	0.0
		b a d c	a2b2	2.9
$P(CH_3)_2(CH_2)_4P^{\bullet}CH_3$	a3b3c3e3	a b c e d	a3b3c3e2	0.0
5		e d c b a	a2d2e3	5.3
		a e d b c	a3c3d2e2	6.1

[a] All structure numbers refer to Fig. S1 of the ESI,† wherein the dihedrals angles are also depicted.



Scheme 1

combinations of all 6-fold rotations of each bond, leaving 6^4 structures to be optimised.

In practice, one might consider only 3^3 structures corresponding to 3-fold rotations (a resolution of 120°) about bonds *a*–*c*, with rotation *d* omitted due to symmetry. As noted above, in a more general case the full conformation space will contain $(360°/R)^N$ conformations, where R is the resolution and N the number of rotatable bonds. If it is possible through symmetry or chemical knowledge to simplify the full conformational space further (for example, one might wish to use a resolution of 180° for the rotation of the double bond *b*) this is easily accommodated by the EDTS algorithm. To illustrate this point, we will assume for the present example that the full conformation space for molecule **1** is that illustrated in Fig. 2a, in which bonds *a* and *c* are considered at a 120° resolution and bond *b* is considered at a 180° resolution.

Having decided which rotations to include in a full search of the conformational space, the next step in the algorithm is to perform a linear search of this conformational space. However, unlike a true linear search (as in Fig. 2b), in this initial scan we do not update values of the torsional angles after each rotation. Rather, we generate starting structures having all possible values for rotation *a*, but with bonds *b*–*c* unchanged, we generate starting structures having all possible values for rotation *b*, but with bonds *a* and *c* unchanged, and then all rotations of bond *c* with *a*–*b* unchanged. This allows us to establish which rotations have the largest effect on the conformational energy. In the present example, if we designate the three rotations of bond *a* as *a*1, *a*2 and *a*3 (where a1 corresponds to the original value, and *a*2 and *a*3 to increments of 120° and 240°, respectively), the two rotations of bond *b* as *b*1 and *b*2, and the three of *c* as *c*1, *c*2 and *c*3, the structures to be optimised would consist of {**a1**b1c1}, {**a2**b1c1}, {**a3**b1c1}, {a1**b2**c1}, {a1b1**c2**} and {a1b1**c3**}. The results of the linear search are then ranked in order of increasing energy and the relative energies of the conformations are calculated. On this basis, one of two scenarios is identified: (i) more than one conformation lies within a certain tolerance level (EC1, which we have set at 3 kJ mol^{-1} based on the empirical studies below) of the minimum energy structure, or (ii) one conformation (which we term a "leading conformation") is significantly lower than the others by this tolerance level.

In scenario (i) we have no leading conformation and it is therefore necessary to choose our starting structures for the next linear search very carefully. To this end, a full conformational search is next performed, but only using the lower half of the conformational space. For example, supposing the ranking of the conformations from the initial linear search was {**a2**b1c1}, {a1**b2**c1}, {a1b1**c3**}, {**a1**b1c1}, {a1b1**c2**} and {**a3**b1c1}, we would examine all combinations of torsional angles a2, b2 and c3, with all other angles unchanged. The species to be optimised would thus be: {**a2b2**c1}, {**a2**b1**c3**}, {a1**b2c3**} and {**a2b2c3**}, which would be added to the existing pool of structures.

Having performed the full search on the lower half of the conformational space, the energies of all existing species are again ranked. On the basis of this ranking one selects either all species within the tolerance level (EC2, set at 4 kJ mol^{-1}) of the lowest energy structure, or (if there are too many) the lowest NMAX species (where NMAX is set at 5 based on our empirical studies). Using these selected species as starting structures, one then performs a linear search of the remaining conformational space, but this time rotations are considered in the order in which they appeared in the initial energy ranking. Moreover, after each individual rotation is performed, all species are ranked and, if necessary, the starting structures are updated. It should be noted that, since a full search has already been performed on the first half of these rotations, the first rotation in the new linear search is actually lowest species of the top half of the rotations. The search terminates once the last rotation is completed.

In the present example, the original order was {**a2**b1c1}, {a1**b2**c1}, {a1b1**c3**}, {**a1**b1c1}, {a1b1**c2**} and {**a3**b1c1} and a full search was performed on a2, b2 and c3; hence the linear search is performed by applying a1, c2 and a3 (in that order) to the new lowest energy structure(s). Supposing after the full search on a2, b2 and c3, the ranking was {**a2**b1c1}, {a1**b2**c1}, {**a2b2**c1}, {**a2**b1**c3**}, {a1**b2c3**}, {**a2b2c3**}, {a1b1**c3**}, {**a1**b1c1}, {a1b1**c2**} and {**a3**b1c1}, and supposing only {a1**b2**c1} lay within 4 kJ mol^{-1} of the lowest energy structure, {**a2**b1c1}, one would perform a linear search on these two. The first structures would be generated from a1 applied to {a1**b2**c1} and {**a2**b1c1}; that is, {**a1b2**c1} and {**a1**b1c1} (which we already have from the original linear search). The next two would be from c2 applied to {a1**b2**c1} and {**a2**b1c1}; that is, {a1**b2c2**} and {**a2**b1**c2**}, and so on until there are no further rotations to examine.

In scenario (ii) there is a leading conformation and it is used as the single starting structure for the subsequent linear search. In other words, the full search on the lowest half of the conformational space is omitted entirely. As in scenario (i) the linear search is again performed on the remaining rotations in their ranked order, and using the lowest energy species as a starting structure. After each individual rotation is performed, the energies of all species are ranked and the starting structure(s) is updated if necessary. In our current example, the initial linear search yielded the order {**a2**b1c1}, {a1**b2**c1}, {a1b1**c3**}, {**a1**b1c1}, {a1b1**c2**} and {**a3**b1c1}. Supposing {**a2**b1c1} is significantly lower than the others, we take this structure and apply rotation b2 to give {**a2b2**c1}. If this structure lies within 4 kJ mol^{-1} of the minimum, we then continue our linear search on both {**a2b2**c1} and {**a2**b1c1}. We next apply rotation c3 to give {**a2b2c3**} and {**a2**b1**c3**}, and so on. The search again terminates once all rotations are considered. The fully worked examples of the algorithm for both scenarios are given in Appendix S1 of the Supporting Information.

To implement EDTS algorithm computationally one needs only a means of systematically generating the starting structures for a full tree search method. One then selectively runs batches of geometry optimisations as dictated by the

algorithm, using any standard computational chemistry software at a level of theory appropriate for the system at hand. After each step in the process, a simple spreadsheet can be used to rank the energies of the conformations, thereby enabling the user to determine the next batch of optimisations to run. In principle, the entire process could be implemented computationally; however, in our experience the interactive approach is relatively straightforward and leads to the most efficient use of computer time.

4. Algorithm testing

Optimisation of parameters

In implementing the EDTS algorithm, one needs to set values for the parameters (labelled EC1, EC2 and NMAX in Fig. 2) that govern how many starting conformations are considered in the search. The first of these, the energy cut-off value EC1, determines whether there is a clear leading conformation, or whether a full search needs to be performed on the lower half of the conformational space to identify starting structure(s) for the subsequent linear search. The second closely related value, EC2, is applied after each step in the subsequent linear search to determine which starting structures are included in the next step. The third parameter NMAX is used to limit the number of starting structures in the event that the EC2 parameter leads to the inclusion of an impractically large number of starting structures. If EC1 and EC2 are vanishingly small, and NMAX is set at 1, the EDTS algorithm collapses to a linear search; if EC1 and EC2 are extremely large and NMAX is set at the total number of available conformations, the EDTS algorithm expands to the full tree search algorithm. To be successful, intermediate values for EC1, EC2 and NMAX should be chosen so that the number of conformations to be optimised is limited as much as possible but without compromising the accuracy of the final result. To this end, we initially performed the EDTS algorithm on 4 selected structures from our test set, using various values for the three parameters. The results for one typical case (structure *2*) are shown in Table 2. The results for other 3 species (structures 1, 3 and 4) are given in Tables S1–S3 of the ESI.†

From Table 2, it is seen that, in all cases the algorithm yielded either the lowest energy conformation or, in the worst case, the second lowest conformation (which was only 0.7 kJ mol^{-1} above the global minimum). Thus, at least for the values examined, the accuracy of the algorithm appears to be relatively insensitive to values of the EC1, EC2 and NMAX. Based on the data in Table 2, the minimal parameters needed to obtain the global minimum are EC1 = 3 kJ mol^{-1}, EC2 = 4 kJ mol^{-1} and NMAX = 5, and these have been adopted as our optimal parameters. More generally, we have found that if there is one leading conformation then it is usually lower in energy than the others by more than 10 kJ mol^{-1}. In that scenario, following the right-hand side on the algorithm flow chart in Fig. 1 only makes minor refinements to this conformation and in that way ensures that the lowest-energy conformation has not been missed. When there are, instead, a few conformations with energies within a couple kJ mol^{-1} of the lowest energy structure, optimisation of all of the conformations within the lower half of the conformational space produces the final leading conformation. In all cases studied, the rotations that appear in this half of the conformational space are always present in the lowest-energy structure. In most cases, the search may be stopped there but the algorithm checks the rest of the rotations to eliminate the possibility of missing out on the lowest-energy one. The NMAX criterion is usually set to 5, as including more structures in the search will only expand the conformational space and one would like to keep this as small as possible. By doing this, there is a danger that if there are multiple low energy structures within the tolerance level, the global minimum structure may be missed. However, this disadvantage is not relevant in thermochemical studies, since it is the energy of the conformation used that is important. If the final conformational energy is very close to that of the global minimum then accurate thermodynamical quantities can still be obtained. The results for other 3 species (structures 1, 3 and 4) confirmed the selected values for the EC1, EC2 and NMAX criteria: EC1 = 3 kJ mol^{-1}, EC2 = 4 kJ mol^{-1} and NMAX = 5.

Table 2 Performance of the EDTS algorithm for $CH_3CH(COOCH_3)SC^{\bullet}(CH_3)SCH(COOCH_3)CH_3$ (*2*) for various values of EC1, EC2 and NMAX[a]

NMAX	EC1	EC2	Lowest conf	Nalg
5	**3.0**	3.0	Conf 2	23 (16%)
		4.0	**Conf 1**	**37 (26%)**
		5.0	Conf 1	37 (26%)
	4.0	3.0	Conf 1	25 (17%)
		4.0	Conf 1	37 (26%)
		5.0	Conf 1	37 (26%)
	5.0	3.0	Conf 1	29 (20%)
		4.0	Conf 1	29 (20%)
		5.0	Conf 1	37 (26%)
8	3.0	3.0	Conf 2	23 (16%)
		4.0	Conf 1	43 (30%)
		5.0	Conf 1	43 (30%)
	4.0	3.0	Conf 1	26 (18%)
		4.0	Conf 1	43 (30%)
		5.0	Conf 1	43 (30%)
	5.0	3.0	Conf 1	30 (21%)
		4.0	Conf 1	43 (30%)
		5.0	Conf 1	43 (30%)

[a] In this table, conf 1 is the global minimum and conf 2 is the second lowest conformation which lies 0.7 kJ mol^{-1} above the global minimum. Nalg is the number of conformers optimized during the search and the number in parentheses is the percentage of the full conformation space that this corresponds to. All structure numbers refer to Fig. S1 of the ESI.† The optimal values are shown in bold.

Test set and computational details

To evaluate the EDTS algorithm we firstly selected a series of 14 molecules as the test set-1, including as a reference the 4 species (*1–4*) for which the parameters (EC1, EC2 and NMAX) were optimised. The species were drawn from our own studies of polymerisation processes and include open- and closed-shell species, stable structures and transition structures, and neutral and charged species. A range of functional groups, including phenyl rings, esters, thioesters and phosphines are represented in the test set. In order to check that the size of the conformational space does not represent a limitation to the

efficiency of the algorithm, we also selected a test set-2 that includes 8 species such as polymers, peptides, drugs, and natural products, which are significantly larger in size than the test set-1 and contain from 8 to 11 rotatable bonds. Most of the species selected have polar terminal groups that are capable of undergoing through-space interactions, situations that normally present difficulties for build-up based algorithms. The global minimum energy conformations of each species are illustrated in Fig. S1 and S2 of the ESI,† and corresponding coordinates for both the starting conformations and global minima are also provided.

For each molecule a full conformational search was performed using a tree search algorithm. For most species in the test set-1, geometry optimisations were performed at the B3-LYP/6-31G(d) level of theory; however for the anionic species (*11*) B3-LYP/6-31 + G(d) was used, while for adduct radicals of xanthates (*8* and *9*) HF/6-31G(d) was used. All the species in the test set-2 were optimised at the AM1 semi-empirical level of theory to reduce the computational cost for performing the full conformational space. The resolution was typically set at 120°; however, bonds involving sp^2 centres were scanned at 180° and rotations that led to symmetry equivalent structures were omitted from the analysis. Where relevant, we also further simplified the full conformational space by making the assumption that the ester linkages are always *cis* and not *trans*. In some cases we also simplified the conformational space by omitting some of the rotations (such as some of the phenyl rotations). The EDTS algorithm as described above was applied to the same conformational space at the same level of theory so that a consistent comparison could be made. Various linear searches, in which the same full conformational space was selectively sampled according to the algorithm illustrated in Fig. 2b, were also carried out for selected structures (Table 1), and these were also performed at the same level of theory. The torsional angles that were varied during the full and linear searches are labelled in Fig. S1 and S2 of the ESI.†

All geometry optimisations were performed in Gaussian 03.[37] All algorithms were implemented interactively using a simple Fortran program to generate the starting structures for the full conformational search (from which starting structures for the simplified searches were selected), and spreadsheets were used to the sort energies and organise results at each stage of the search.

Performance of EDTS

Having found optimal values of EC1, EC2 and NMAX parameters, we next implemented both the EDTS algorithm and the full tree search algorithm for our full test of 22 species. In each case we compared the lowest energy conformation obtained in each search and the number of conformations optimised during search (see Tables 3 and 4). In most of the cases, the EDTS algorithm successfully found the correct global minimum conformation. For the rest of the species the EDTS algorithm was able to find conformations whose total electronic energies are within chemical accuracy (1 kcal mol^{-1}) from the true global minima. This would still allow for accurate calculations of kinetics and thermochemistry.

The number of conformations requiring optimisation in the EDTS algorithm depends not only on the size of the conformational space, but also upon whether or not a leading conformation was found at the first step of the search, and upon the number of starting structures that are retained at each step of the subsequent linear search. This in turn depends upon both the chemistry of the system (for example, how strongly the arrangement of one substituent affects that of another) and how close the specific starting structure chosen for the search was to the final global minimum. As a result, there was a wide variation in both the number of optimisations required by the algorithm and the fraction of the full conformational space that was explored. Nonetheless, in all cases the EDTS algorithm substantially reduced the number of conformations requiring optimisation when compared with a

Table 3 Performance of the EDTS algorithm on the test set-1 using the optimal values of EC1, EC2 and NMAX[a]

Species	Ntot	Nalg	Species	Ntot	Nalg
CH_3–CH($COOCH_3$)–CH_2–C•H($COOCH_3$)	36	21 (58%)	TS[CH_3CH($COOCH_3$)CH_2CH($COOCH_3$)• + CH_2=CH$COOCH_3$]	54	34 (63%)
4			**7**		
TS[CH_3CH($COOCH_2CH_3$)CH_2CH($COOCH_2CH_3$)• + CH_2=CH$COOCH_2CH_3$]	54	27 (50%)	CH_3–CH($COOCH_3$)–S–C•(CH_3)–S–CH_2CN	72	27 (38%)
6			**3**		
CH_3O–C(=O)–CH_2–SC•(OC_2H_5)–SCH_3	144	28 (16%)	P(CH_3)$_2$(CH_2)$_3$P•CH_3	81	32 (40%)
8			**10**		
C•H(Ph)–CH_2–CH(Ph)–CH_2–CH(CH_3)Ph	81	33 (41%)	CH_3O–C(=O)–CH_2–SC•($OC(CH_3)_3$)–SCH_3	96	29 (30%)
14			**9**		
CN$(CH_3)_2$S–C•(Ph)–SCH(Ph)–CH_2–CH(Ph)–CH_2–C$(CH_3)_2$CN	108	27 (25%)	CH_3–CH($COOCH_3$)–S–C•(CH_3)–S–CH($COOCH_3$)–CH_3	144	25 (17%)
13			**2**		
S=C(CH_3)–SCH($COOCH_3$)CH_2–CH($COOCH_3$)–CH_3	216	17 (8%)	P(CH_3)$_2$(CH_2)$_4$P•CH_3	243	45 (19%)
1			**5**		
S=C(Ph)–SCH(Ph)–CH_2–CH(Ph)–CH_2–C$(CH_3)_2$Ph	243	50 (21%)	P(CH_3)$_2$(CH_2)$_4$P$^-$(CH_3)	243	40 (17%)
12			**11**		

[a] In this table, Ntot is the total number of conformations to be explored in the complete conformation space and Nalg is the number of conformations needed for the new algorithm to find the global minimum. The percentage of the conformation space explored is given in parentheses. All structure numbers refer to Fig. S1 of the ESI.† The optimal values of EC1, EC2 and NMAX are 3 kJ mol^{-1}, 4 kJ mol^{-1} and 5, respectively.

Table 4 Performance of the EDTS algorithm on the test set-2 using the optimal values of EC1, EC2 and NMAX[a]

Species	Ntot	Nalg	ΔE (kJ mol^{-1}) global-local (ranking in bracket)
$(CH_3)_2C(CN)(CH_2C(CH_3)COOCH_3)_2SC^{\bullet}(Ph)SC(CH_3)_2CN$ **15**	26 244	82 (0.3%)	2.98 (4)
$H(CH_2C(CH_3)COOCH_3)_3SC^{\bullet}(Ph)SC(CH_3)_3$ **16**	78 732	85 (0.1%)	2.55 (10)
His–arg **17**	78 732	103 (0.1%)	3.72 (12)
Thr–lys **18**	39 366	244 (0.6%)	0.00 (1)
Oleocanthal **19**	78 732	491 (0.6%)	0.33 (41)
Pantothenic acid **20**	17 496	195 (1.1%)	0.00 (1)
Taxol **21**	17 496	58 (0.3%)	0.00 (1)
Viagra **22**	972	70 (7.2%)	0.00 (1)

[a] In this table, Ntot is the total number of conformations to be explored in the complete conformation space and Nalg is the number of conformations needed for the new algorithm to find the global minimum. The percentage of the conformation space explored is given in parentheses. All structure numbers refer to Fig. S2 of the ESI.† The optimal values of EC1, EC2 and NMAX are 3, 4 and 5 kJ mol^{-1}, respectively.

full search, and this fraction tended to decrease as the conformational space increased. For example, for 7 molecules having 100 to 1000 conformations in their full space, the fraction of conformational space explored ranged from 7–25%; for the other species having their full space in the range of 10 000 to 80 000 conformations the fraction reduces to only 0.1–1.1%.

The efficiency of the EDTS algorithm would be expected to increase further as the conformational space expands, as the method (in its worst case scenario) effectively scales as 2^N conformations for a resolution of 120°, compared with 3^N for the corresponding systematic search. That is, the efficiency scales as $(2/3)^N$ with the number of rotatable bonds in the corresponding full search. (A derivation of this result is supplied in Appendix S2 of the ESI.†) For example, for a full conformational space of 3^{10} structures (10 bonds at 120° resolution), the corresponding worst-case EDTS algorithm (in which a linear search on all of the space is followed by full search on half of the space, and then 5 additional linear searches on the rest of the space) would require less than 1083 starting structures—less than 2% of the full conformational space. Importantly, such searches, whilst expensive, are nonetheless practical whereas the full searches (of nearly 60 000 conformations) are not. Since the algorithm maintains the reliability of a full search, it does effectively expand the range of systems for which accurate thermochemistry can be studied.

Of course, this accuracy comes at a cost, and the algorithm remains too expensive for larger systems. For example, for a species with 20 rotatable bonds (for which the full conformational space at a resolution of 120° would contain 3^{20} structures), less than 0.03% of the conformational space would be explored in a worst-case EDTS algorithm; however, this would nonetheless translate to around a million starting structures. The EDTS algorithm is therefore not currently competitive as an alternative to the approximate stochastic algorithms for these larger systems, but rather is intended to bridge the gap between the moderately large systems for which accurate thermochemistry is now possible and the size limitations imposed by the practicality of full conformational searches.

5. Conclusions

In the present work, we have introduced a new systematic algorithm, energy-directed tree search (EDTS) for exploring the conformational space of molecules. The algorithm has been designed to reliably locate the global minimum at a fraction of the cost of a full conformational search, and in this way extend the range of chemical systems for which accurate thermochemistry can be studied. The algorithm was tested for a set of 22 large molecules, including open- and closed-shell species, stable structures and transition structures, and neutral and charged species, incorporating a range of functional groups (such as phenyl rings, esters, thioesters and phosphines) and covering polymers, peptides, drugs, and natural products. In most of the cases the global minimum energy structure was obtained at a substantially reduced computational cost when compared with the corresponding full conformational search. For the rest of the species the EDTS algorithm was able to find conformations whose total electronic energies are within chemical accuracy from the true global minima. This would still allow for accurate calculations of kinetics and thermochemistry.

Acknowledgements

We gratefully acknowledge financial support from the Australian Research Council and generous allocations of computer time from the Australian Partnership for Advanced Computing and the Australian National University Supercomputer Facility.

References

1 M. Lipton and W. C. Still, *J. Comput. Chem.*, 1988, **9**, 343–355.
2 J. M. Goodman and W. C. Still, *J. Comput. Chem.*, 1991, **12**, 1110–1117.
3 I. Kolossvary and W. C. Guida, *J. Comput. Chem.*, 1993, **14**, 691–698.
4 C.-S. Wang, *J. Comput. Chem.*, 1997, **18**, 277–289.

5 M. L. Coote, E. I. Izgorodina, E. H. Krenske, M. Busch and C. Barner-Kowollik, *Macromol. Rapid Commun.*, 2006, **27**, 1015–1022.
6 M. Saunders, *J. Am. Chem. Soc.*, 1987, **109**, 3150–3152.
7 G. Chang, W. E. Guida and W. C. Still, *J. Am. Chem. Soc.*, 1989, **111**, 4379.
8 R. S. Judson, M. E. Colvin, J. C. Meza, A. Huffer and D. Gutierrez, *Int. J. Quantum Chem.*, 1992, **44**, 277–290.
9 K. W. Foreman, A. T. Phillips, J. B. Rosen and K. A. Dill, *J. Comput. Chem.*, 1999, **20**, 1527–1532.
10 S. B. Ozkan and H. Meirovitch, *J. Phys. Chem. B*, 2003, **107**, 9128–9131.
11 K. A. Dill, A. T. Phillips and J. B. Rosen, *J. Comput. Biol.*, 1997, **4**, 227.
12 D. Cvijovic and J. Klinowski, *Science*, 1995, **267**, 664.
13 A. F. Stanton, R. E. Bleil and S. Kais, *J. Comput. Chem.*, 1997, **18**, 594–599.
14 H. Senderowitz and W. C. Still, *J. Comput. Chem.*, 1998, **19**, 1736–1745.
15 J. C. Meza and M. L. Martinez, *J. Comput. Chem.*, 1994, **15**, 627–632.
16 R. S. Judson, E. P. Jaeger, A. M. Tresurywala and M. L. Peterson, *J. Comput. Chem.*, 1993, **14**, 1407–1414.
17 S. Kirkpatrick, C. D. Gelatt, Jr and M. P. Vecchi, *Science*, 1983, **220**, 671.
18 S. R. Wilson, W. Cui, J. W. Moskowitz and K. E. Schmidt, *Tetrahedron Lett.*, 1988, 4343.
19 S. D. Morley, D. E. Jackson, M. R. Saunders and G. G. Vinter, *J. Comput. Chem.*, 1992, **13**, 693–703.
20 J. Lee, H. A. Scheraga and S. Rackovsky, *J. Comput. Chem.*, 1997, **18**, 1222–1232.
21 P. Liu and B. J. Berne, *J. Chem. Phys.*, 2003, **118**, 2999–3005.
22 P. Amara, D. Hsu and J. Straub, *J. Phys. Chem.*, 1993, **97**, 6715.
23 A. Finnila, M. Gomez, C. Sebenik, C. Stenson and J. Doll, *Chem. Phys. Lett.*, 1994, **219**, 343.
24 I. Kolossvary and W. C. Guida, *J. Am. Chem. Soc.*, 1996, **118**, 5011–5019.
25 C.-e. Chang and M. K. Gilson, *J. Comput. Chem.*, 2003, **24**, 1987–1998.
26 K. D. Gibson and H. A. Scheraga, *J. Comput. Chem.*, 1987, **8**, 826.
27 M. R. Pincus, R. D. Klausner and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 1982, **79**, 5107.
28 B. E. Hingerty, S. Figueroa, T. L. Hayden and S. Broyde, *Biopolymers*, 1989, **28**, 1195.
29 A. T. Bruni, V. B. P. Leite and M. M. C. Ferreira, *J. Comput. Chem.*, 2002, **23**, 222–236.
30 A. R. Leach and K. Prout, *J. Comput. Chem.*, 1990, **11**, 1193–1205.
31 P. Tuffery, C. Etchebest, S. Hazout and R. Lavery, *J. Comput. Chem.*, 1993, **14**, 790–798.
32 M. Saunders, K. N. Houk, Y.-D. Wu, W. C. Still, M. Lipton, G. Chang and W. C. Guida, *J. Am. Chem. Soc.*, 1990, **112**, 1419–1427.
33 A. K. Ghose, E. P. Jaeger, P. J. Kowalczyk, M. L. Peterson and A. M. Treasurywala, *J. Comput. Chem.*, 1993, **14**, 1050–1065.
34 A. M. Treasurywala, E. P. Jaeger and M. L. Peterson, *J. Comput. Chem.*, 1996, **17**, 1171–1182.
35 J. C. Meza, R. S. Judson, T. R. Faulkner and A. M. Treasurywala, *J. Comput. Chem.*, 1996, **17**, 1142–1151.
36 P. Serra, A. F. Stanton, S. Kais and R. E. Bleil, *J. Chem. Phys.*, 1997, **106**, 7170–7177.
37 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez and J. A. Pople, *GAUSSIAN 03, (Revision B.03)*, Gaussian Inc., Pittsburgh PA, 2003.

