WIREs
COMPUTATIONAL MOLECULAR SCIENCE
WILEY

The SMFA program for quantum chemistry calculations on large molecules

Journal:	*WIREs Computational Molecular Science*
Manuscript ID	CMS-729.R1
Wiley - Manuscript type:	Software Focus
Date Submitted by the Author:	n/a
Complete List of Authors:	Kobayashi, Rika; Australian National University, ANU Supercomputer Facility Addicoat, Matthew; Australian National University, Research School of Chemistry Gilbert, Andrew; Australian National University Amos, Roger; Australian National University Collins, MIchael; Australian National University, Research School of Chemistry
Keywords:	quantum chemistry, macro molecules
Choose 1-3 topics to categorize your article:	Ab Initio Electronic Structure Methods (CCAB) < Electronic Structure Theory (CCAA), Quantum Chemistry (CGAB) < Software (CGAA), Computational Biochemistry and Biophysics (CAAC) < Structure and Mechanism (CAAA)

Article Title: The SMFA program for quantum chemistry calculations on large molecules

Article Type: Software Focus

Authors:

Rika Kobayashi
Australian National University, Leonard Huxley Bldg 56, Mills Road, Canberra, ACT 2601, Australia. Email: Rika.Kobayashi@anu.edu.au, ORCID: 0000-0002-0672-833X

Matthew A. Addicoat
Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. Email: matthew.addicoat@ntu.ac.uk

Andrew T. B. Gilbert
Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. Email: andrew.iqmol@gmail.com, ORCID: 0000-0002-8211-9261

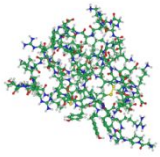
Roger D. Amos
Australian National University, Leonard Huxley Bldg 56, Mills Road, Canberra, ACT 2601, Australia. Email: rogerdamos@gmail.com, ORCID: 0000-0003-2868-7994

Michael A. Collins*
Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. Email: collins@rsc.anu.edu.au, ORCID: 0000-0003-1920-5105

Abstract

SMFA is a general program package for performing quantum chemistry calculations on large molecules, using an energy-based fragmentation approach. The program can calculate electronic energies, energy gradients and second derivatives; perform geometry optimization; find first order saddle points (transition states); perform energy optimized scans along a user-defined path; and evaluate various molecular properties. The program can use any of the following quantum chemistry packages: GAMESS(US), GAUSSIAN, NWChem and Q-Chem. In addition, SMFA provides a number of utility programs that, *inter alia*, calculate vibrational frequencies and infrared spectra with isotopic substitutions, the electrostatic potential on the solvent-accessible-surface, and isodesmic and higher order near-iso-energetic reaction schemes. Calculations of the electronic energy and related properties can be carried out using a scheme that provides a computation time that is linearly dependent on the size of the molecule or, if the user has enough processing units available, in a walltime that is independent of the size of the molecule.

Graphical/Visual Abstract and Caption



Graphics Caption: SMFA enables quantum chemistry calculations for proteins

Introduction

The field generally known as quantum chemistry is concerned with the calculation of the electronic structure of molecules. Over many decades, many computational methods have been developed to approximate the total energy of the electrons in a molecule, and associated molecular properties. With this knowledge, one can explain known chemical processes and predict previously unknown processes. Quantum chemistry has proven to be a powerful tool in chemistry generally. However, these computational methods are difficult to apply to large molecules, as the computational time (and memory) scale up rapidly with molecular size. If "N" represents some measure of molecular size (say, the total number of basis functions used to describe the electronic wavefunction), a widely used but not very reliable method (density functional theory) requires a computation time that scales as $O(N^3)$ - $O(N^4)$, while the most commonly used reliable method (a "coupled cluster" method) scales as $O(N^7)$. Substantial effort has been made to reduce these "scaling powers" to near linear scaling $O(N^1)$ through the use of highly parallelized codes. However, computation times remain very high. Consequently, the most reliable methods can only be applied to small molecules (containing perhaps tens of atoms), and even the less reliable methods cannot easily be applied to molecules containing hundreds or thousands of atoms.

Over the last decade, so-called fragmentation methods have been developed which aim to avoid the high power scaling problem of traditional quantum chemistry methods. [1 2]

SMFA (Systematic Molecular Fragmentation by Annihilation) is a general purpose program package which aims to approximate quantum chemistry calculations of molecular energies and properties for large organic, inorganic and biological molecules. The approximation employed, systematic molecular fragmentation by annihilation, [2 - 10] decomposes a molecule into relative small pieces (fragments), on which quantum chemistry calculations are carried out.

The number of fragments is linearly proportional to the number of chemical functional groups in the molecule. The size of the fragments is independent of the size of the whole molecule. The energy and properties of the molecule can be accurately approximated from a combination, respectively, of the energy and properties of these small fragments. Hence, if calculations on these fragments are carried out sequentially, the computation time is linearly proportional to the number of chemical functional groups in the molecule $O(N^1)$. However, these fragment calculations are completely independent of one another, hence if sufficient processing units are available, the "walltime" for any calculation only depends on the time required to perform a calculation on the largest fragment, which is independent of the total number of chemical functional groups $O(N^0)$. Thus, the SMFA approach provides a route to achieving high levels of parallelization that is independent of the level of theory used, therefore avoiding the implementation overhead of parallelizing each method separately.

THE SMFA METHOD

SMFA is based on a completely automated procedure for decomposing a molecule into small fragments. The method has been described in detail elsewhere. [6 9 10] To summarise, the computer algorithm can be briefly described as follows.

(1) Bonds and functional groups are defined, consistent with normal chemical concepts. A molecule is viewed as a collection of groups connected by single bonds.

(2) Beginning with some arbitrary group (call it A):

(i) Remove group A from the molecule;

(ii) leave A in place, but remove all groups separated from A by more than a specified number of bonds (call it "Level" bonds); and

(iii) remove group A and all the groups in (ii) from the molecule.

(iv) The molecule is represented by the sum of the fragments produced in steps (i) and (ii),

minus the fragments produced in step (iii).

(v) Taking each fragment in (iv), repeat steps (i) to (iii), and continue with every new fragment created until there is no fragment that has groups separated by more than Level bonds.

This procedure yields a set of fragments $\{F_n\}$ (N_{frag} in number) with integer coefficients $\{f_n\}$. Hydrogen atom caps are appended to the atoms with dangling bonds, with bond lengths calculated using the covalent radii of the atoms involved. In this way, the molecule, *M*, is represented by a sum of fragments, and the "bonded" energy, E_b , is a sum of the energies of the fragments.

$$M \rightarrow \sum_{n=1}^{N_{frag}} f_n F_n \qquad (1)$$

$$E_b = \sum_{n=1}^{N_{frag}} f_n E\left(F_n\right) \qquad (2)$$

Eq. (2) neglects the interactions between groups that are separated by more than Level bonds. These "nonbonded" interactions are accounted for as follows. First, a Level = 1 fragmentation is evaluated (comprising $N^{(1)}_{frag}$ fragments), and the nonbonded interactions, E_{nb}, are given by the interactions of these Level = 1 fragments:

$$E_{nb} = \sum_{n_1=1}^{N^{(1)}_{frag}-1} \sum_{n_2=n_1+1}^{N^{(1)}_{frag}} f^{(1)}_{n_1} f^{(1)}_{n_2} E\left[F^{(1)}_{n_1} \leftrightarrow F^{(1)}_{n_2}\right]_{allowed} \qquad (3)$$

where $E\left[F^{(1)}_{n_1} \leftrightarrow F^{(1)}_{n_2}\right]_{allowed}$ denotes the interaction energy of Level = 1 fragments that do not include group interactions that have already been accounted for in the bonded energy of Eq. (2). These fragment–fragment interaction energies are evaluated using ab initio calculations if the fragment–fragment separation is less than some specified tolerance, denoted by d_{tol} (the nearest atom-atom distance divided by the sum of the corresponding Van der Waals radii). For

longer separations, the interaction energy is evaluated using perturbation theory [based on ab initio calculations of charge distributions, static polarizability and the imaginary frequency polarizability] as the sum of an electrostatic interaction, a dispersion interaction and a contribution from induction. When formally charged groups are present in the molecule, or a polar solvent is present, the significant induction effect is accounted for by carrying out all ab initio calculations in the presence of appropriate embedded charges.[9]

This is a completely automated procedure in the SMFA program. If the user selects larger and larger values of the integer parameter Level, the fragments, F_n, are larger, and the total energy [Eq. (2) plus Eq. (3) plus the perturbative contributions] can be seen to converge (to near the correct value for the ab initio method and basis set which the user has chosen). Accurate energy gradients and higher derivatives are obtained by differentiating Eqs (1) and (2).

Similarly, since any property of a molecule can be evaluated as the derivative of the molecular Hamiltonian in the presence of an external field, equations similar to Eqs (1) and (2) have been derived for any molecular property. The method for obtaining such properties from the SMFA program is described below (and in detail in the User's Guide that accompanies the program).

The SMFA method shares some similar features and some significant differences from other "energy-based" molecular fragmentation methods. Several of these methods have been reviewed and compared in the review article of Ref. 2. In particular, SMFA tends to produce a similarly accurate approximation to the molecular energy as other approaches but with somewhat smaller fragment sizes. This is probably due to the fact that the fragmentation method relies primarily on the bonded connectivity of a molecule, rather than a distance-based algorithm.

It is important to note that SMFA (like the other approaches reviewed in Ref. 2) is

essentially an ab initio quantum chemistry method, rather than a "QM/MM" method. There is no distinction in SMFA between an "inner" part of the molecule that is treated by quantum electronic structure methods and an "outer" region which is treated using molecular mechanics. Hence there is no arbitrary or intuitive "user-defined" region that is the only part of the molecule that is treated correctly using first principles methods. The whole molecule (including any solvent) is treated from first principles. The essential approximation made is that "chemistry is local": That is, the influence that some part of the molecule has on the electronic structure of another part of the molecule decays in magnitude with the distance and bonded connectivity of the two regions.

A major advantage of SMFA is that it can be applied systematically. The essential approximation can be tested simply by repeating the application of SMFA with increasing values of the parameter Level. If the SMFA approximation is valid, then the values of the energy or property obtained should converge as the value of Level increases.

ACCURACY of SMFA

There have been several reports on the accuracy of SMFA in various applications. [4 - 7, 9] To illustrate these results, we note that for a set of 96 neutral organic molecules containing 18 to 80 atoms, the mean absolute error in the energy (MAEE) for a range of *ab initio* methods and basis sets was 2.1 to 4.4 kJ mol^{-1}, for *Level* = 3, and d_{tol} = 1.1.[5] A set of larger molecules, containing 45 to 180 atoms, was considered in Ref. 9: 24 peptides and sugars (with extensive hydrogen bonding) showed an MAEE of 14.9, 5.6 and 3.9 kJ mol^{-1}, for Levels 3, 4, and 5, respectively; 14 ions and zwitterions had an MAEE of 8.2, 4.5 and 2.9 kJ mol^{-1}, for Levels 3, 4, and 5, respectively; 20 conformers of a protein containing 246 atoms (with 3 formally charged groups) had a mean absolute error in their relative energies of 4.6 kJ mol^{-1} at Level 3. Moreover,

for these 20 conformers, the root-mean-square error in the energy gradient falls below about 0.0003 au for *Level* = 3 and *Level* = 4, if $d_{tol} > 2$.

There is no established trend for how the accuracy of SMFA (or any other energy-based fragmentation approach) varies as the size of the molecule increases. A qualitative argument (see Section 5.2 in Ref. 2) suggests that the error in the energy might increase as $O(N^{1/2})$, where N represents the number of functional groups in the molecule. However, so far no such trend has been observed, as the variation of the error with the chemical composition and structure of the molecules tested masks any trend with molecular size. A practical difficulty in exploring this subject is the difficulty of obtaining "exact" results for very large molecules, as such calculations become impossible to carry out as the molecular size is increased. Fortunately, the systematic character of SMFA allows some estimate of the error in the energy or property, as the user can systematically increase the value of Level and observe convergence of the energy or property value.

QUANTUM CHEMISTRY METHODS

SMFA carries out the quantum chemistry calculations using one or other of the GAMESS(US) [11], GAUSSIAN[12], NWChem[13] or Q-Chem[14] program packages (hereafter called the quantum chemistry package). The user can select either of these available program packages. SMFA also uses the DALTON [15] program package for some calculations that are not available on one or more of the four packages above, and the DALTON package must also be available for use. In addition to direct use of these packages to calculate the energies and properties of molecular fragments, quantum chemistry data is also used to evaluate the interactions between distantly separated parts of the molecule by perturbation theory. [16 17]

SMFA is suitable for use with any single configuration ab initio quantum chemistry method that is available in the quantum chemistry package (for example, Hartree Fock, Möller-

Plesset methods, coupled cluster methods, etc). The user can also choose any density functional theory (DFT) method available in the quantum chemistry package. At present, SMFA is not suitable for use with multi-configuration methods.

APPLICABILITY

SMFA views a molecule as a collection of chemical functional groups. A group consists of heavy atoms that are connected by multiple bonds (double, triple etc) and all associated hydrogen atoms. Hence, a molecule appears as a set of groups that are connected by *single* bonds. Individual groups are never fragmented in the SMFA process. So, for example, a single benzene ring is not fragmented in SMFA, as it is a single chemical group. Similarly a graphene molecule of any size is not fragmented in SMFA, and the program is not usefully applicable to such molecules. Otherwise, SMFA can be usefully applied to molecules which consist of many chemical functional groups that are connected by single bonds.

Most importantly, SMFA makes it possible to apply reliable quantum chemistry methods to large molecules; methods that can accurately describe chemical reaction enthalpies and activation energies. Reaction pathways can be explored to obtain a quantitative understanding of mechanisms in large molecules, including biological molecules. As illustrated in the examples below, SMFA can be usefully applied to molecules ranging in size from tens to thousands of atoms.

At the smaller end of this range, the advantage is not only a saving in computer time, but the ability to apply very reliably accurate levels of quantum chemistry methods with large basis sets to obtain very reliable information about the mechanism and energetics of chemical reactions. Results that would be difficult to obtain in any other way.

At the high end of molecular size, very reliably accurate levels of quantum chemistry methods with large basis sets can still be applied to study chemical reactions. If the user has

sufficient processing units available, the walltime for highly reliable calculations on such large molecules is no more than that for relatively small molecules. Hence, first principle calculations (rather than say molecular mechanics or QM/MM approaches) can be applied to biological molecules, including an atomic level description of the solvent.

Moreover, reliable levels of quantum chemistry can be applied to the estimation of molecular properties that provide reliable information about molecular structure.

One should note that the current SMFA program produces the electronic energy and other properties for a given molecular configuration. The SMFA program does not automatically produce free energies that are associated with an ensemble of molecular configurations. The user must supply a molecular configuration as input; including the structure of any solvent molecules. SMFA can then be used to locate minimum energy configurations, transition states, and minimum energy paths for a user-selected process.

OPERATION

SMFA requires input from the user, as indicated at the top of Figure 1. The input is achieved via a sequence of questions and prompts from the program to which the user responds interactively. For most calculations, the input required from the user is very little more than would be required to use one of the quantum chemistry packages; namely, the value of Level to be used, and the value of the parameter (d_{tol}) that defines the distance beyond which perturbation theory is used to evaluate interactions between parts of the molecule. For many calculations, the user can ignore the optional input of Figure 1. However, if the system contains metal atoms, or polar solvent molecules, or a (possibly constrained) optimisation is involved, then some extra input is required. The user can also change the default definitions that define the bonding in a molecule.

The flowchart in Figure 1 indicates the process followed by SMFA. The steps on the right are automated. After the fragmentation step, SMFA reports a recommendation for the optimum number of processors to the output file. The program also produces a sample "input deck" for the largest fragment which will be calculated. The user may wish to run test calculations using this "input deck" with various levels of electronic structure theory to evaluate what calculations are feasible and to estimate memory and time requirements. Using this information, the user then enters the final part of the input which is the number of processors to be used, the total memory and disk space to be used and the time limit.

INPUT

- Values of Level and d_{tol}
- File of Cartesian coordinates
- Type of quantum chemistry calculation
- Basis set

OPTIONAL INPUT

- Change bonding definitions
- Specify charges on metal atoms
- Apply embedded charges
- Specify number of optimization steps
- Define a constrained path

SYSTEM PARAMETERS

- Set the number of processors, memory and time limit

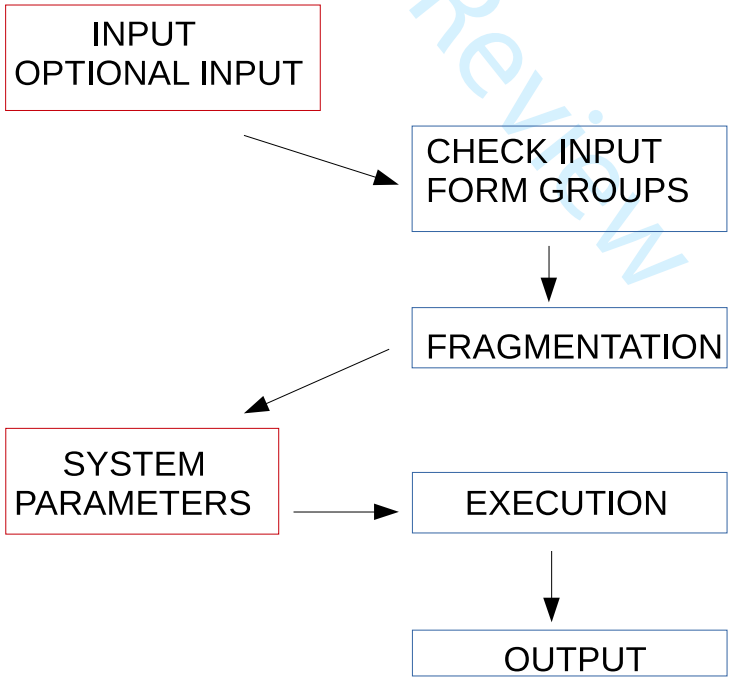


Figure 1. SMFA requires input from the user, as indicated above. Input is achieved via a sequence of questions and responses. The flow chart indicates the stages at which user input is required (on the left), while the stages on the right are automated.

STRUCTURAL OPTIMIZATION

SMFA provides three types of molecular structure optimization: energy minimisation, first-order saddle point (transition state) determination, and energy minimization along a user-defined path. The path is defined by constraints on any combination of bond lengths, bond angles and dihedral angles. Each type of optimization evaluates the hessian at the initial geometry and thereafter uses energy gradients in conjunction with an updated approximate hessian. The code is based on established methods in the literature. [18-20]

PROPERTIES

The program directly provides calculation of the molecular vibrational frequencies, dipole polarizability, and hyperpolarizability. However, any molecular property that is available in the quantum chemistry package can be evaluated by SMFA with the addition of a user-defined script. The User's Guide provides the information required for the composition of a simple script to extract any desired property.

In addition, SMFA can calculate the electrostatic potential (ESP) of a molecule on the solvent accessible surface. The format of the calculated data is designed to interface with graphics programs (for example, VMD[21]). This type of graphic can be useful in obtaining a qualitative understanding of how a small charged molecule might interact with a protein, for example.

SMFA provides harmonic vibrational frequencies (and zero point vibrational energy), infrared intensities and simulated infrared spectrum (with a user-defined line width). Moreover,

SMFA provides frequencies, infrared intensities and simulated infrared spectrum with isotopic substitutions. The user can choose to substitute particular nuclei or all nuclei of a particular element.

Finally, SMFA provides a simple facility to measure atom-atom distances, angles and dihedral angles for a given molecular structure.

ISO-ENERGETIC REACTIONS

SMFA provides a utility program that can be used to compose isoenergetic reaction schemes, without actually having to perform any quantum chemistry calculations. Such reaction schemes can be useful in a number of ways.

An isodesmic chemical reaction [22] is one in which the number and type of chemical bonds is the same for both reactants and products. For example, reactants and products have the same number of C-C, C-H, C=O bonds, etc. Since, the heat of formation of molecules is mostly determined by the number and type of bonds, the heat of reaction is near zero for an isodesmic reaction. This fact allows one to estimate the heat of formation of one species in the reaction, if the heat of formation of all other species is known. A homodesmotic reaction [23] is similar, except that in addition to the same number and types of bonds in both reactants and products, the neighbouring substituents of those bonds are the same. For example, not only is the number of C-C bonds the same, but the number of C-C-X bonds (with a given substituent X) are the same for reactants and products (for all X). Given an even closer correspondence in bonding for reactants and products, a homodesmotic reaction has a heat of reaction which is even more likely to be near zero.

As it turns out, SMFA naturally produces instances of such reactions when it fragments a molecule. The simplest application of SMFA, that is using Level = 1, produces a set of molecular fragments which, with a simple rearrangement, represents an isodesmic reaction involving the

original molecule. Using Level = 2, produces a homodesmotic reaction involving the original molecule. Higher values of Level lead to even more reliably iso-energetic reactions. The SMFA program outputs a description of such reactions, in files that give the Cartesian coordinates of the reactants and products, the coefficients of these molecules in the reaction, the InChI (IUPAC International Chemical Identifier) for each species, and a graphics file that can be used to draw simple 2-dimensional structures of the species.

Once such files have been created for two such near-iso-energetic reactions, SMFA provides the means to subtract one reaction from another to generate an additional near-iso-energetic reaction scheme; and hence with repetition ad infinitum.

SOLVENT

SMFA does not allow the use of implicit solvation via continuum models. However, explicit solvent molecules may be included in the structure, at little additional computational cost if the solvent molecules are small. For example, a structure that contains a large organic or biological molecule solvated in water will be fragmented into small structures that will include parts of the solute, water molecules interacting with parts of the solute, and perhaps many structures that contain only groups of water molecules. These latter fragments are unlikely to be the largest fragments in the set. Hence, using many processors, such solvent fragments are unlikely to contribute to the apparent computation walltime.

EXAMPLES

There are several papers in the literature which demonstrate the accuracy of SMFA for the calculation of the energy and other properties for moderately large molecules. [2, 4 - 10] Some details were included in the section on accuracy above. Hence, herein two examples are presented that rather serve to demonstrate the computational efficiency of SMFA for both moderate sized and large molecules.

Table 1 presents computation times for two small proteins, containing 224 and 1031 atoms, respectively. In the Protein Data Bank (PDB) [24], these molecules have identifiers 1xv4 and 2n37, respectively. Although these are very small proteins in biological terms, they would be considered quite large molecules for the application of quantum chemistry. In column 1 of this table, we compare computation times for the vibrational frequencies of molecule 1xv4, via a direct calculation for the whole molecule versus calculation with SMFA (for a value of Level of 3 and d_{tol} = 1.1). Both calculations were carried out for a geometry optimised at B3LYP/6-31G(d), a common quantum chemistry method for larger molecules. Comparing the frequencies calculated for the whole molecule with those calculated with SMFA shows a mean absolute difference of just 1.2 cm^{-1}. The whole molecule calculation was carried out using the most efficient means available: In parallel on a single node containing 16 cpus. The SMFA program indicated that 32 cpus should be used for maximum efficiency and minimum walltime, so the SMFA calculations were carried out on 32 cpus in two nodes. The table shows that SMFA reduces the total cpu time by a factor of about 8.7, and reduces the walltime by a factor of about 15.5. The power of SMFA is demonstrated by the data in column 2. The molecule 2n37 is sufficiently large that calculation of the frequencies at the B3LYP/6-31G(d) level of quantum chemistry is beyond the computational resources available to us. In fact, it was not even possible to converge a calculation of the energy of this molecule at B3LYP/6-31G(d). In contrast, SMFA calculates the frequencies without difficulty. The SMFA program indicated that 160 cpus should be used for maximum efficiency and minimum walltime. Comparing 1xv4 with 2n37: The number of basis functions increases by a factor of 4.73; the total cpu time increases by a factor of 5.0; the walltime increases by only a factor of 1.2. Since the structure of these two proteins are of course quite different, one cannot expect exact polynomial scaling in cpu times when comparing these two molecules. However, Table 1 strongly indicates that the total cpu time

scales only linearly with the size of the molecule, and that the walltime is independent of the size of the molecule, if sufficient processors are available, up to the number of processors recommended by SMFA.

Table 1. Selected data for the calculation of vibrational frequencies with the B3LYP/6-31G(d) method, for two small proteins, with PDB identifiers 1xv4 and 2n37. "Molecule" denotes calculations carried out on the whole molecule, while "SMFA" denotes calculations carried out using SMFA with Level = 3 and d_{tol} = 1.1. It is important to note that since these calculations were carried out on a shared facility, system delays result in variation of the walltime. The walltimes shown here should be taken as only indicative to within a few minutes.

	1xv4	2n37
Number of Atoms	224	1031
Number of Electrons	820	3934
Number of Basis Functions	1865	8821
Molecule: Number of CPUs	16	-
Molecule: Total CPU Time (mins)	9565.02	-
Molecule: Walltime (mins)	599.92	-
SMFA: Number of CPUs	32	160
SMFA: Total CPU Time (mins)	1101.93	5556.6
SMFA: Walltime (mins)	38.75	45.3

In order to minimise the walltime and employ multiple processors efficiently, SMFA first estimates the minimum number of processors (call it N_{proc}) required to ensure that all processors are fully occupied while the largest fragment is processed (on one processor). The largest N_{proc} fragments are then assigned to N_{proc} processors (one each). Thereafter, as each calculation

completes, the now idle processor is assigned the next fragment from an ordered list of fragments. In this way, SMFA aims to minimise the percentage of time that processors are idle.

One should note that the walltime for SMFA is mainly determined by the time required to perform the quantum chemistry calculation on the largest fragment. This fragment can vary in size and composition from one molecule to another, and even between two conformers of the same molecule, if the distribution of hydrogen bonds is different for the two conformers.

It is worthwhile to show that SMFA may also be useful for much smaller molecules than those in Table 1. A series of crown ethers, cyclic molecules with the formula $(CH_2OCH_2)_n$, were considered with n = 4,..., 9. These molecules contain between 28 and 63 atoms. The geometries of these molecules were optimised at the B3LYP/6-31G(d) level of electronic structure calculation. Generally, these structures are not "flat" but quite buckled ring structures, changing shape significantly as the repeat number, n, increases from 4 to 9. Nevertheless, this series of molecules provides a succession of roughly analogous structures of increasing size. Figures 2 and 3 show how the times for RI-MP2/cc-pVQZ single point calculations of the electronic energy vary for this sequence of structures. Whole molecule calculations could not be completed for n = 9, due to excessive memory requirements.

Figure 2 presents the total computation times for calculations carried out on a single node of 16 processors. The time for a direct calculation of the whole molecules (black dots) are compared with the times for calculations via SMFA (blue dots for Level = 2, and green dots for Level = 3). It is clear that for n > 6, both Level = 2 or 3 yield total cpu times below that for the whole molecule.

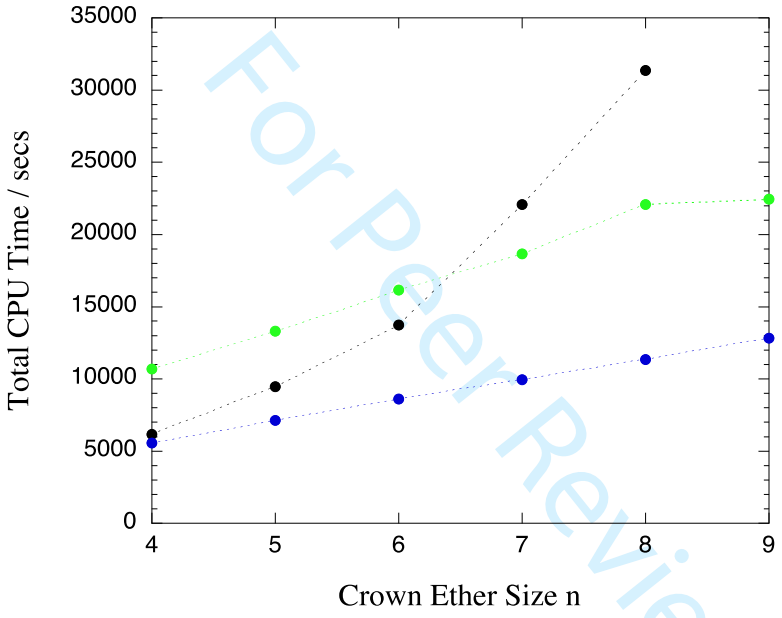


Figure 2. The total computational times for RI-MP2/cc-pVQZ are shown for single point energy calculations on crown ethers, $(CH_2OCH_2)_n$, carried out on a single node of 16 processors. The black symbols correspond to calculations on the whole molecule; blue symbols denote SMFA calculations with Level = 2; green symbols denote Level = 3.

Figure 3 presents the corresponding walltimes for the whole molecule calculation (black), and for the Level = 2 (blue) and Level = 3 (green) calculations. The walltimes for the whole molecule calculations are simply the corresponding total cpu times of Fig. 2, divided by a factor of 16, so these results have been truncated to allow a clearer view of the SMFA results. As noted in Table 1, since the SMFA calculations are performed on a shared facility, system delays result in variations of the walltime reported, and some consequent uncertainty in the data of Fig. 3. Nonetheless, it appears that the walltime for a Level = 2 calculation is always below that for the whole molecule, and the Level 3 calculations also lower for N > 5 (even though the whole molecule calculation is run in parallel on 16 cpus). Linear scaling is reasonably clear for SMFA with Level = 2 or 3. At Level = 3, for n = 4, 5, 6, 7, 8 and 9, SMFA recommends the use of 22, 27, 32, 37, 39, and 44 cpus, respectively, for optimum efficiency and minimum walltime. Since, our system only allows the use of multiple whole nodes, these Level = 3 calculations were repeated with 32, 32, 32, 48, 48, and 48 cpus on two or three nodes, respectively. The resulting walltimes are shown in red in Figure 3. It is apparent that near $O(N^0)$ scaling is achieved by SMFA for these relatively small molecules, if sufficient cpus are available, up to or exceeding the number of cpus recommended by SMFA.

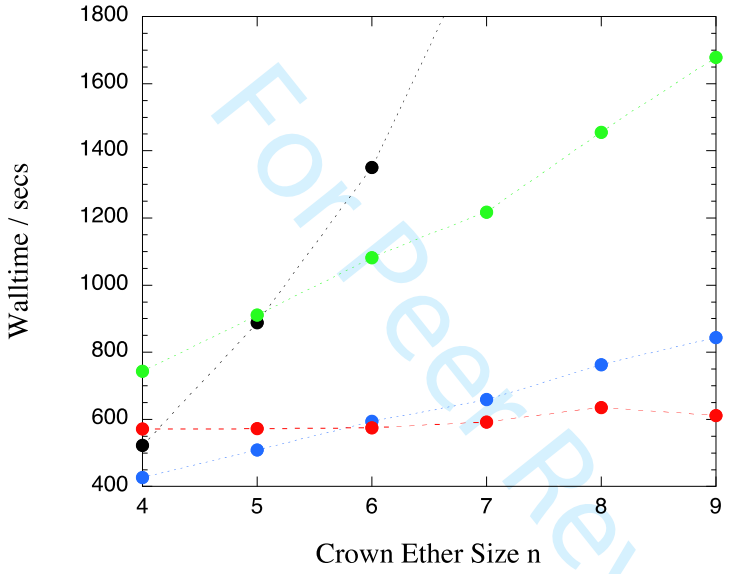


Figure 3. The computation walltimes for RI-MP2/cc-pVQZ single point energy calculations are shown, corresponding to the total cpu times presented in Figure 2. Using a single node with 16 processors, the black symbols denote calculations on the whole molecule; blue symbols denote SMFA with Level = 2; and green symbols denote SMFA with Level = 3. The red symbols denote results for SMFA (Level = 3) where the number of processors is allowed to increase with

molecular size (see the text for details). As noted in Table 1, the walltimes are subject to some uncertainty.

Conclusion

The SMFA program provides the means to use many current quantum chemistry methods of reliable accuracy to estimate the energy, structure and properties of many molecules of moderate to large size. For chemical reactions, reactant, product, and transition state geometries can be found, and reaction paths can be explored. Even when used with a single processing unit, linear scaling of computation time with molecular size is achieved. Moreover, this approach is especially suitable for highly parallel computer systems, where the walltime for the calculation is approximately independent of molecular size if sufficient numbers of cpus are available.

The SMFA program is freely available from the GitHub website, *https://github.com/mickcollins/SMFAPAC.*

Acknowledgments

The authors acknowledge the support of the Australian NCI National Facility where these computations and program development were carried out. The authors also gratefully acknowledge helpful comments on the SMFA program from Dr Terry Frankcombe.

References

(1) Gordon MS, Fedorov DG, Pruitt SR, Slipchenko L. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. Chem. Rev. 2012; 112: 632-672.

(2) Collins MA, Bettens RPA. Energy-Based Molecular Fragmentation Methods. Chem. Rev. 2015; 115: 5607-5642.

(3) Deev V, Collins MA. Approximate ab initio energies by systematic molecular fragmentation. J. Chem. Phys. 2005; 122: 154102.

(4) Collins MA, Deev VA. Accuracy and efficiency of electronic energies from systematic molecular fragmentation. J. Chem. Phys. 2006; 125: 104104.

(5) Addicoat MA, Collins MA. Accurate treatment of non-bonded interactions within systematic molecular fragmentation. J. Chem. Phys. 2009; 131: 104103.

(6) Collins MA. Systematic fragmentation of large molecules by annihilation. Phys. Chem. Chem. Phys. 2012; 14: 7744-7751.

(7) Pruitt SR, Addicoat MA, Collins MA, Gordon MS. The fragment molecular orbital and systematic molecular fragmentation methods applied to water clusters. Phys. Chem. Chem. Phys. 2012; 14: 7752-7764.

(8) Reid DM, Collins MA. Molecular electrostatic potentials by systematic molecular fragmentation. J. Chem. Phys. 2013; 139: 184117.

(9) Collins MA. Molecular forces, geometries and frequencies by systematic molecular fragmentation including embedded charges. J. Chem. Phys. 2014; 141: 094108.

(10) Collins MA, Cvitkovic MW, Bettens RPA. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. Acc. Chem. Res. 2014; 47: 2776-2785.

(11) Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH,

et al. General atomic and molecular electronic structure system. J. Comput. Chem. 1993; 14: 1347-1363.

(12) Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian 09; Revision A.1. Gaussian Inc., Wallingford CT, 2009.

(13) Valiev M, Bylaska EJ, Govind N, Kowalski K, Straatsma TP, van Dam HJJ, et al. NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. Comput. Phys. Comm. 2010; 181: 1477-1489.

(14) Shao Y, Gan Z, Epifanovsky E, Gilbert ATB, Wormit M, Kussman J , et.al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. Mol. Phys. 2015; 113: 184-215.

(15) Aidas K, Angel C, Bak KL, Bakken V, Bast R, et. al. The Dalton quantum chemistry program system. WIREs - Comp. Mol. Sci. 2014; 4: 269-289.

(16) Stone AJ. *The Theory of Intermolecular Forces*. Clarendon, Oxford, 1996.

(17) Stone AJ. Distributed Multipole Analysis: Stability for Large Basis Sets. J. Chem. Theor. Comput. 2005; 1: 1128-1132.

(18) Banerjee A, Adams N, Simons J, Shepard, R. Search for Stationary Points on Surfaces. J. Phys. Chem. 1985; 89: 52-57.

(19) Birkholtz AB, Schlegel HB. Exploration of some refinements to geometry optimization methods. Theor. Chem. Acc. 2016; 135: 84.

(20) Bofill JM. Updated Hessian Matrix and the Restricted Step Method for Locating Transition Structures. J. Comp. Chem. 1994; 15: 1-11.

(21) VMD Visual Molecular Dynamics, http://www.ks.uiuc.edu/Research/vmd/

(22) Hehre WJ, Ditchfield R, Radom L, Pople JA. Molecular orbital theory of the

electronic structure of organic compounds. V. Molecular theory of bond separation. J. Am. Chem Soc. 1970; 92: 4796-4801.

(23) George P, Trachtman M, Bock CW, Brett AM. An alternative approach to the problem of assessing stabilization energies in cyclic conjugated hydrocarbons. Theor. Chim. Acta, 1975, 38: 121-129.

(24) Protein-Data-Bank, https://www.rcsb.org/

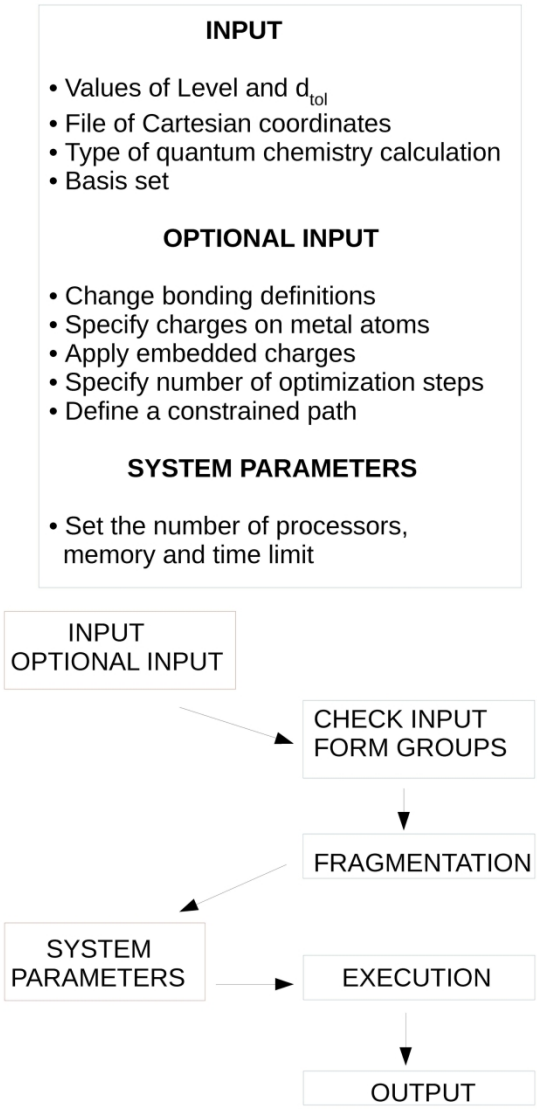


Figure 1. SMFA requires input from the user, as indicated above. Input is achieved via a sequence of questions and responses. The flow chart indicates the stages at which user input is required (on the left), while the stages on the right are automated.

209x296mm (300 x 300 DPI)

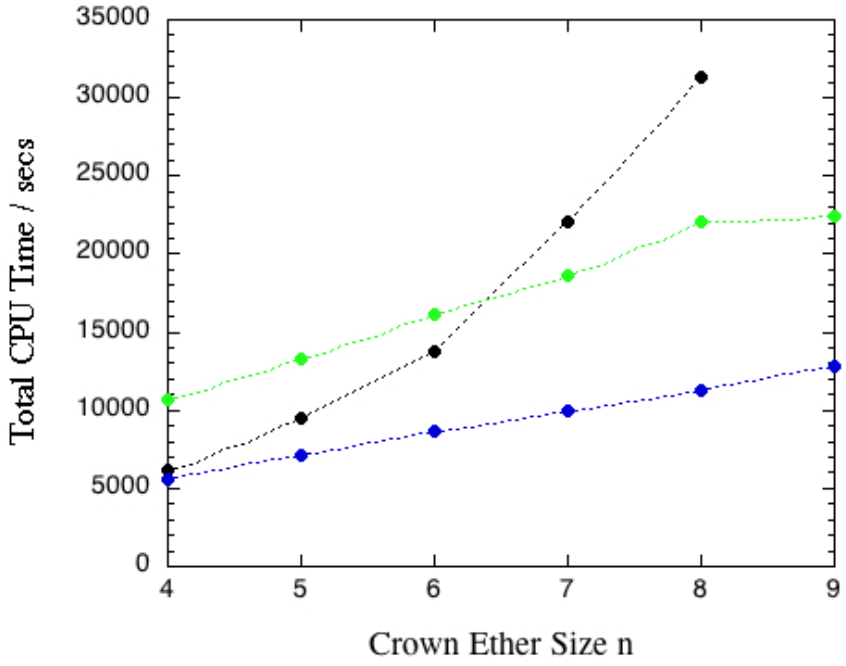


Figure 2. The total computational times for RI-MP2/cc-pVQZ are shown for single point energy calculations on crown ethers, (CH2OCH2)n, carried out on a single node of 16 processors. The black symbols correspond to calculations on the whole molecule; blue symbols denote SMFA calculations with Level = 2; green symbols denote Level = 3.

190x156mm (72 x 72 DPI)

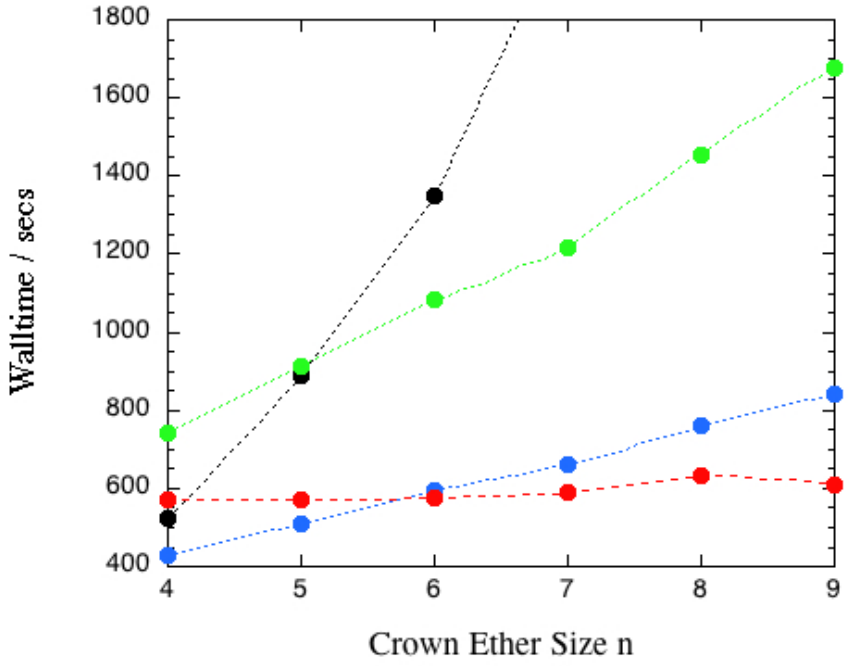


Figure 3. The computation walltimes for RI-MP2/cc-pVQZ single point energy calculations are shown, corresponding to the total cpu times presented in Figure 2. Using a single node with 16 processors, the black symbols denote calculations on the whole molecule; blue symbols denote SMFA with Level = 2; and green symbols denote SMFA with Level = 3. The red symbols denote results for SMFA (Level = 3) where the number of processors is allowed to increase with molecular size (see the text for details). As noted in Table 1, the walltimes are subject to some uncertainty.

190x156mm (72 x 72 DPI)

Article Title: The SMFA program for quantum chemistry calculations on large molecules

Article Type: Software Focus

Authors:

Rika Kobayashi
Australian National University, Leonard Huxley Bldg 56, Mills Road, Canberra, ACT 2601, Australia. Email: Rika.Kobayashi@anu.edu.au, ORCID: 0000-0002-0672-833X

Matthew A. Addicoat
Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. Email: matthew.addicoat@ntu.ac.uk

Andrew T. B. Gilbert
Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. Email: andrew.iqmol@gmail.com, ORCID: 0000-0002-8211-9261

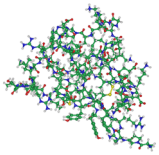
Roger D. Amos
Australian National University, Leonard Huxley Bldg 56, Mills Road, Canberra, ACT 2601, Australia. Email: rogerdamos@gmail.com, ORCID: 0000-0003-2868-7994

Michael A. Collins*
Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia. Email: collins@rsc.anu.edu.au, ORCID: 0000-0003-1920-5105

Abstract

SMFA is a general program package for performing quantum chemistry calculations on large molecules, using an energy-based fragmentation approach. The program can calculate electronic energies, energy gradients and second derivatives; perform geometry optimization; find first order saddle points (transition states); perform energy optimized scans along a user-defined path; and evaluate various molecular properties. The program can use any of the following quantum chemistry packages: GAMESS(US), GAUSSIAN, NWChem and Q-Chem. In addition, SMFA provides a number of utility programs that, *inter alia*, calculate vibrational frequencies and infrared spectra with isotopic substitutions, the electrostatic potential on the solvent-accessible-surface, and isodesmic and higher order near-iso-energetic reaction schemes. Calculations of the electronic energy and related properties can be carried out using a scheme that provides a computation time that is linearly dependent on the size of the molecule or, if the user has enough processing units available, in a walltime that is independent of the size of the molecule.

Graphical/Visual Abstract and Caption



Graphics Caption: SMFA enables quantum chemistry calculations for proteins

Introduction

The field generally known as quantum chemistry is concerned with the calculation of the electronic structure of molecules. Over many decades, many computational methods have been developed to approximate the total energy of the electrons in a molecule, and associated molecular properties. With this knowledge, one can explain known chemical processes and predict previously unknown processes. Quantum chemistry has proven to be a powerful tool in chemistry generally. However, these computational methods are difficult to apply to large molecules, as the computational time (and memory) scale up rapidly with molecular size. If "N" represents some measure of molecular size (say, the total number of basis functions used to describe the electronic wavefunction), a widely used but not very reliable method (density functional theory) requires a computation time that scales as $O(N^3)$ - $O(N^4)$, while the most commonly used reliable method (a "coupled cluster" method) scales as $O(N^7)$. Substantial effort has been made to reduce these "scaling powers" to near linear scaling $O(N^1)$ through the use of highly parallelized codes. However, computation times remain very high. Consequently, the most reliable methods can only be applied to small molecules (containing perhaps tens of atoms), and even the less reliable methods cannot easily be applied to molecules containing hundreds or thousands of atoms.

Over the last decade, so-called fragmentation methods have been developed which aim to avoid the high power scaling problem of traditional quantum chemistry methods. [1 2]

SMFA (Systematic Molecular Fragmentation by Annihilation) is a general purpose program package which aims to approximate quantum chemistry calculations of molecular energies and properties for large organic, inorganic and biological molecules. The approximation employed, systematic molecular fragmentation by annihilation, [2 - 10] decomposes a molecule into relative small pieces (fragments), on which quantum chemistry calculations are carried out.

The number of fragments is linearly proportional to the number of chemical functional groups in the molecule. The size of the fragments is independent of the size of the whole molecule. The energy and properties of the molecule can be accurately approximated from a combination, respectively, of the energy and properties of these small fragments. Hence, if calculations on these fragments are carried out sequentially, the computation time is linearly proportional to the number of chemical functional groups in the molecule $O(N^1)$. However, these fragment calculations are completely independent of one another, hence if sufficient processing units are available, the "walltime" for any calculation only depends on the time required to perform a calculation on the largest fragment, which is independent of the total number of chemical functional groups $O(N^0)$. Thus, the SMFA approach provides a route to achieving high levels of parallelisation that is independent of the level of theory used, therefore avoiding the implementation overhead of parallelising each method separately.

THE SMFA METHOD

SMFA is based on a completely automated procedure for decomposing a molecule into small fragments. The method has been described in detail elsewhere. [6 9 10] To summarise, the computer algorithm can be briefly described as follows.

(1) Bonds and functional groups are defined, consistent with normal chemical concepts. A molecule is viewed as a collection of groups connected by single bonds.

(2) Beginning with some arbitrary group (call it A):

(i) Remove group A from the molecule;

(ii) leave A in place, but remove all groups separated from A by more than a specified number of bonds (call it "Level" bonds); and

(iii) remove group A and all the groups in (ii) from the molecule.

(iv) The molecule is represented by the sum of the fragments produced in steps (i) and (ii),

minus the fragments produced in step (iii).

(v) Taking each fragment in (iv), repeat steps (i) to (iii), and continue with every new fragment created until there is no fragment that has groups separated by more than Level bonds.

This procedure yields a set of fragments $\{F_n\}$ (N_{frag} in number) with integer coefficients $\{f_n\}$. Hydrogen atom caps are appended to the atoms with dangling bonds, with bond lengths calculated using the covalent radii of the atoms involved. In this way, the molecule, *M*, is represented by a sum of fragments, and the "bonded" energy, E_b, is a sum of the energies of the fragments.

$$M \rightarrow \sum_{n=1}^{N_{frag}} f_n F_n \qquad (1)$$

$$E_b = \sum_{n=1}^{N_{frag}} f_n E\left(F_n\right) \qquad (2)$$

Eq. (2) neglects the interactions between groups that are separated by more than Level bonds. These "nonbonded" interactions are accounted for as follows. First, a Level = 1 fragmentation is evaluated (comprising $N^{(1)}_{frag}$ fragments), and the nonbonded interactions, E_{nb}, are given by the interactions of these Level = 1 fragments:

$$E_{nb} = \sum_{n_1=1}^{N^{(1)}_{frag}-1} \sum_{n_2=n_1+1}^{N^{(1)}_{frag}} f^{(1)}_{n_1} f^{(1)}_{n_2} E\left[F^{(1)}_{n_1} \leftrightarrow F^{(1)}_{n_2}\right]_{allowed} \qquad (3)$$

where $E\left[F^{(1)}_{n_1} \leftrightarrow F^{(1)}_{n_2}\right]_{allowed}$ denotes the interaction energy of Level = 1 fragments that do not include group interactions that have already been accounted for in the bonded energy of Eq. (2). These fragment–fragment interaction energies are evaluated using ab initio calculations if the fragment–fragment separation is less than some specified tolerance, denoted by d_{tol} (the

nearest atom-atom distance divided by the sum of the corresponding Van der Waals radii). For longer separations, the interaction energy is evaluated using perturbation theory [based on ab initio calculations of charge distributions, static polarizability and the imaginary frequency polarizability] as the sum of an electrostatic interaction, a dispersion interaction and a contribution from induction. When formally charged groups are present in the molecule, or a polar solvent is present, the significant induction effect is accounted for by carrying out all ab initio calculations in the presence of appropriate embedded charges.[9]

This is a completely automated procedure in the SMFA program. If the user selects larger and larger values of the integer parameter Level, the fragments, F_n, are larger, and the total energy [Eq. (2) plus Eq. (3) plus the perturbative contributions] can be seen to converge (to near the correct value for the ab initio method and basis set which the user has chosen). Accurate energy gradients and higher derivatives are obtained by differentiating Eqs (1) and (2).

Similarly, since any property of a molecule can be evaluated as the derivative of the molecular Hamiltonian in the presence of an external field, equations similar to Eqs (1) and (2) have been derived for any molecular property. The method for obtaining such properties from the SMFA program is described below (and in detail in the User's Guide that accompanies the program).

The SMFA method shares some similar features and some significant differences from other "energy-based" molecular fragmentation methods. Several of these methods have been reviewed and compared in the review article of Ref. 2. In particular, SMFA tends to produce a similarly accurate approximation to the molecular energy as other approaches but with somewhat smaller fragment sizes. This is probably due to the fact that the fragmentation method relies primarily on the bonded connectivity of a molecule, rather than a distance-based algorithm.

It is important to note that SMFA (like the other approaches reviewed in Ref. 2) is essentially an ab initio quantum chemistry method, rather than a "QM/MM" method. There is no distinction in SMFA between an "inner" part of the molecule that is treated by quantum electronic structure methods and an "outer" region which is treated using molecular mechanics. Hence there is no arbitrary or intuitive "user-defined" region that is the only part of the molecule that is treated correctly using first principles methods. The whole molecule (including any solvent) is treated from first principles. The essential approximation made is that "chemistry is local": That is, the influence that some part of the molecule has on the electronic structure of another part of the molecule decays in magnitude with the distance and bonded connectivity of the two regions.

A major advantage of SMFA is that it can be applied systematically. The essential approximation can be tested simply by repeating the application of SMFA with increasing values of the parameter Level. If the SMFA approximation is valid, then the values of the energy or property obtained should converge as the value of Level increases.

ACCURACY of SMFA

There have been several reports on the accuracy of SMFA in various applications. [4 - 7, 9] To illustrate these results, we note that for a set of 96 neutral organic molecules containing 18 to 80 atoms, the mean absolute error in the energy (MAEE) for a range of *ab initio* methods and basis sets was 2.1 to 4.4 kJ mol^{-1}, for *Level* = 3, and d_{tol} = 1.1.[5] A set of larger molecules, containing 45 to 180 atoms, was considered in Ref. 9: 24 peptides and sugars (with extensive hydrogen bonding) showed an MAEE of 14.9, 5.6 and 3.9 kJ mol^{-1}, for Levels 3, 4, and 5, respectively; 14 ions and zwitterions had an MAEE of 8.2, 4.5 and 2.9 kJ mol^{-1}, for Levels 3, 4, and 5, respectively; 20 conformers of a protein containing 246 atoms (with 3 formally charged groups) had a mean absolute error in their relative energies of 4.6 kJ mol^{-1} at Level 3. Moreover,

for these 20 conformers, the root-mean-square error in the energy gradient falls below about 0.0003 au for *Level* = 3 and *Level* = 4, if $d_{tol} > 2$.

There is no established trend for how the accuracy of SMFA (or any other energy-based fragmentation approach) varies as the size of the molecule increases. A qualitative argument (see Section 5.2 in Ref. 2) suggests that the error in the energy might increase as $O(N^{1/2})$, where N represents the number of functional groups in the molecule. However, so far no such trend has been observed, as the variation of the error with the chemical composition and structure of the molecules tested masks any trend with molecular size. A practical difficulty in exploring this subject is the difficulty of obtaining "exact" results for very large molecules, as such calculations become impossible to carry out as the molecular size is increased. Fortunately, the systematic character of SMFA allows some estimate of the error in the energy or property, as the user can systematically increase the value of Level and observe convergence of the energy or property value.

~~The SMFA procedure is entitled "systematic" because fragmentation of the molecule can be carried out at a sequence of integer values of the parameter "Level". Higher values of Level result in larger fragments and more reliable estimates of energy and property values. Convergence of such energy/property values can be observed as the value of Level is increased.~~

QUANTUM CHEMISTRY METHODS

SMFA carries out the quantum chemistry calculations using one or other of the GAMESS(US) [11], GAUSSIAN[12], NWChem[13] or Q-Chem[14] program packages (hereafter called the quantum chemistry package). The user can select either of these available program packages. SMFA also uses the DALTON [15] program package for some calculations that are not available on one or more of the four packages above, and the DALTON package must also be available for

use. In addition to direct use of these packages to calculate the energies and properties of molecular fragments, quantum chemistry data is also used to evaluate the interactions between distantly separated parts of the molecule by perturbation theory. [16 17]

SMFA is suitable for use with any single configuration ab initio quantum chemistry method that is available in the quantum chemistry package (for example, Hartree Fock, Möller-Plesset methods, coupled cluster methods, etc). The user can also choose any density functional theory (DFT) method available in the quantum chemistry package. At present, SMFA is not suitable for use with multi-configuration methods.

APPLICABILITY

SMFA views a molecule as a collection of chemical functional groups. A group consists of heavy atoms that are connected by multiple bonds (double, triple etc) and all associated hydrogen atoms. Hence, a molecule appears as a set of groups that are connected by *single* bonds. Individual groups are never fragmented in the SMFA process. So, for example, a single benzene ring is not fragmented in SMFA, as it is a single chemical group. Similarly a graphene molecule of any size is not fragmented in SMFA, and the program is not usefully applicable to such molecules. Otherwise, SMFA can be usefully applied to molecules which consist of many chemical functional groups that are connected by single bonds.

Most importantly, SMFA makes it possible to apply reliable quantum chemistry methods to large molecules; methods that can accurately describe chemical reaction enthalpies and activation energies. Reaction pathways can be explored to obtain a quantitative understanding of mechanisms in large molecules, including biological molecules. As illustrated in the examples below, SMFA can be usefully applied to molecules ranging in size from tens to thousands of atoms.

At the smaller end of this range, the advantage is not only a saving in computer time, but the ability to apply very reliably accurate levels of quantum chemistry methods with large basis sets to obtain very reliable information about the mechanism and energetics of chemical reactions. Results that would be difficult to obtain in any other way.

At the high end of molecular size, very reliably accurate levels of quantum chemistry methods with large basis sets can still be applied to study chemical reactions. If the user has sufficient processing units available, the walltime for highly reliable calculations on such large molecules is no more than that for relatively small molecules. Hence, first principle calculations (rather than say molecular mechanics or QM/MM approaches) can be applied to biological molecules, including an atomic level description of the solvent.

Moreover, reliable levels of quantum chemistry can be applied to the estimation of molecular properties that provide reliable information about molecular structure.

One should note that the current SMFA program produces the electronic energy and other properties for a given molecular configuration. The SMFA program does not automatically produce free energies that are associated with an ensemble of molecular configurations. The user must supply a molecular configuration as input; including the structure of any solvent molecules. SMFA can then be used to locate minimum energy configurations, transition states, and minimum energy paths for a user-selected process.

OPERATION

~~SMFA requires the user to input the data which is required by the quantum chemistry package: type of calculation, basis set, and so on. The molecular geometry must be provided in a file of Cartesian coordinates. The input is achieved via a sequence of questions and prompts from the program to which the user responds interactively. The user then requests the program to automatically check the input and perform the fragmentation process. The program outputs~~

~~(to an output file) some details about the fragmentation process (purely for information) and outputs a recommendation for the optimum number of processors required for a parallel calculation, and a sample "input deck" for the largest fragment which will be calculated. Using this information, the user then enters the final part of the input which is the number of processors to be used, the total memory and disk space to be used and the time limit. Given the instruction to proceed, the program automatically performs all necessary calculations and appends the final results to the output file.~~

~~For most calculations, the input required from the user is very little more than would be required to use one of the quantum chemistry packages; namely, the value of Level to be used, and the value of a parameter that defines the distance beyond which perturbation theory is used to evaluate interactions between parts of the molecule. When metal atoms are present in the molecule, the user must define the formal charge on these metals. The user also has the option of changing the program defaults that define the bonding between atoms.~~

SMFA requires input from the user, as indicated at the top of Figure 1. The input is achieved via a sequence of questions and prompts from the program to which the user responds interactively. For most calculations, the input required from the user is very little more than would be required to use one of the quantum chemistry packages; namely, the value of Level to be used, and the value of the parameter (d_{tol}) that defines the distance beyond which perturbation theory is used to evaluate interactions between parts of the molecule. For many calculations, the user can ignore the optional input of Figure 1. However, if the system contains metal atoms, or polar solvent molecules, or a (possibly constrained) optimisation is involved, then some extra input is required. The user can also change the default definitions that define the bonding in a molecule.

The flowchart in Figure 1 indicates the process followed by SMFA. The steps on the right are automated. After the fragmentation step, SMFA reports a recommendation for the optimum number of processors to the output file. The program also produces a sample "input deck" for the largest fragment which will be calculated. The user may wish to run test calculations using this "input deck" with various levels of electronic structure theory to evaluate what calculations are feasible and to estimate memory and time requirements. Using this information, the user then enters the final part of the input which is the number of processors to be used, the total memory and disk space to be used and the time limit.

INPUT

- Values of Level and d_{tol}
- File of Cartesian coordinates
- Type of quantum chemistry calculation
- Basis set

OPTIONAL INPUT

- Change bonding definitions
- Specify charges on metal atoms
- Apply embedded charges
- Specify number of optimization steps
- Define a constrained path

SYSTEM PARAMETERS

- Set the number of processors, memory and time limit

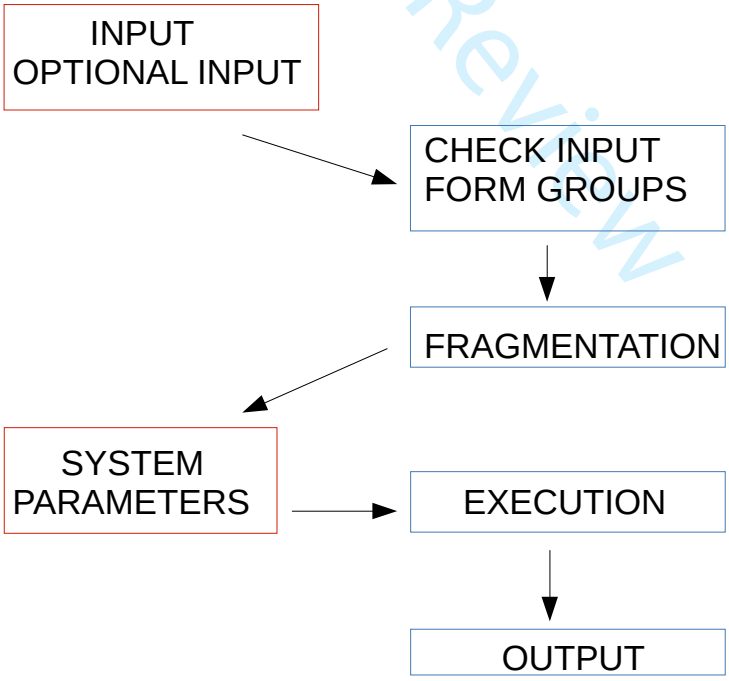


Figure 1. SMFA requires input from the user, as indicated above. Input is achieved via a sequence of questions and responses. The flow chart indicates the stages at which user input is required (on the left), while the stages on the right are automated.

STRUCTURAL OPTIMIZATION

SMFA provides three types of molecular structure optimization: energy minimisation, first-order saddle point (transition state) determination, and energy minimization along a user-defined path. The path is defined by constraints on any combination of bond lengths, bond angles and dihedral angles. Each type of optimization evaluates the hessian at the initial geometry and thereafter uses energy gradients in conjunction with an updated approximate hessian. The code is based on established methods in the literature. [18-20]

PROPERTIES

The program directly provides calculation of the molecular vibrational frequencies, dipole polarizability, and hyperpolarizability. However, any molecular property that is available in the quantum chemistry package can be evaluated by SMFA with the addition of a user-defined script. The User's Guide provides the information required for the composition of a simple script to extract any desired property.

In addition, SMFA can calculate the electrostatic potential (ESP) of a molecule on the solvent accessible surface. The format of the calculated data is designed to interface with graphics programs (for example, VMD[21]). This type of graphic can be useful in obtaining a qualitative understanding of how a small charged molecule might interact with a protein, for example.

SMFA provides harmonic vibrational frequencies (and zero point vibrational energy), infrared intensities and simulated infrared spectrum (with a user-defined line width). Moreover,

SMFA provides frequencies, infrared intensities and simulated infrared spectrum with isotopic substitutions. The user can choose to substitute particular nuclei or all nuclei of a particular element.

Finally, SMFA provides a simple facility to measure atom-atom distances, angles and dihedral angles for a given molecular structure.

ISO-ENERGETIC REACTIONS

SMFA provides a utility program that can be used to compose isoenergetic reaction schemes, without actually having to perform any quantum chemistry calculations. Such reaction schemes can be useful in a number of ways.

An isodesmic chemical reaction [22] is one in which the number and type of chemical bonds is the same for both reactants and products. For example, reactants and products have the same number of C-C, C-H, C=O bonds, etc. Since, the heat of formation of molecules is mostly determined by the number and type of bonds, the heat of reaction is near zero for an isodesmic reaction. This fact allows one to estimate the heat of formation of one species in the reaction, if the heat of formation of all other species is known. A homodesmotic reaction [23] is similar, except that in addition to the same number and types of bonds in both reactants and products, the neighbouring substituents of those bonds are the same. For example, not only is the number of C-C bonds the same, but the number of C-C-X bonds (with a given substituent X) are the same for reactants and products (for all X). Given an even closer correspondence in bonding for reactants and products, a homodesmotic reaction has a heat of reaction which is even more likely to be near zero.

As it turns out, SMFA naturally produces instances of such reactions when it fragments a molecule. The simplest application of SMFA, that is using Level = 1, produces a set of molecular fragments which, with a simple rearrangement, represents an isodesmic reaction involving the

original molecule. Using Level = 2, produces a homodesmotic reaction involving the original molecule. Higher values of Level lead to even more reliably iso-energetic reactions. The SMFA program outputs a description of such reactions, in files that give the Cartesian coordinates of the reactants and products, the coefficients of these molecules in the reaction, the InChI (IUPAC International Chemical Identifier) for each species, and a graphics file that can be used to draw simple 2-dimensional structures of the species.

Once such files have been created for two such near-iso-energetic reactions, SMFA provides the means to subtract one reaction from another to generate an additional near-iso-energetic reaction scheme; and hence with repetition ad infinitum.

SOLVENT

SMFA does not allow the use of implicit solvation via continuum models. However, explicit solvent molecules may be included in the structure, at little additional computational cost if the solvent molecules are small. For example, a structure that contains a large organic or biological molecule solvated in water will be fragmented into small structures that will include parts of the solute, water molecules interacting with parts of the solute, and perhaps many structures that contain only groups of water molecules. These latter fragments are unlikely to be the largest fragments in the set. Hence, using many processors, such solvent fragments are unlikely to contribute to the apparent computation walltime.

EXAMPLES

There are several papers in the literature which demonstrate the accuracy of SMFA for the calculation of the energy and other properties for moderately large molecules. [2, 4 - 10] Some details were included in the section on accuracy above. ~~In general, SMFA attains so-called chemical accuracy (a few kJ mol^{-1}) for values of 3 or 4 for the parameter Level.~~ Hence, herein

two examples are presented that rather serve to demonstrate the computational efficiency of SMFA for both moderate sized and large molecules.

Table 1 presents computation times for two small proteins, containing 224 and 1031 atoms, respectively. In the Protein Data Bank (PDB) [24], these molecules have identifiers 1xv4 and 2n37, respectively. Although these are very small proteins in biological terms, they would be considered quite large molecules for the application of quantum chemistry. In column 1 of this table, we compare computation times for the vibrational frequencies of molecule 1xv4, via a direct calculation for the whole molecule versus calculation with SMFA (and a value of Level of 3 and d_{tol} = 1.1). ~~Previous calculations showed that the energy of this molecule converged to within a few kJ mol^{-1} by Level = 3.~~ Both calculations were carried out for a geometry optimised at B3LYP/6-31G(d), a common quantum chemistry method for larger molecules. Comparing the frequencies calculated for the whole molecule with those calculated with SMFA shows a mean absolute difference of just 1.2 cm^{-1}. The whole molecule calculation was carried out using the most efficient means available: ~~In parallel on a 28 cpu node.~~ In parallel on a single node containing 16 cpus. The SMFA program indicated that 32 cpus should be used for maximum efficiency and minimum walltime, so the SMFA calculations were carried out on 32 cpus in two nodes. The processors of both machines are very close in speed. The table shows that SMFA reduces the total cpu time by a factor of about ~~7.3~~8.7, and reduces the walltime by a factor of about ~~7.8~~15.5. The power of SMFA is demonstrated by the data in column 2. The molecule 2n37 is sufficiently large that calculation of the frequencies at the B3LYP/6-31G(d) level of quantum chemistry is beyond the computational resources available to us. In fact, it was not even possible to converge a calculation of the energy of this molecule at B3LYP/6-31G(d). In contrast, SMFA calculates the frequencies without difficulty. The SMFA program indicated that 160 cpus should be used for maximum efficiency and minimum walltime. Comparing 1xv4 with 2n37: The

number of basis functions increases by a factor of 4.73; the total cpu time increases by a factor of ~~5.26~~5.0; the walltime increases by only a factor of ~~1.25~~1.2. Since the structure of these two proteins are of course quite different, one cannot expect exact polynomial scaling in cpu times when comparing these two molecules. However, Table 1 strongly indicates that the total cpu time scales only linearly with the size of the molecule, and that the walltime is independent of the size of the molecule, if sufficient processors are available, up to the number of processors recommended by SMFA.

Table 1. Selected data for the calculation of vibrational frequencies with the B3LYP/6-31G(d) method, for two small proteins, with PDB identifiers 1xv4 and 2n37. "Molecule" denotes calculations carried out on the whole molecule, while "SMFA" denotes calculations carried out using SMFA with Level = 3 and d_{tol} = 1.1. It is important to note that since these calculations were carried out on a shared facility, system delays result in variation of the walltime. The walltimes shown here should be taken as only indicative to within a few minutes.

	1xv4	2n37
Number of Atoms	224	1031
Number of Electrons	820	3934
Number of Basis Functions	1865	8821
Molecule: Number of CPUs	~~28~~16	-
Molecule: Total CPU Time (mins)	~~7793.8~~9565.02	-
Molecule: Walltime (mins)	~~280.9~~599.92	-
SMFA: Number of CPUs	32	160
SMFA: Total CPU Time (mins)	~~1056.4~~1101.93	5556.6
SMFA: Walltime (mins)	~~36.3~~38.75	45.3

In order to minimise the walltime and employ multiple processors efficiently, SMFA first estimates the minimum number of processors (call it N_{proc}) required to ensure that all processors are fully occupied while the largest fragment is processed (on one processor). The largest N_{proc} fragments are then assigned to N_{proc} processors (one each). Thereafter, as each calculation completes, the now idle processor is assigned the next fragment from an ordered list of fragments. In this way, SMFA aims to minimise the percentage of time that processors are idle.

One should note that the walltime for SMFA is mainly determined by the time required to perform the quantum chemistry calculation on the largest fragment. This fragment can vary in size and composition from one molecule to another, and even between two conformers of the same molecule, if the distribution of hydrogen bonds is different for the two conformers.

It is worthwhile to show that SMFA may also be useful for much smaller molecules than those in Table 1. A series of crown ethers, cyclic molecules with the formula $(CH_2OCH_2)_n$, were considered with n = 4,..., 9. These molecules contain between 28 and 63 atoms. The geometries of these molecules were optimised at the B3LYP/6-31G(d) level of electronic structure calculation. Generally, these structures are not "flat" but quite buckled ring structures, changing shape significantly as the repeat number, n, increases from 4 to 9. Nevertheless, this series of molecules provides a succession of roughly analogous structures of increasing size. Figures ~~1~~2 and ~~2~~3 show how the times for RI-MP2/cc-pVQZ calculations of the electronic energy vary for this sequence of structures. Whole molecule calculations could not be completed for n = 9, due to excessive memory requirements.

Figure ~~1~~2 presents the total computation times for calculations carried out on a single node of 16 processors. The time for a direct calculation of the whole molecules (black dots) are compared with the times for calculations via SMFA (blue dots for Level = 2, and green dots for

Level = 3). It is clear that for n > 6, both Level = 2 or 3 yield total cpu times below that for the whole molecule.

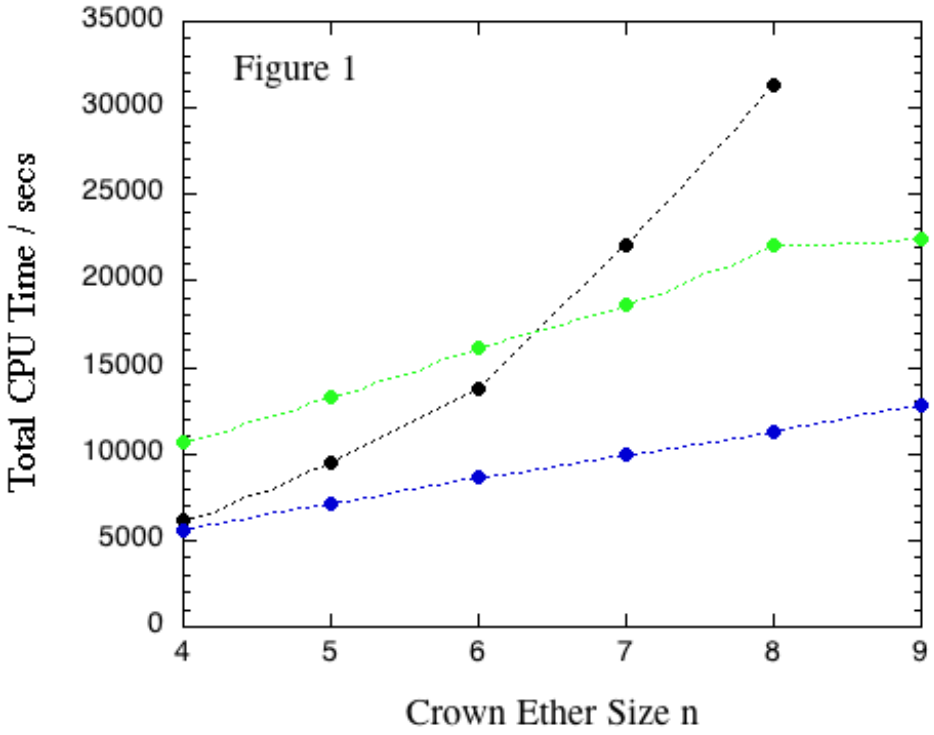


Figure ~~1~~2. The total computational times for RI-MP2/cc-pVQZ are shown for single point energy calculations on crown ethers, $(CH_2OCH_2)_n$, carried out on a single node of 16 processors. The black symbols correspond to calculations on the whole molecule; blue symbols denote SMFA calculations with Level = 2; green symbols denote Level = 3.

~~Figure 2 presents the corresponding walltimes for the whole molecule calculation (black), and for the Level = 2 (blue) and Level = 3 (green) calculations. While the walltime for a Level = 2 calculation is always below that for the whole molecule, the Level 3 calculations also lower for N > 5 (even though the whole molecule calculation is run in parallel on 16 cpus). It~~

~~appears that the whole molecule calculation does not scale linearly with molecular size, but that linear scaling is reasonably clear for SMFA with Level = 2 or 3. At Level = 3, for n = 4, 5, 6, 7, 8 and 9, SMFA recommends the use of 22, 27, 32, 37, 39, and 44 cpus, respectively, for optimum efficiency and minimum walltime. Since, our system only allows the use of multiple whole nodes, these Level = 3 calculations were repeated with 32, 32, 32, 48, 48, and 48 cpus on two or three nodes, respectively. The resulting walltimes are shown in red in Figure 2. It is apparent that near O(N^0) scaling is achieved by SMFA for these relatively small molecules.~~

Figure 3 presents the corresponding walltimes for the whole molecule calculation (black), and for the Level = 2 (blue) and Level = 3 (green) calculations. The walltimes for the whole molecule calculations are simply the corresponding total cpu times of Fig. 2, divided by a factor of 16, so these results have been truncated to allow a clearer view of the SMFA results. As noted in Table 1, since the SMFA calculations are performed on a shared facility, system delays result in variations of the walltime reported, and some consequent uncertainty in the data of Fig. 3. Nonetheless, it appears that the walltime for a Level = 2 calculation is always below that for the whole molecule, and the Level 3 calculations also lower for N > 5 (even though the whole molecule calculation is run in parallel on 16 cpus). Linear scaling is reasonably clear for SMFA with Level = 2 or 3. At Level = 3, for n = 4, 5, 6, 7, 8 and 9, SMFA recommends the use of 22, 27, 32, 37, 39, and 44 cpus, respectively, for optimum efficiency and minimum walltime. Since, our system only allows the use of multiple whole nodes, these Level = 3 calculations were repeated with 32, 32, 32, 48, 48, and 48 cpus on two or three nodes, respectively. The resulting walltimes are shown in red in Figure 3. It is apparent that near O(N^0) scaling is achieved by SMFA for these relatively small molecules, if sufficient cpus are available, up to or exceeding the number of cpus recommended by SMFA.

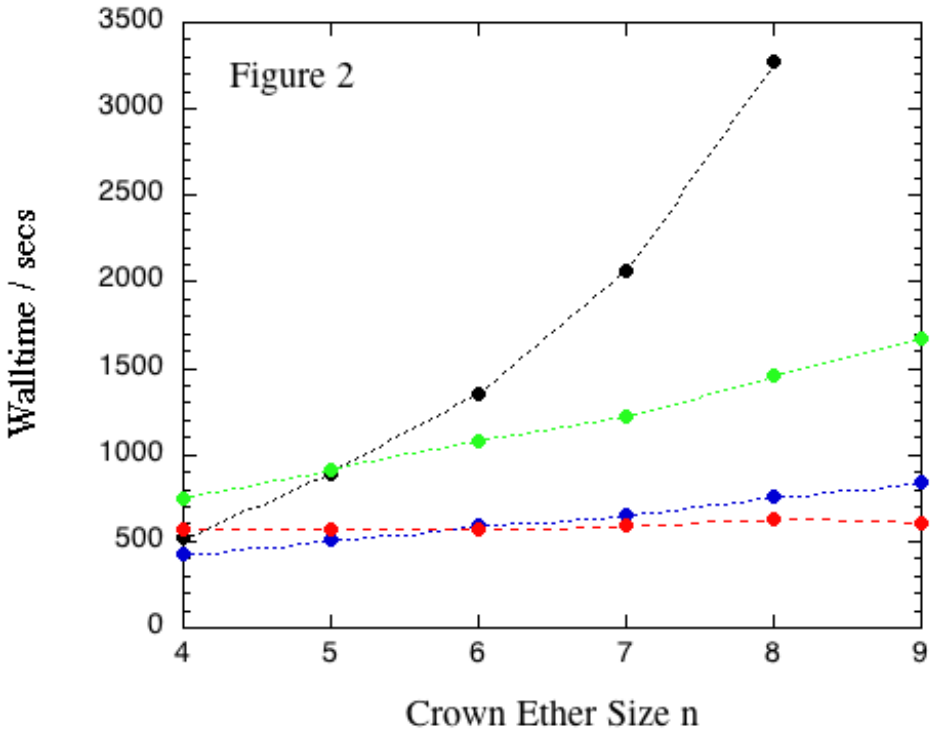


Figure 2
Walltime / secs
3500
3000
2500
2000
1500
1000
500
0
4
5
6
7
8
9
Crown Ether Size n

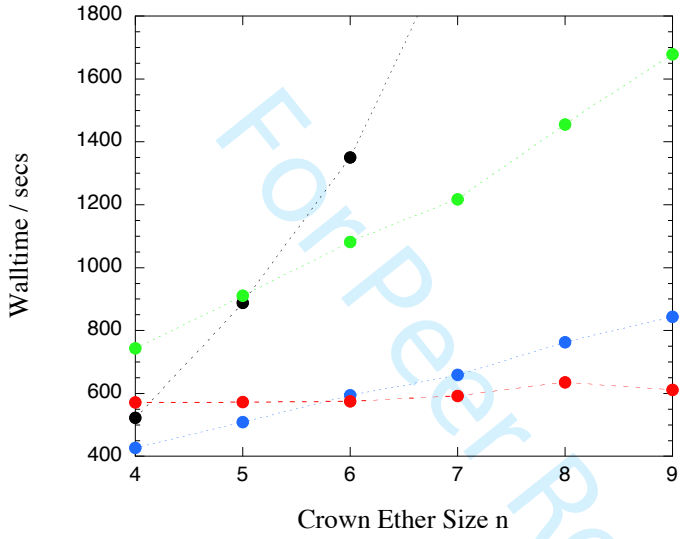


Figure ~~23~~ 23. The computation walltimes for RI-MP2/cc-pVQZ are shown, corresponding to the total cpu times presented in Figure 1. Using a single node with 16 processors, the black symbols denote calculations on the whole molecule; blue symbols denote SMFA with Level = 2; and green symbols denote SMFA with Level = 3. The red symbols denote results for SMFA (Level = 3) where the number of processors is allowed to increase with molecular size (see the text for details). As noted in Table 1, the walltimes are subject to some uncertainty.

Conclusion

The SMFA program provides the means to use many current quantum chemistry methods of reliable accuracy to estimate the energy, structure and properties of many molecules of moderate to large size. For chemical reactions, reactant, product, and transition state geometries can be found, and reaction paths can be explored. Even when used with a single processing unit, linear scaling of computation time with molecular size is achieved. Moreover, this approach is especially suitable for highly parallel computer systems, where the walltime for the calculation is approximately independent of molecular size if sufficient numbers of cpus are available.

The SMFA program is freely available from the GitHub website, *https://github.com/mickcollins/SMFAPAC*.

Acknowledgments

The authors acknowledge the support of the Australian NCI National Facility where these computations and program development were carried out. The authors also gratefully acknowledge helpful comments on the SMFA program from Dr Terry Frankcombe.

References

(1) Gordon MS, Fedorov DG, Pruitt SR, Slipchenko L. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. Chem. Rev. 2012; 112: 632-672.

(2) Collins MA, Bettens RPA. Energy-Based Molecular Fragmentation Methods. Chem. Rev. 2015; 115: 5607-5642.

(3) Deev V, Collins MA. Approximate ab initio energies by systematic molecular fragmentation. J. Chem. Phys. 2005; 122: 154102.

(4) Collins MA, Deev VA. Accuracy and efficiency of electronic energies from systematic molecular fragmentation. J. Chem. Phys. 2006; 125: 104104.

(5) Addicoat MA, Collins MA. Accurate treatment of non-bonded interactions within systematic molecular fragmentation. J. Chem. Phys. 2009; 131: 104103.

(6) Collins MA. Systematic fragmentation of large molecules by annihilation. Phys. Chem. Chem. Phys. 2012; 14: 7744-7751.

(7) Pruitt SR, Addicoat MA, Collins MA, Gordon MS. The fragment molecular orbital and systematic molecular fragmentation methods applied to water clusters. Phys. Chem. Chem. Phys. 2012; 14: 7752-7764.

(8) Reid DM, Collins MA. Molecular electrostatic potentials by systematic molecular fragmentation. J. Chem. Phys. 2013; 139: 184117.

(9) Collins MA. Molecular forces, geometries and frequencies by systematic molecular fragmentation including embedded charges. J. Chem. Phys. 2014; 141: 094108.

(10) Collins MA, Cvitkovic MW, Bettens RPA. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. Acc. Chem. Res. 2014; 47: 2776-2785.

(11) Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, Jensen JH,

et al. General atomic and molecular electronic structure system. J. Comput. Chem. 1993; 14: 1347-1363.

(12) Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian 09; Revision A.1. Gaussian Inc., Wallingford CT, 2009.

(13) Valiev M, Bylaska EJ, Govind N, Kowalski K, Straatsma TP, van Dam HJJ, et al. NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. Comput. Phys. Comm. 2010; 181: 1477-1489.

(14) Shao Y, Gan Z, Epifanovsky E, Gilbert ATB, Wormit M, Kussman J , et.al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. Mol. Phys. 2015; 113: 184-215.

(15) Aidas K, Angel C, Bak KL, Bakken V, Bast R, et. al. The Dalton quantum chemistry program system. WIREs - Comp. Mol. Sci. 2014; 4: 269-289.

(16) Stone AJ. *The Theory of Intermolecular Forces*. Clarendon, Oxford, 1996.

(17) Stone AJ. Distributed Multipole Analysis: Stability for Large Basis Sets. J. Chem. Theor. Comput. 2005; 1: 1128-1132.

(18) Banerjee A, Adams N, Simons J, Shepard, R. Search for Stationary Points on Surfaces. J. Phys. Chem. 1985; 89: 52-57.

(19) Birkholtz AB, Schlegel HB. Exploration of some refinements to geometry optimization methods. Theor. Chem. Acc. 2016; 135: 84.

(20) Bofill JM. Updated Hessian Matrix and the Restricted Step Method for Locating Transition Structures. J. Comp. Chem. 1994; 15: 1-11.

(21) VMD Visual Molecular Dynamics, http://www.ks.uiuc.edu/Research/vmd/

(22) Hehre WJ, Ditchfield R, Radom L, Pople JA. Molecular orbital theory of the

electronic structure of organic compounds. V. Molecular theory of bond separation. J. Am. Chem Soc. 1970; 92: 4796-4801.

(23) George P, Trachtman M, Bock CW, Brett AM. An alternative approach to the problem of assessing stabilization energies in cyclic conjugated hydrocarbons. Theor. Chim. Acta, 1975, 38: 121-129.

(24) Protein-Data-Bank, https://www.rcsb.org/