The Journal of Chemical Physics

Harmonizing accuracy and efficiency: A pragmatic approach to fragmentation of large molecules

Cite as: J. Chem. Phys. **149**, 064112 (2018); https://doi.org/10.1063/1.5036595
Submitted: 17 April 2018 . Accepted: 26 July 2018 . Published Online: 13 August 2018

Subodh S. Khire, Libero J. Bartolotti, and Shridhar R. Gadre

View Online Export Citation CrossMark

Harmonizing accuracy and efficiency: A pragmatic approach to fragmentation of large molecules

Subodh S. Khire,[1] Libero J. Bartolotti,[2] and Shridhar R. Gadre[1,a]

[1] *Interdisciplinary School of Scientific Computing, Savitribai Phule Pune University, Pune 411007, India*
[2] *Department of Physical and Computational Chemistry, East Carolina University, Greenville, North Carolina 27858, USA*

(Received 17 April 2018; accepted 26 July 2018; published online 13 August 2018)

Fragmentation methods offer an attractive alternative for *ab initio* treatment of large molecules and molecular clusters. However, balancing the accuracy and efficiency of these methods is a tight-rope-act. With this in view, we present an algorithm for automatic molecular fragmentation within Molecular Tailoring Approach (MTA) achieving this delicate balance. The automated code is tested out on a variety of molecules and clusters at the Hartree-Fock (HF)- and Møller-Plesset second order perturbation theory as well as density functional theory employing augmented Dunning basis sets. The results show remarkable accuracy and efficiency *vis-à-vis* the respective full calculations. Thus the present work forms an important step toward the development of an MTA-based black box code for implementation of HF as well as correlated quantum chemical calculations on large molecular systems. *Published by AIP Publishing.* https://doi.org/10.1063/1.5036595

I. INTRODUCTION

A goal of theoretical and computational chemistry is to predict the electronic structure, properties, and reactivity of molecular systems accurately, which could be used to supplement/complement experimental findings. Several Quantum Mechanical (QM) theories as well as Monte Carlo (MC), Molecular Dynamics (MD) simulation methods, etc. are available for achieving this. With the advents in computational hardware and software, sequential- and parallel codes are widely used by the scientific community in the form of black box packages. Among all these methods, QM investigations on molecular systems provide fundamental insights and molecular level understanding of the corresponding experimental results.

Widely used QM theories[1–4] include the Hartree-Fock (HF) method, Density Functional Theory (DFT), Møller-Plesset (MP) theory, Coupled Cluster (CC) method, etc. They mainly differ in the way of constructing the wave function/electron density for solving the Schrödinger equation approximately, yet accurately. The computational expense of a QM method can be gauged from the level of theory and number of basis functions, N. The HF and (most of the) DFT computations formally scale as $O(N^3$ or $N^4)$ whereas the correlated theories show a higher scaling. For instance, Møller-Plesset second order perturbation theory (MP2) and coupled cluster singles and doubles with perturbative triples, viz., CCSD(T) methods, scale as $O(N^5)$ and $O(N^7)$, respectively.[1,5–7] It should be noted that pursuing such QM investigations, especially the correlated ones, requires large computational resources. For instance, Xantheas and co-workers reported a CCSD(T)-level full calculation (FC) on the $(H_2O)_{16}$ cluster using 120 000 processors from CRAY XT5 partition at the ORNL Leadership Computing Facility.[8] Thus, probing larger molecular systems (containing ≥100 atoms) using such high-scaling correlated theories is a task which is either impossible or demands huge computational resources. However, doing such computations with sufficient accuracy using contemporary multi-core off-the-shelf hardware demands innovative approaches.

To overcome the computational power requirements and to cut down the scaling complexities of the methods, several fragmentation-based methods have been proposed. In the 1970's and 1980's, Christoffersen and co-workers[9] demonstrated the use of fragmentation methods for *ab initio* treatment of large biological molecules, albeit using only the HF theory with a sub-minimal basis-set. A decade later, the divide-and-conquer method proposed by Yang *et al.*[10] in 1992 and the molecular tailoring approach (MTA) suggested by Gadre *et al.*[11] in 1994 paved the way for treating large molecules by *ab initio* methods. This was followed by the development of many other fragmentation-based methods by several research groups. These included the Fragment Molecular Orbital (FMO) by research groups of Kitaura and Federov[12] in 1999; molecular fractionation with conjugate caps (MFCC)[13] by Zhang and co-workers in 2003; and the systematic molecular fragmentation (SMF)[14] method of Collins and co-workers in 2004. The generalized energy-based fragmentation (GEBF)[15] approach by Li and co-workers and Molecules in Molecules (MIM)[16] procedure by Raghavachari *et al.* in 2011 are among many other recent fragmentation-based studies. Comprehensive reviews of all these methods are available in the literature.[17–19] The central theme of the fragmentation-based methods is to divide the large parent molecular system into subsystems (fragments) and treat only the subsystems computationally in order to estimate the electronic properties of the parent system.

[a] gadre@unipune.ac.in

On the other hand, Pulay[20,21] proposed the method incorporating local and density fitting approximations. This technique is further extended for probing the electronic properties of large molecular systems by the groups of Werner and Neese.[22,23] To the best of the authors' knowledge, for a well-constructed density fitting basis set, density fitting errors are typically 0.15 kJ/mol or 0.06 mH per atom. However, it is stated that this error can be reduced by employing correcting expressions.[24,25] Nevertheless, all such local methods are essentially based on approximations and are thus prone to error vis-à-vis their full calculation (FC) counterparts.

As stated by Yuan *et al.*,[26] "fragmentation based methods basically differ in construction of subsystems (fragments)." The redundancy of computations and special skills needed for using a fragmentation method is pointed out by some workers.[27] Miliordos and Xantheas[28] expressed their skepticism about the fragmentation-based approaches when they pointed out the "need of doing full calculations of energy and molecular geometry for validating the fragment based approach." However, such a comparison with full calculations (FC) is impossible at a highly correlated level of theory *on any contemporary hardware* for molecular systems containing over ∼200 atoms and/or 10 000 basis functions. Does one simply resign and admit that such computations should not be attempted or adopt a pragmatic yet cautious approach toward fragmentation-based calculations that demonstrate an art-of-the possible?

Toward a positive end, some recent studies point out the lacunae in fragment-based methods and offer useful suggestions for their betterment. For instance, Gordon *et al.*[17] have stated the following in their review article on fragment-based methods: "Despite the very considerable progress, they remain underused; there may be several reasons for this. Some of the software developments are only locally implemented, making it difficult for most interested users to utilize the methods." Werner *et al.*[27] pointed out the ease of fragmentation schemes. However, they opined that such an approach requires "redundant calculations." Collins and Bettens[18] have brought to attention the need for fragmentation methods to move beyond the demonstration phase and to create as well as disseminate user-friendly programs. Raghavachari and Saha[19] have mentioned that "fragmentation schemes are user-specific requiring substantial user-intervention, precluding their use as a broad black box code."

Thus, to bring in the user-friendliness, a much-warranted step would be building a utility to generate good-quality fragmentation schemes automatically. Also, a clear-cut *priori* appraisal of the effect of arbitrary fragmentation on energetics is desirable. Such a utility and its benchmarking will be indeed helpful to an uninitiated user.

With this in view, the present work reports the development of an automatic fragmentation utility within the Molecular Tailoring Approach (MTA). Furthermore, the consistency in MTA energies vis-à-vis the respective full calculation (FC) ones at the prescribed level of theory and basis-set by employing automatically generated fragments of a reasonably good quality is also assessed. The benchmark calculations are reported at the HF, DFT, and MP2 level theory using large basis-sets. The benchmark set of systems consists of medium- to large-sized molecules/clusters comprising of ∼50 to 190 atoms and ∼1600 to 7300 basis functions.

II. METHODOLOGY

A. Molecular tailoring approach

The Molecular Tailoring Approach (MTA) was first proposed in our group in 1994 with a limited purpose of estimating the electrostatic properties of closed-shell large molecules.[11] Over the last two decades, the scope of MTA was extended for the estimation of molecular energy, geometry optimization, the estimation of the Hessian matrix, and the computation of vibrational infrared and Raman spectra.[29–33] As of now, the code has been thoroughly benchmarked for several large molecular systems including weakly bound molecular aggregates.[31–37] The MTA has the following intrinsic advantages. First, the fragment computations are independent to each other, making the procedure ideal for parallelization.[35,36] Second, the MTA code can be used in principle, in conjunction with any black box code[38–40] working at the back-end, leading to a plug-in version of MTA.[37] The basic idea of MTA is to virtually break a parent molecule into two or more overlapping sub-systems (fragments), until the latter become simple enough to be computed directly. Furthermore, a desired electronic property, P, of a parent molecule is estimated by patching the results from individual sub-systems (fragments) employing the set inclusion-exclusion principle, used as an approximation; *vide* the following equation:

$$P = \sum_{i} P^{F_i} - \sum_{i<j} P^{F_i \cap F_j} + \cdots + (-1)^k \sum_{i<j<\cdots<n} P^{F_i \cap F_j \cap \cdots \cap F_n}. \tag{1}$$

Here, P^{Fi} denotes the property associated with i-th fragment, $P^{Fi\cap Fj}$ represents the value of property P of the binary overlap between the i-th and j-th fragments, and k is the order of overlap between the fragments.

In 2012, a grafting correction[41] was introduced in order to reduce the error occurring due to the missing interactions within MTA. This procedure is indeed similar in spirit to that used in the methods[42,43] such as ONIOM, G2, etc. In this correction, the contribution of the missing interatomic interactions due to fragmentation is estimated from the difference between the energies of MTA and FC calculations at a suitable lower basis (LB) set keeping the fragmentation scheme unaltered, by using the following equation:

$$P^{HB} = P^{HB}_{MTA} + \left(P^{LB}_{FC} - P^{LB}_{MTA}\right). \tag{2}$$

Here, P^{HB} is the electronic property of the parent molecule estimated at a higher basis (HB) set after effecting the grafting correction, P^{HB}_{MTA} is the property computed by the MTA [using Eq. (1)] procedure at the HB. P^{LB}_{FC} and P^{LB}_{MTA} are properties computed at the LB set by doing the full and MTA [using Eq. (1)] calculation, respectively. In other words, the difference between the property of the parent molecule at LB without and with fragmentation is added to the MTA-based property of the parent molecule at HB. This procedure essentially hangs on our earlier observation[37] that the error does not show large dependence on the basis set employed.

B. Automatic fragmentation procedure

Within MTA methodology, the parent molecule can be fragmented in numerous ways. However, it needs a tailor's skill for making quality fragments leading to accurate estimation of electronic property vis-à-vis its FC counterpart. To circumvent this limitation, the automatic procedure for fragmenting the molecule within MTA was first developed by Babu and Gadre[44] followed by Gadre and Ganesh.[45] In this method,[45] at the first stage, the fragments centered at non-hydrogen atoms are generated by putting spheres of radius (in Å) RADCUT (Radius Cutoff) on these atoms. Thus, the total number of fragments made in the first round is equal to the total number of non-hydrogen atoms in the parent molecule. At later stages, these fragments are merged on the basis of a common atomic block between the fragments, until the values of the parameters MAXSZ (maximum allowed number of atoms in a fragment) and MINSZ (minimum allowed number of atoms in a fragment) are roughly achieved as per the intricacy of the parent molecule. While finalizing the scheme, the program ensures that each atom is present in at least one of the fragments. The fragments thus generated are termed as the main fragments, and the quality of the fragmentation scheme is defined in terms of a parameter called R-Goodness (Rg). The Rg of a fragmentation scheme is calculated by taking the following steps.

(i) Rg of an atom in a main fragment is the maximum radius of the sphere centered on that atom such that all the atoms lying in this sphere lie in the fragment.

(ii) If the reference atom is present in two or more fragments, an identical procedure is followed for all these fragments, which generates different radii, the maximum value of which is called as the overall Rg of the reference atom.

(iii) Likewise, the Rg value for every atom in the parent molecule in a fragmentation scheme is computed. The minimum of all these values is called as the Rg value of that scheme.

Thus, in-general, the larger the Rg value of a scheme, the better the chemical environment of each atom mimicked in that fragmentation scheme. This completes the process of generating the main fragments along with the Rg value of the scheme. With the use of these main fragments, the overlap fragments are generated and the corresponding order of overlap is calculated. In summary, by varying RADCUT, MAXSZ, and MINSZ, different fragmentation schemes are generated. This entire procedure will henceforth be termed as the fragmentor module.

In order to generate many fragmentation schemes for a given large closed-shell molecular system, we have written an external script. The script systematically varies the parameters RADCUT and MAXSZ. The parameter RADCUT is varied, in the case of molecular clusters, between 2.5 Å and 3.5 Å with an increment (ΔR) by 0.1 Å. The RADCUT range is chosen from 3.0 Å to 4.0 Å for molecular cases. Furthermore, for the given total number of atoms (NA), MAXSZ is varied within range NA/5 to NA/2 and MINSZ is fixed to NA/5. The increment ΔSZ for MAXSZ is NA/4-NA/5 for molecular systems, while it is taken as the total number of atoms in the monomer for a molecular cluster. However, the user can supersede the above prescribed default values as per the requirement and/or extent of a molecule/cluster. For each triplet of RADCUT, MINSZ, and MAXSZ, the fragmentor module is called for generating the fragments. It further provides the Rg of each atom from which the average Rg (ARg) for the scheme is calculated. The Economy Ratio (ER) for every scheme is approximately estimated; *vide* the following equation:

$$ER = \frac{\sum_{i=0}^{nfrag} N_i^j}{NA^j}. \tag{3}$$

Here nfrag is the total of fragments (main and overlap) generated and N_i is the total number of atoms in the i-th fragment. The value of j is taken as 3 in the case of HF/DFT and 5 in the case of MP2 calculations. The pool of schemes thus generated is assessed further on the basis of ER, Rg, ARg, size, and number of fragments. The procedure is depicted schematically in Fig. 1 and is expected to work for many "normal" molecules and all weakly bound molecular clusters. However, for molecules that are somewhat complex from the fragmentation point-of-view, one may use the in-house developed MetaStudio[46] package for effecting fragmentation. MetaStudio has a pencil tool with a graphical interface to facilitate the fragmentation procedure. The fragmentation schemes generated by this software can be easily fed into the fragmentor module for further processing, followed by MTA-based test calculations.

C. Screening and benchmarking of fragmentation schemes

With the help of the above procedure, a pool of fragmentation schemes is generated for the parent molecule. How does one pick reliable and efficient schemes from this pool? For this purpose, the fragmentation scheme pool is initially sorted with the best ER in descending order and top half schemes are selected at the first stage. This ensures the maximum efficiency of the chosen schemes. Selected schemes are further sorted with Rg as the primary key and ARg as the secondary key, in descending order. Finally, top 5 unique Rg schemes are finalized. In case there are less than 5 schemes with unique Rg, schemes with top ARg values are chosen. Cognizance of missed atom pair interactions which lie within a distance of 3.0–5.0 Å is also taken while making the selection. This completes the qualitative analysis of the schemes.

Moving on further, for making a final selection from this set of schemes, a further quantitative analysis in terms of energetics is carried out. For this purpose, a "grafting" correction is applied for medium-sized molecules, using Eq. (2), with 6-31G and 6-31+G(d) as LB and HB, respectively, at the selected level of theory. With this choice of HB and LB, computations (FC and MTA) can be easily carried out with a small wall-clock time employing a lower end hardware, viz., the core i3/i7-based machine. For larger parent systems, a smaller pair of basis-sets could be used. The absolute difference between MTA and FC energies at the 6-31+G(d) basis is used for making the choice of the "best" performing scheme. The final calculation at the actual higher basis set (for which FC is either extremely difficult or impossible with off-the-shelf hardware) is done with this scheme. This whole screening procedure ensures the faithfulness of the scheme qualitatively as well

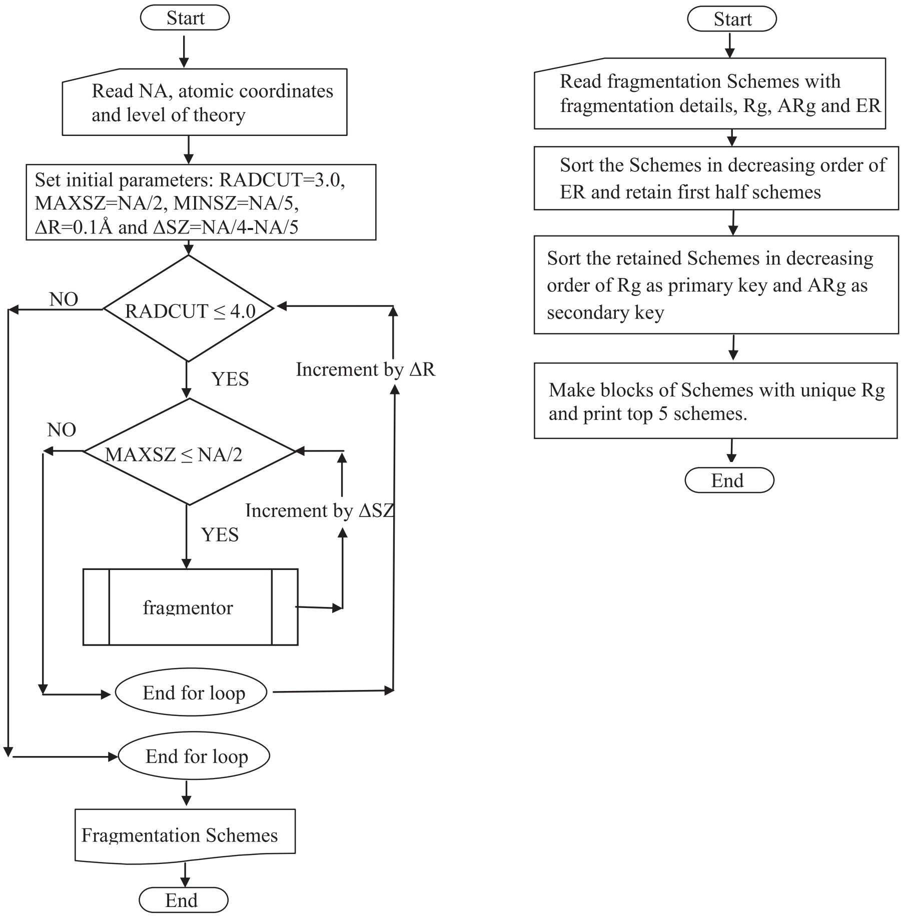
Start
Read NA, atomic coordinates and level of theory
Set initial parameters: RADCUT=3.0, MAXSZ=NA/2, MINSZ=NA/5, ΔR=0.1Å and ΔSZ=NA/4-NA/5
NO
RADCUT ≤ 4.0
YES
Increment by ΔR
NO
MAXSZ ≤ NA/2
Increment by ΔSZ
YES
fragmentor
End for loop
End for loop
Fragmentation Schemes
End
Start
Read fragmentation Schemes with fragmentation details, Rg, ARg and ER
Sort the Schemes in decreasing order of ER and retain first half schemes
Sort the retained Schemes in decreasing order of Rg as primary key and ARg as secondary key
Make blocks of Schemes with unique Rg and print top 5 schemes.
End

FIG. 1. Flow chart depicting an overview of the systematic automatic fragmentation scheme generation algorithm (left) using MTA guidelines and algorithm for screening of the schemes (right) for large molecules. The fragmentor is an external module for fragmentation. See text for details.

as quantitatively, finally converging on to one fragmentation scheme for the desired calculation at a higher basis-set.

III. RESULTS AND DISCUSSION

In order to automatically generate several fragmentation schemes and to *a priori* judge their efficacy and accuracy for calculating molecular energies, a variety of medium-to large-sized molecular systems are chosen. These test systems are assessed at the HF/DFT level of theory using high level Dunning as well as Pople basis sets, employing off-the-shelf hardware and Gaussian16-interfaced utility of MTA. A few sample calculations employing MP2 level theory are also reported. The choice of high-level Dunning basis-sets is made considering their suitability for wave function as well as DFT-based correlated calculations and their ease for basis-set extrapolation.

The test cases include chemically diverse closed-shell systems covering a broad range of large molecules and weakly bonded molecular clusters. Table I lists the test systems, along with the level of theory, the basis-set used (as HB), and the total number of associated basis functions. The initial geometries for the protein 1YJP, taxol, and vancomycin are taken from the published studies of Saha and Raghavachari,[47]

TABLE I. Test molecules/clusters, level of theory, number of atoms (NA), and number of basis functions (NBF) at aug-cc-pVNZ (aVNZ), where N = D, T, Q, respectively, and 6-31+G(d) basis set. See text for details.

			NBF			
Molecule/cluster	Level of theory	NA	aVDZ	aVTZ	aVQZ	6-31+G(d)
Pincer tweezer (PT)	B97D	98	1666	3542	6412	1148
1YJP	HF	107	1807	3845	6966	1238
Taxol	ωB97XD	113	1885	4025	7306	1280
α-cyclodextrin	B3LYP	126	2058	4416		1374
γ-cyclodextrin	B3LYP	168	2744	5888		1832
Vancomycin	B3LYP	176	3006	6379		2077
$(H_2O)_{64}$	ωB97XD	192	2624	5888		1472
$(H_2O)_{32}$	MP2	96	1312	2944		736
$(H_2O)_{16}$	MP2	48	656	1472	2751	368

TABLE II. Analysis of 5 fragmentation schemes for two molecules: R-Goodness (Rg); average Rg (ARg); full calculation (E_{FC}) and MTA (E_{MTA}) energies (a.u.) computed at the 6-31+G(d) basis set as HB with 6-31G as LB; and absolute difference (ΔE), in E_{FC} and E_{MTA} in mH. See text for details.

Molecule	Level of theory	Rg	ARg	E_{FC}	E_{MTA}	ΔE (mH)
1YJP	HF	2.73	4.38	−3140.767 47	−3140.766 30	1.17
		2.63	4.16		−3140.766 51	0.97
		2.50	4.27		−3140.766 83	0.65
		2.43	4.29		−3140.767 07	0.40
		2.38	4.07		−3140.766 36	1.11
α-cyclodextrin	B3LYP	2.71	3.85	−3664.638 89	−3664.641 59	2.70
		2.56	3.48		−3664.641 67	2.78
		2.53	4.44		−3664.642 00	3.11
		2.44	4.19		−3664.641 22	2.33
		1.91	3.89		−3664.641 69	2.80

Ganesh *et al.*,[36] and Bykov *et al.*,[48] respectively. Geometries of α-cyclodextrin and γ-cyclodextrin wherein sugar molecules are bound in a cyclic manner are taken from Ref. 49 and that of the pincer tweezer (PT) molecule is adopted from Ref. 50. In the case of DFT calculations, the B3LYP functional and two dispersion corrected functionals, viz., B97D and ωB97XD, are selected for benchmarking. Most of the geometries are optimized at the respective levels of theory using the 6-31+G(d) basis-set. The sources of other geometries are listed out separately later. The Cartesian coordinates and depiction for all the test cases are given in Table TS1 and Figure FS1 of the supplementary material, respectively.

To demonstrate the automatic fragmentation code and to probe the energetics of molecular clusters, some water clusters, viz., $(H_2O)_{16}$, $(H_2O)_{32}$, and $(H_2O)_{64}$, are selected. The geometries of $(H_2O)_{16}$, $(H_2O)_{32}$, and $(H_2O)_{64}$ are taken from the literature by Yoo *et al.*,[8] Sahu *et al.*,[32] and Yuan *et al.*,[26] respectively. As it is normally recommended to use correlated theory for water clusters, the first two clusters are explored at the MP2 level. The largest water cluster, viz., $(H_2O)_{64}$, is tested out employing the ωB97XD functional. It may be noted that the number of atoms in all the test cases range from 98 to 192 and the number of functions for the aug-cc-pVTZ (aVTZ) basis-set for all the molecular cases for HF/DFT calculations lies between 3542 and 6379. For MP2 level theory, systems with a maximum of 2944 basis-functions at aVTZ level are tested.

The screening of fragmentation schemes (outlined in Sec. II) is carried out, and the results are displayed for two sample test cases in Table II. For rest of the test cases, fragmentation details along with energetics are given in Tables TS2 and TS3 of the supplementary material.

It may be noticed that the schemes with sufficiently diverse Rg values lead to comparable ΔE-values for a given test case, out of which the scheme with the lowest one is chosen for a further higher level calculation. For example, for 1YJP as well as α-cyclodextrin, the scheme at serial number 4 turns out to be the "best" one *vis-à-vis* the FC at 6-31+G(d) basis set. These schemes are taken up for actual calculations employing larger basis-sets (aVDZ, aVTZ, etc.). The aim of this screening is to enable calculations at a higher level of theory and/or basis set with a balance between accuracy and efficiency. In the present work, with use of the "best" scheme, the energy calculation for all test molecules using the aVDZ basis set at the respective level of theory is performed. For the use of aVDZ as HB, the respective cc-pVDZ (VDZ) is selected as LB for the grafting correction. The hardware employed is mainly an

TABLE III. Energetics of the "best" fragmentation scheme for the test molecules: R-Goodness (Rg); average Rg (ARg); full calculation- and MTA energies (a.u.); the associated wall-clock times (min); and absolute difference (ΔE) in FC and MTA energies (mH), computed at aVDZ basis. See text for details.

				MTA		FC		
Name	Level of theory	Rg	ARg	Energy	Time	Energy	Time	ΔE
Pincer tweezer (PT)	B97D	2.47	4.53	−2636.662 95	114[a]	−2636.663 29	130[a]	0.34
1YJP	HF	2.43	4.29	−3141.154 50	59[a]	−3141.154 30	156[a]	0.20
Taxol	ωB97XD	2.15	3.98	−2929.050 96	168	−2929.053 34	361	2.38
α-cyclodextrin	B3LYP	2.44	4.19	−3665.113 38	61	−3665.114 16	174	0.78
γ-cyclodextrin	B3LYP	2.83	3.78	−4886.826 41	110	−4886.826 64	345	0.23
Vancomycin	B3LYP	3.41	6.11	−5779.482 29	384	−5779.481 55	517	0.74
$(H_2O)_{64}$	ωB97XD	2.97	4.16	−4891.506 64	173	−4891.507 49	385	0.85
$(H_2O)_{32}$	MP2	2.96	4.16	−2440.953 45	202	−2440.952 43	675	1.02
$(H_2O)_{16}$	MP2	2.82	3.70	−1220.451 34	10	−1220.451 08	78	0.26

[a] Wall-clock time on an i7-based computer.

TABLE IV. Energetics of the "best" fragmentation scheme for the test molecules: MTA and FC energies (a.u.), absolute difference (ΔE) in these energies (mH) computed at the aVTZ basis set, and wall-clock timings (min) on the 16-core computer. See text for details.

		MTA		FC		
Name	Level of theory	Energy	Time	Energy	Time	ΔE
Pincer tweezer (PT)	B97D	−2637.230 19	638[a]	−2637.230 43	853	0.24
1YJP	HF	−3141.831 78	849[a]	−3141.831 81	1 255	0.03
Taxol	ωB97XD	−2929.697 35	1439	−2929.698 79	4 044	1.44
α-cyclodextrin	B3LYP	−3665.991 92	577	−3665.992 20	2 035	0.28
γ-cyclodextrin	B3LYP	−4887.994 15	1107	−4887.994 02	10 054	0.13
Vancomycin	B3LYP	−5780.657 83	3437	−5780.657 74	10 568	0.08
$(H_2O)_{64}$	ωB97XD	−4892.767 40	1324	−4892.767 48	2 844	0.08
$(H_2O)_{32}$	MP2	−2443.113 31	4897[b]	−2443.112 98[c,d]	22 414[c,d]	0.33
$(H_2O)_{16}$	MP2	−1221.535 79	202	−1221.536 14[c,d]	1 178[c,d]	0.35

[a] Wall-clock time on an i7-based computer.
[b] The FC at LB done on a 20-core computer with the wall-clock time included.
[c] Calculations done on a 20-core computer.
[d] Calculations done using G09 package, whereas for all the others, G16 was used.

Intel-based 16 core machine with 32 GB memory, while some FC computations are done on Intel xeon processor based 20 core machines. Furthermore, for bringing out the power of MTA, 8 core machines are used for the first two test cases. Table III summarizes the energy and associated wall-clock timings for all the test molecules. For larger test cases, MTA is seen to reduce the wall-clock time for HF/DFT calculations typically by a factor of 1.5 to 3 in comparison with the respective FC ones. Furthermore, the energy differences for most of the test cases are in millihartree (mH). In the case of the taxol molecule, a somewhat larger error is observed, which may be attributed to the poor Rg of the fragmentation scheme and efficiency-based screening. For MP2-level calculations, the time advantage is substantial (cf. Table III), still retaining high accuracy of the energies.

With this impressive performance of MTA for obtaining accurate energies using limited hardware, we further benchmarked the energies of test cases at the aVTZ basis set as HB with cc-pVTZ (VTZ) as LB. Table IV displays the results for all the test molecules. With the enhancement in the basis set, the absolute difference between FC and MTA energies is also seen to be reduced. In the case of 1YJP, not only MTA energies are remarkable but also the wall-clock timings using a lower-end 8 core machine are quite impressive.

TABLE V. Energetics by employing the "best" fragmentation scheme for the test molecules: MTA energies (a.u.) and the associated wall-clock times (mins) for aVQZ basis. See text for details.

		MTA	
Name	Level of theory	Energy	Time
Pincer tweezer (PT)	B97D	−2637.403 23	8 900[a]
1YJP	HF	−3142.013 62	9 951[a]
Taxol	ωB97XD	−2929.903 46	21 678[b]
$(H_2O)_{16}$	MP2	−1221.896 23	3 439[b]

[a] Wall-clock time on an i7-based computer. See text for details.
[b] Wall-clock time on a 16-core computer. See text for details.

Full calculations on large molecule/clusters using the aug-cc-pVQZ (aVQZ) basis set are either impossible using off-the-shelf hardware or require huge wall-clock time. Enthused by the impressive MTA results obtained with aVDZ and aVTZ basis sets, we embark up on a study exploring the energetics of the test molecules by employing the aVQZ basis-set. Grafting correction is made employing cc-pVQZ (VQZ) as LB. For this purpose, we have chosen four of the test cases, viz., PT, 1YJP, taxol, and $(H_2O)_{16}$, as they represent four levels of theory, i.e., B97D, HF, ωB97XD, and MP2, respectively. Table V summarizes the results of MTA energies with aVQZ basis and the corresponding wall-clock timings for the chosen test cases. It is gratifying to note that MTA indeed enables these high-level computations using only the off-the-shelf hardware. We trust that the energies reported in Table V will match their FC counterparts within 1 mH when the latter become available.

IV. CONCLUDING REMARKS

In recent years, fragment-based methods have gained attention as they enable a*b initio* calculations on large molecular systems at a reduced computational cost. Apparently the arbitrary nature of fragments and lack of an *a priori* judgment of the accuracy of the results restrict the use of these methods. In the present work, a major step toward overcoming this restriction is taken within the well-entrenched molecular tailoring approach (MTA) for treating large molecular systems at the *ab initio* level. The prime innovation of this study is the development of an automated algorithm for generating a pool of fragmentation schemes by systematically varying three parameters, viz., RADCUT, MAXSZ, and MINSZ. The time-advantage of the schemes thus generated is gauged by the efficiency ratio (ER), and the accuracy is qualitatively appraised in terms of the parameters Rg and ARg. Use of these three parameters enables *a priori* shortlisting of a few schemes endowed with a balance of accuracy and efficiency. These shortlisted schemes are subjected to further quantitative analysis employing 6-31+G(d) basis as HB. The one giving the best agreement of the MTA energy with the FC one at 6-31+G(d) basis is recommended for further calculations at the desired larger basis-sets, e.g., aVDZ or aVTZ. It may be noticed that the screening procedure is a minor overhead and can be readily implemented on a low-end hardware, such as an i3/i7-based PC.

This algorithm was applied to a variety of medium-to large-sized molecular systems at HF, DFT, and MP2 level theory. The best fragmentation scheme is subjected to an MTA calculation with aVNZ as HB employing VNZ (N = D, T, Q) as LB. All calculations on test molecules show a very good time advantage for HF/DFT and a further impressive performance for MP2 theory. The accuracy of energy *vis-à-vis* the respective FC one is remarkable, with the error being 1 mH or less for most of the test cases examined. All these MTA calculations are performed on a relatively small hardware such as 8- or 16-core off-the shelf machines.

It is rather crucial to identify the LB for achieving better accuracy within MTA. We have found that aVNZ and VNZ basis-sets are made for each other for grafting purpose and so is the pair 6-31+G(d) and 6-31G. We have gone ahead and produced MTA-based results for some of the test molecules, viz., PT, 1YJP, taxol, and $(H_2O)_{16}$, at aVQZ basis set with B97D, HF, ωB97XD, and MP2 level of theories, respectively. We trust that the errors with the respective FC energy values will lie typically within the 1 mH error bar.

In the present study, we have reported test calculations at HF/DFT on molecular systems containing up to 192 (first/second-row) atoms and 6379 basis functions. Use of MTA would readily enable such calculations on systems with 500 atoms and/or 10 000 basis functions. For MP2 computations, the use of MTA would permit dealing with 200 atoms and/or 4000 basis functions. It may also be worth to test out the conjecture that MTA, in conjunction with local/density fitting methods, would enable MP2 calculations on even larger systems.

We have clearly demonstrated that the delicate balance between the accuracy and efficiency can be achieved with automated fragmentation. Our earlier work[32,35,36] has shown the effectiveness of MTA in geometry optimization and calculation of vibrational IR/RAMAN spectra for large molecules. We trust that the present work is indeed a crucial step toward the use of MTA as a black box code for the user community in quantum chemistry.

SUPPLEMENTARY MATERIAL

See supplementary material for Cartesian coordinates of molecules/clusters studied in the article along with the depiction of molecules and fragmentation details.

ACKNOWLEDGMENTS

S.R.G. is thankful to the Department of Science and Technology, Govt. of India for the award of J. C. Bose Fellowship (No. SR/S2/JCB-41/2006). The authors are thankful to the Center for Development of Advanced Computing (C-DAC) for providing computational facility and support under Project No. CDAC-CHE-20100163. Thanks are also due to the "Consortium for Computational Studies" at East Carolina University for supplying computational resources. Authors are grateful to the former members of the TCG Group at Pune for discussions.

[1] A. Szabo and N. S. Ostlund, *Morden Quantum Chemistry* (McGraw-Hill, New York, 1989).
[2] R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).
[3] I. N. Levine, *Quantum Chemistry* (Pearson Prentice Hall, New Delhi, 2009).
[4] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley, Weinheim, 2001); D. S. Sholl and J. A. Steckel, *Density Functional Theory, a Practical Introduction* (Wiley, Hoboken, 2009).
[5] D. L. Strout and G. E. Scuseria, J. Chem. Phys. **102**, 8448 (1993).
[6] W. J. Hehre, L. Random, and P. V. R. Schleyer, *Ab Initio Molecular Orbital Theory* (Wiley, New York, 1986).
[7] J. Almlöf, K. Faegri, and K. Korsell, J. Comput. Chem. **3**, 385 (1982).
[8] S. Yoo, E. Apra, X. C. Zeng, and S. S. Xantheas, J. Phys. Chem. Lett. **1**, 3122 (2010).
[9] D. Spangler and R. E. Christoffersen, Int. J. Quantum Chem. **17**, 1075 (1980).
[10] C. Lee and W. Yang, J. Chem. Phys. **96**, 2408 (1992); W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).
[11] S. R. Gadre, R. N. Shirsat, and A. C. Limaye, J. Phys. Chem. **98**, 9165 (1994).
[12] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi, Chem. Phys. Lett. **313**, 701 (1999).
[13] D. W. Zhang and J. Z. H. Zhang, J. Chem. Phys. **119**, 3599 (2003).
[14] V. Deev and M. A. Collins, J. Chem. Phys. **122**, 154102 (2005).
[15] S. H. Li, W. Li, and T. Fang, J. Am. Chem. Soc. **127**, 7215 (2005).
[16] N. J. Mayhall and K. Raghavachari, J. Chem. Theory Comput. **7**, 1336 (2011).
[17] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko, Chem. Rev. **112**, 632 (2012).
[18] M. A. Collins and R. P. A. Bettens, Chem. Rev. **115**, 5607 (2015).
[19] K. Raghavachari and A. Saha, Chem. Rev. **115**, 5643 (2015).
[20] P. Pulay, Chem. Phys. Lett. **100**, 151 (1983).
[21] S. Saebø and P. Pulay, J. Chem. Phys. **86**, 914 (1987).
[22] M. Schwilk, Q. Ma, C. Köppl, and H.-J. Werner, J. Chem. Theory Comput. **13**, 3650 (2017).
[23] P. Pinski and F. Neese, J. Chem. Phys. **148**, 031101 (2018).
[24] D. P. Tew, J. Chem. Phys. **148**, 011102 (2018).
[25] F. Weigend, Phys. Chem. Chem. Phys. **8**, 1057 (2006).
[26] D. Yuan, Y. Li, Z. Ni, P. Pulay, W. Li, and S. Li, J. Chem. Theory Comput. **13**, 2696 (2017).
[27] H.-J. Werner, G. Knizia, C. Krause, M. Schwilk, and M. Dornbach, J. Chem. Theory Comput. **11**, 484 (2015).
[28] E. Miliordos and S. S. Xantheas, J. Chem. Phys. **142**, 234303 (2015).
[29] M. Elango, V. Subramanian, A. P. Rahalkar, S. R. Gadre, and N. Sathyamurthy, J. Phys. Chem. **112**, 7699 (2008).
[30] A. P. Rahalkar, V. Ganesh, and S. R. Gadre, J. Chem. Phys. **129**, 234101 (2008).
[31] N. Sahu, S. D. Yeole, and S. R. Gadre, J. Chem. Phys. **138**, 104101 (2013).
[32] N. Sahu, S. S. Khire, and S. R. Gadre, Mol. Phys. **113**, 2970 (2015).
[33] N. Sahu, G. Singh, A. Nandi, and S. R. Gadre, J. Phys. Chem. A **120**, 5706 (2016).
[34] S. R. Gadre, S. D. Yeole, and N. Sahu, Chem. Rev. **114**, 12132 (2014).
[35] N. Sahu and S. R. Gadre, Acc. Chem. Res. **47**, 2739 (2014).
[36] V. Ganesh, R. K. Dongare, P. Balanarayan, and S. R. Gadre, J. Chem. Phys. **125**, 104109 (2006).
[37] A. P. Rahalkar, M. Katouda, S. R. Gadre, and S. Nagase, J. Comput. Chem. **31**, 2405 (2010).
[38] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari,

A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, GAUSSIAN 16, Revision A.03, Gaussian, Inc., Wallingford, CT, 2016.

[39] Package GAMESS, see M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, J. Comput. Chem. **14**, 1347 (1993).

[40] Package NWCHEM, see M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus, and W. A. de Jong, Comput. Phys. Commun. **181**, 1477 (2010).

[41] J. P. Furtado, A. P. Rahalkar, S. Shanker, P. Bandyopadhyay, and S. R. Gadre, J. Phys. Chem. Lett. **3**, 2253 (2012).

[42] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, and K. Morokuma, Chem. Rev. **115**, 5678 (2015).

[43] L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, J. Chem. Phys. **94**, 7221 (1991).

[44] K. Babu and S. R. Gadre, J. Comput. Chem. **24**, 484 (2003).

[45] S. R. Gadre and V. Ganesh, J. Theor. Comput. Chem. **5**, 835 (2006).

[46] V. Ganesh, J. Comput. Chem. **30**, 661 (2009).

[47] A. Saha and K. Raghavachari, J. Chem. Theory Comput. **11**, 2012 (2015).

[48] D. Bykov, T. Petreko, R. Izsak, S. Kossmaan, U. Becker, E. Valeev, and F. Neese, Mol. Phys. **113**, 13 (2015).

[49] M. M. Deshmukh, L. J. Bartolotti, and S. R. Gadre, J. Comput. Chem. **32**, 2996 (2011).

[50] S. Grimme, Chem. Eur. J. **18**, 9955 (2012).