Enabling \textit{ab initio} Hessian and frequency calculations of large molecules

Anuja P. Rahalkar, V. Ganesh,\textsuperscript{a) b) } and Shridhar R. Gadre\textsuperscript{a, b)}  

\textit{Department of Chemistry, University of Pune, Pune 411007, India}

(Received 28 April 2008; accepted 14 August 2008; published online 15 December 2008)

A linear scaling method, termed as cardinality guided molecular tailoring approach, is applied for the estimation of the Hessian matrix and frequency calculations of spatially extended molecules. The method is put to test on a number of molecular systems largely employing the Hartree–Fock and density functional theory for a variety of basis sets. To demonstrate its ability for correlated methods, we have also performed a few test calculations at the Møller–Plesset second order perturbation theory. A comparison of central processing unit and memory requirements for medium-sized systems with those for the corresponding full \textit{ab initio} computation reveals substantial gains with negligible loss of accuracy. The technique is further employed for a set of larger molecules, Hessian and frequency calculations of which are not possible on commonly available personal-computer-type hardware. © 2008 American Institute of Physics.

[DOI: 10.1063/1.2978387]

I. INTRODUCTION

The Hessian matrix, comprising of second order partial derivatives of the molecular energy with respect to nuclear coordinates, is one of the most commonly used entity for characterizing the potential energy surfaces of chemical systems and reactions. This can also be used for effecting speedier geometry optimization of molecules. Further, it is essential for computing the vibrational frequencies of a molecule and is also crucial for confirming the local minima or identifying the transition state. However, for \textit{ab initio} quantum chemical methods, computation of the Hessian matrix is currently an expensive proposition in terms of central processing unit (CPU) and memory requirements. For instance, evaluation of analytic Hessian for Hartree–Fock (HF) method formally scales\textsuperscript{1,2} as $O(N^4)$ for a system represented using $N$ basis functions. When the Hessian is computed numerically, as in the case of density functional theory (DFT), at least $3M+1$ energy and gradient evaluations are necessary for a system containing $M$ atoms.

To surmount this scalability issue, computational chemists have developed a variety of means and tools for fast evaluation of Hessian and vibrational frequencies at the \textit{ab initio} level of theory. For instance, Deglmann \textit{et al.}\textsuperscript{3} proposed various schemes for evaluating frequencies of large molecules using DFT and tested it on several linear molecules and aromatic sheets. Key features of their work involve efficient numerical quadrature, integral prescreening based on rigorous estimates, and exploitation of point group symmetry of the molecule under study. Benchmark results in this work\textsuperscript{3} show that molecules with up to 100 heavy atoms can be treated on a personal computer (PC). Comparisons with experimental values of infra-red (IR) shifts are also made. These techniques are available in the later versions of the TURBOMOLE (Ref. 4) package. Another package,\textsuperscript{5} Q-CHEM, incorporates a scalable Hessian evaluation code and has been shown to function well for medium-sized systems containing less than 100 atoms.

A recent paper by Alexeev \textit{et al.}\textsuperscript{1} discusses a new parallel algorithm for Hessian calculation and has been shown to scale well over a large number of processors for systems such as luciferin and capreomycin. The reported Hessian calculations in all the cases were at the HF level of theory. Their new distributed data analytic method is able to scale over large numbers of processors quite efficiently. Using an advanced HP IA64 based supercomputer interconnected by a very high speed network, the largest of their presented test cases, capreomycin with 778 basis functions, took about 75 min on 256 processors for the Hessian and frequency calculation at the HF/6-31G($d$) level of theory. Izmaylov and Scuseria\textsuperscript{6} also recently proposed a method for efficient evaluation of analytical vibrational frequencies in the HF and DFT frameworks for periodic systems. They also applied it for evaluating the vibrational frequencies of various periodic nonconducting systems such as boron nitride sheet and bulk diamond, mostly employing the 6-31G($d,p$) basis set. However, the methods presented in this paper\textsuperscript{7} are specifically meant for solids and periodic systems and thus employ periodic boundary conditions.

Neugebauer \textit{et al.}\textsuperscript{7} developed methods of approximations in calculations of IR and Raman frequencies and applied their method on buckminsterfullerene at various basis sets. Their study is probably the first computational study on IR and Raman spectra of $C_{60}$ molecule at varying basis sets using numerical Hessian derived from analytical gradients. In a recent study, Pathak and Rastogi\textsuperscript{8} were able to compute IR frequencies for a variety of polycyclic aromatic hydrocarbons (PAHs) at the B3LYP/4-31G level of theory and compared their results with the corresponding experimental ones. All of their calculations were performed on high performance machines using standard GAMESS (Ref. 9) package.

\textsuperscript{a)} Also at Interdisciplinary School of Scientific Computing, University of Pune.

\textsuperscript{b)} Electronic mail: gadre@chem.unipune.ernet.in.
One of the largest of the calculations that they performed were on PAH C_{96}H_{24} at the B3LYP/4-31G level of theory with 912 basis functions. Although the number of basis functions employed is indeed large, the authors’ use of symmetry, prevalent in the case of PAHs, considerably reduces the required computational cost.

Besley and Metcalf,\textsuperscript{10} in a recent interesting study, demonstrated the computation of frequencies from partial Hessians for amide I band of a few polypeptides and proteins at the HF and DFT levels of theory. Their method includes the effects of only carbon, oxygen, and nitrogen present on the backbone chain of amino acids. This approximation basically exploits the nature of the amide I band, which is largely localized on carbonyl groups of the backbone amide residues. The mean absolute error that they observed for a set of model polypeptides was of the order of 15 cm\(^{-1}\) at the HF/STO-3G level of theory. For the model test cases using their method, a time of 10\%–15\% of the calculation of the full Hessian was observed, which certainly is a large gain in terms of memory requirements. This method too is integrated with the Q-CHEM package.\textsuperscript{3}

To the best of the authors’ knowledge, one of the largest frequency calculation using standard \textit{ab initio} package\textsuperscript{31} is reported in a recent paper by Requena \textit{et al.}\textsuperscript{12} They reported frequency calculation of \(\beta\)-carotene and its derivatives, viz., capsanthin and capsorubin, at the B3LYP/6-311G\((d,p)\) level of theory, involving a maximum of 1096 basis functions.

Though parallelization of conventional Hessian calculation procedure does make it easier to handle larger molecules, it in no way addresses the core issue of nonlinear scaling. This practically limits the usability of these methods to systems with typically less than 800 basis functions even scaling. This practically limits the usability of these methods for molecules, it in no way addresses the core issue of nonlinear computation. Therefore, the technique presented here is applicable to other levels of theory also. This is illustrated with a couple of test cases at the Moller–Plesset second order perturbation theory (MP2). For all the calculations presented here, the parallel CG-MTA scheme is used.\textsuperscript{32} This code is developed in line with various other codes developed earlier\textsuperscript{34–36} in our laboratory.

\section*{II. METHODOLOGY}

As stated earlier, instead of carrying out a full \textit{ab initio} calculation on a spatially extended large system, CG-MTA breaks it down into a series of calculations on a set of overlapping fragments. A detailed description of the CG-MTA scheme can be found elsewhere,\textsuperscript{31–33} while the most important steps are briefly described here. The method starts with the generation of a set of fragments using the requested parameters, viz., \(R\) goodness and maximum allowed size of a fragment.\textsuperscript{16,31–32} Due to the possibility of a multiple fragmentation scheme, effective implementation of CG-MTA requires a quantitative definition of the quality of a given fragmentation scheme. This is achieved by defining minimum \(R\) goodness of a fragmentation scheme, details of which can be found elsewhere.\textsuperscript{16,31–33} A better \(R\)-goodness value generally translates to better quantitative estimates of energy and gradients. Our earlier benchmarks have shown that generally an \(R\) goodness of 3–4 Å produces sufficiently accurate results (up to millihartree accuracy for energy). Once a fragmentation scheme is chosen, the total energy \((E)\) along with its derivatives with respect to the Cartesian coordinates of a supermolecule represented using \(k\) fragments can be estimated based on the cardinality expression as given below:

\[
E = \sum_i E^f_i - \sum_i \sum_j E^{f(i)f(j)} + \cdots + (-1)^{k-1} \sum_i \sum_j \cdots \sum_k E^{f(i)f(j)\cdots f(k)}. \tag{1}
\]

Here, \(E^f_i\) refers to the energy of fragment \(f_i\), \(E^{f(i)f(j)}\) refers to the energy of the overlap fragment \(f_i \cap f_j\), and so on. Using the above expression for energy, the gradients can now be estimated as

\[
\frac{\partial E}{\partial X^f_\mu} = \sum_i \frac{\partial E^f_i}{\partial X^f_\mu} - \sum_i \sum_j \frac{\partial E^{f(i)f(j)}_\mu}{\partial X^{f(i)f(j)}_\mu} + \cdots + (-1)^{k-1} \sum_i \sum_j \cdots \sum_k \frac{\partial E^{f(i)f(j)\cdots f(k)}_\mu}{\partial X^{f(i)f(j)\cdots f(k)}_\mu}. \tag{2}
\]

Note that in the above equation, \(X^f_\mu\) and \(X^{f(i)f(j)}_\mu\) refer to the
coordinates of atom \( \mu \) in the fragment \( f_i \) and the overlap fragment \( f_i \cap f_j \), etc.

A similar expression is used for computing the \((\mu, \mu')\)th element of the Hessian matrix:

\[
H_{\mu\mu'} = \sum H_{\mu\mu'}^{f_i} - \sum H_{\mu\mu'}^{f_i \cap f_j} + \cdots + (-1)^{k-1} \sum H_{\mu\mu'}^{f_i \cap f_j \cdots f_k}.
\]

Here, \( H_{\mu\mu'} \) refers to the Hessian element corresponding to second derivative of energy with respect to coordinates \( X_\mu \) and \( X_{\mu'} \), i.e., \( \partial^2 E / \partial X_\mu \partial X_{\mu'} \), and so on.

Once the Hessian matrix is set up employing the CG-MTA scheme [Eq. (3)], the corresponding vibrational frequencies are extracted by diagonalizing the mass-weighted Hessian matrix \( H_m \):

\[
H_m = M^{-1/2} H M^{-1/2},
\]

\[
H_{\mu\mu} L_k = \lambda_k L_k.
\]

Here, \( M \) is the atomic mass array, and \( L_k \) and \( \lambda_k \) are the displacements and frequencies, respectively.

### III. RESULTS AND DISCUSSION

The majority of the calculations reported here have been performed on a cluster of Intel Core2 Quad processor at 2.4 GHz with 4 Gbyte random access memory (RAM) and 250 Gbytes of disk each. A few of the calculations reported in this work are performed on a cluster of Pentium Dual Core at 2.8 GHz with 2 Gbytes of RAM and 80 Gbytes of local disk. A distributed version of the CG-MTA code\(^{16,31,32}\) is used for all the calculations owing to the embarrassingly parallel nature of the algorithm.

Figure 1 displays the structures of test cases investigated in this paper. Table I summarizes the molecular systems, along with the level of theory, basis set used, and number of basis functions involved in the calculation. The spatially extended molecules used as test cases are chemically and biologically diverse in nature and the calculations are performed without enforcing symmetry. The fragmentation details of the systems used as test cases are given in Table I. For all the test cases, a fragmentation scheme with a minimum \( R \) goodness of at least 3 Å is employed to ensure the reliability of the calculations performed.\(^{16,33}\) The largest fragment size and the scaling factor (see Table I) provide an indirect measure of
that this factor should be less than 5. It should be pointed out
resources as well as time required for the calculation. For
have substantial chemical or biological significance. A fat-
to-water molecules, especially if a smaller basis set is employed.
size is more than half of the parent molecule. This is due to
tum chemical package, GAMESS. Even in such cases, our
the “actual calculation” henceforth
than 600 basis functions at the specified level of theory, mak-

The first three systems in Table I, viz., the cluster of 37
water molecules, α-tocopherol, and folic acid, are small but
have substantial chemical or biological significance. A fat-
soluble antioxidant, α-tocopherol is the form of vitamin E
that is preferentially absorbed in humans. Folic acid is a form
of water soluble vitamin B. All these systems involve less
than 600 basis functions at the specified level of theory, making
it possible to run the full frequency calculation (called the “actual
calculation” henceforth) using a standard quantum
chemical package, GAMESS. Even in such cases, our
method offers notable advantage in terms of computational
resources as well as time required for the calculation. For
instance, in the case of α-tocopherol, the total time taken for
this calculation is about 11 h with CG-MTA as against the
staggering 50 h taken for the actual job on the same hard-

<table>
<thead>
<tr>
<th>System</th>
<th>[level/basis]</th>
<th>NA (NB)</th>
<th>NF</th>
<th>AV</th>
<th>LF</th>
<th>GP, GN</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H₂O)₃₇</td>
<td>[HF/6-31G]</td>
<td>111 (481)</td>
<td>8</td>
<td>54</td>
<td>63 (273)</td>
<td>4.1, 4.2</td>
<td>4.9</td>
</tr>
<tr>
<td>α-tocopherol</td>
<td>[HF/6-31G(d)]</td>
<td>81 (565)</td>
<td>6</td>
<td>38</td>
<td>44 (322)</td>
<td>4.5, 4.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Folic acid (a)</td>
<td>[B3LYP/6-31G]</td>
<td>51 (326)</td>
<td>4</td>
<td>28</td>
<td>33 (213)</td>
<td>4.1, 4.1</td>
<td>2.0</td>
</tr>
<tr>
<td>Folic acid (b)</td>
<td>[B3LYP/6-31G(d, p)]</td>
<td>51 (575)</td>
<td>4</td>
<td>28</td>
<td>33 (375)</td>
<td>4.1, 4.1</td>
<td>2.0</td>
</tr>
<tr>
<td>(H₃BO₃)₁₀</td>
<td>[HF/6-31++G(d, p)]</td>
<td>280 (3670)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scheme 1 | 60 | 23 | 28 (376) | 3.0, 3.0 | 5.0 |
| Scheme 2 | 40 | 26 | 28 (376) | 4.1, 4.1 | 3.8 |
| Cholesterol | [HF/6-31G(d)] | 74 (650) | 4  | 40 | 42 (370) | 3.4, 3.8 | 2.0 |
| Capreomycin | [HF/6-31+G(d, p)] | 91 (1113) | 11 | 37 | 43 (509) | 3.7, 3.7 | 3.9 |
| α-cyclodextrin | [B3LYP/6-31G(d)] | 126 (1110) |    |    |    |        |    |
| Scheme 1 | 22 | 30 | 36 (280) | 3.3, 3.4 | 4.3 |
| Scheme 2 | 12 | 38 | 41 (329) | 3.3, 3.7 | 3.2 |
| Scheme 3 | 6  | 55 | 55 (461) | 3.6, 3.7 | 2.5 |
| β-carotene (a) | [HF/6-31G(d)] | 96 (712) |    |    |    |        |    |
| Scheme 1 | 10 | 29 | 32 (235) | 3.9, 3.9 | 2.7 |
| Scheme 2 | 7  | 32 | 36 (254) | 4.1, 4.2 | 2.2 |
| β-carotene (b) | [B3LYP/6-31G(d)] | 96 (712) |    |    |    |        |    |
| Scheme 1 | 5  | 39 | 42 (318) | 4.2, 4.2 | 1.9 |
| Scheme 2 | 5  | 36 | 36 (280) | 4.0, 4.0 | 1.8 |

the computational cost of a fragmentation scheme. For
substantial computational advantage, earlier studies 16,33 found
that this factor should be less than 5. It should be pointed out
here that for folic acid and α-tocopherol the largest fragment
size is more than half of the parent molecule. This is due to
the inherently small size of the parent molecule and hence
computational advantage offered by CG-MTA may be small
especially if a smaller basis set is employed.

The first three systems in Table I, viz., the cluster of 37
water molecules, α-tocopherol, and folic acid, are small but
have substantial chemical or biological significance. A fat-
soluble antioxidant, α-tocopherol is the form of vitamin E
that is preferentially absorbed in humans. Folic acid is a form
of water soluble vitamin B. All these systems involve less
than 600 basis functions at the specified level of theory, making
it possible to run the full frequency calculation (called the “actual
calculation” henceforth) using a standard quantum
chemical package, GAMESS. Even in such cases, our
method offers notable advantage in terms of computational
resources as well as time required for the calculation. For
instance, in the case of α-tocopherol, the total time taken for
this calculation is about 11 h with CG-MTA as against the
staggering 50 h taken for the actual job on the same hard-
ware (see Table II). It is also worth noting that the maximum
memory required per node for this job is about 600 Mbytes
for CG-MTA while that for the actual job is 1.8 Gbytes. In
terms of the accuracy of the Hessian computed using CG-
MTA as compared to the actual one, the maximum error is
1.8 × 10⁻⁴ a.u. and root mean square deviation (RMSD) is
1.0 × 10⁻⁴ a.u. for α-tocopherol (see Table III). The other
entries from Table III also show that CG-MTA enables the
computation of Hessian and vibrational frequencies without
significant loss of accuracy.

Important frequencies for α-tocopherol are O—H stretching at 3604.3 cm⁻¹ (3604.4 cm⁻¹ for CG-MTA) and
C—O stretching at 1111.8 cm⁻¹ (1111.7 cm⁻¹ for CG-
MTA). For (H₂O)₃₇, a wide range of O—H stretching frequencies from 3173 to 4106 cm⁻¹ is obtained due to the
extensive hydrogen bonding between the water molecules. The
important frequencies for folic acid are the N—H stretches at 3766.3 and 3621.6 cm⁻¹, O—H stretches in the range of
3715—3762 cm⁻¹, three C==O stretches at 1750.8, 1835.3,
and 1847.6 cm⁻¹, while C—N stretching frequency is observed
at 1677.5 cm⁻¹. For all these cases, the difference
between CG-MTA and actual frequencies is seen to be within

<table>
<thead>
<tr>
<th>System</th>
<th>Level/basis</th>
<th>NB</th>
<th>No. of cores</th>
<th>Actual</th>
<th>CG-MTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H₂O)₃₇</td>
<td>HF/6-31G</td>
<td>481</td>
<td>16</td>
<td>1180</td>
<td>172</td>
</tr>
<tr>
<td>α-tocopherol (d)</td>
<td>HF/6-31G</td>
<td>565</td>
<td>6</td>
<td>3024</td>
<td>666</td>
</tr>
<tr>
<td>Folic acid (a)</td>
<td>B3LYP/6-31G</td>
<td>326</td>
<td>16</td>
<td>450</td>
<td>270</td>
</tr>
<tr>
<td>Folic acid (b)</td>
<td>B3LYP/6-31G(d, p)</td>
<td>575</td>
<td>16</td>
<td>4320</td>
<td>960</td>
</tr>
</tbody>
</table>

The cores are of Pentium Dual core at 2.8 GHz.
1 cm\(^{-1}\). In general, when the complete range of frequencies is compared, the maximum deviation is less than 0.5\% (see Table III).

We next present a few larger test cases for which the actual Hessian and frequency calculations cannot be readily performed on the available hardware. Out of these cases, cholesterol, capreomycin, and \((\text{H}_3\text{BO}_3)_4\) \((\text{a nanotube of 40 orthoarobic acid units involving extensively hydrogen bonded network})\) are treated at the HF level, while \(\alpha\)-cycloextrin is treated at the DFT level of theory, employing the basis sets specified in Table I. CG-MTA-optimized geometry of cholesterol is subjected to frequency analysis. Chemically, cholesterol is a combination of steroid and alcohol and plays an important role in the cell membrane of animal tissues. Frequency corresponding to the O—H stretch is found at 4179.0 cm\(^{-1}\). Similarly, for capreomycin, a scheme with 11 main fragments is used for CG-MTA geometry optimization. Capreomycin is a peptide antibiotic and is used in combination with other antibiotics for treating tuberculosis. The characteristic frequencies of C==O and N—H in peptide linkages are obtained in ranges of 1885.6–1950.1 and 3782.2–3927.4 cm\(^{-1}\), respectively, while the O—H stretch is obtained at 4197.4 cm\(^{-1}\). The frequency analysis of both the cases confirms the local minimum structure with all the real frequencies.

For \((\text{H}_3\text{BO}_3)_4\), a CG-MTA based optimized geometry\(^{37}\) is taken for frequency analysis. This is carried out using two different fragmentation schemes to appraise the internal consistency of the CG-MTA results. These schemes have \(R\)-goodness values of 3 and 4.1 Å, respectively. The results (in terms of the Hessian and IR frequencies) obtained from both the schemes are found to be in excellent agreement (see Table III) with each other. The frequencies from both the schemes are real and match to within 0.5 cm\(^{-1}\). The time taken by scheme 1 is 25 h on five dual core machines with 2 Gbyte RAM each, while that for scheme 2 is 12 h on 13 nodes of the same configuration. It is worth noting here that the number of basis functions involved in these calculations is 3760, for which actual calculations cannot at all be performed on any commonly available hardware resources. For \((\text{H}_3\text{BO}_3)_4\), frequencies within the range of 3970–4190 cm\(^{-1}\) correspond to O—H stretches, while those in the range of 1120–1600 cm\(^{-1}\) are due to B—O stretches.

Within DFT, the Hessian evaluation is generally done numerically. Hence, even if one wants to use a less accurate two-point formula, it is necessary to evaluate energy and gradients at \(3N_a+1\) steps, where \(N_a\) is the number of atoms in the system. Therefore, the conventional computation of Hessian at the DFT level for a large molecular system consisting of more than 100 atoms is very expensive and time consuming. However, within CG-MTA, the calculations are performed on individual fragments (say, \(n_i\) atoms) in which, being much smaller than the original molecule \((N_u\) atoms), the number of steps drastically decrease since \(3n_i+1 \approx 3N_a+1\). Moreover, the fragments being smaller, the effort required for energy-gradient evaluation at each step is also much less than that for the whole molecule.

A cyclic oligosaccharide composed of six glucose units, viz., \(\alpha\)-cycloextrin, is chosen as the next test case at the B3LYP/6-31G(d) level. For CG-MTA-based frequency analysis, \(\alpha\)-cycloextrin is fragmented with three different schemes. Of these schemes, scheme 3 is the best one with a minimum \(R\) goodness of 3.6 Å as both the other schemes have minimum \(R\) goodness of 3.3 Å. Comparisons of the results obtained from schemes 1 and 2 with those from scheme 3 are displayed in Table III, which evidently bring out self-consistency between the results obtained by the method. The O—H stretching frequencies are obtained in the range of 3562–3745 cm\(^{-1}\), while C==O stretching frequencies are in the range of 1108–1111 cm\(^{-1}\) from all the three schemes.

Finally, to probe the applicability of CG-MTA for Hessian calculation at the MP2 level, a couple of small test cases, viz., folic acid and \(\beta\)-carotene, are initially subjected to CG-MTA geometry optimization at the MP2/STO-3G level of theory. These optimized geometries are then used for Hessian and frequency calculation. Folic acid has 179 basis functions for the STO-3G basis, allowing calculations to be performed both by CG-MTA and the conventional method. The actual Hessian and frequency calculation took 44.5 h on...
a Pentium IV at 2.8 GHz with 1 Gbyte of RAM, while CG-MTA calculation took 15.4 h on the same hardware. The maximum memory required for actual calculation is 450 Mbytes while that for CG-MTA is merely 85 Mbytes. In terms of the accuracy of the Hessian matrix, the maximum error is $1.3 \times 10^{-2}$ and all the frequencies (in cm$^{-1}$) are correct to all places before the decimal point. The above geometry is further optimized at the MP2/3-21G level of theory with CG-MTA and then subjected to frequency calculation. This calculation involved 326 basis functions and hence an actual calculation is not possible with the above mentioned hardware. In order to prove the stability of the method at the MP2 level, two different fragmentation schemes with minimum goodness values of 4.0 and 4.8 Å are used for this calculation. The maximum deviation in the Hessian elements is 0.049 a.u. and RMSD is 6.9 Å.

The maximum goodness values of 4.0 and 4.8 Å are used for this calculation. The maximum deviation in the Hessian elements is 0.049 66 and RMSD is 6.9 Å. The calculation involved 326 basis functions and hence an actual calculation is not possible with the above mentioned hardware. In order to prove the stability of the method at the MP2 level, two different fragmentation schemes with minimum goodness values of 4.0 and 4.8 Å are used for this calculation. The maximum deviation in the Hessian elements is 0.049 66 and RMSD is 6.9 Å. The calculation involved 326 basis functions and hence an actual calculation is not possible with the above mentioned hardware. In order to prove the stability of the method at the MP2 level, two different fragmentation schemes with minimum goodness values of 4.0 and 4.8 Å are used for this calculation. The maximum deviation in the Hessian elements is 0.049 66 and RMSD is 6.9 Å.

The CG-MTA calculations are performed using two different fragmentation schemes with $R$-goodness values of 3.9 and 4.1 Å, respectively. The time advantage offered by CG-MTA is worth pointing out here. CG-MTA-based calculations using both of these schemes are completed within 50 min on 20 cores. The results from both the schemes are consistent with each other and the maximum error in the frequencies obtained using these two schemes is 2.7 cm$^{-1}$. These results bring out the applicability of CG-MTA for higher correlated methods, which is being pursued further in our laboratory.

### IV. CONCLUDING REMARKS

The present work has explored the use of CG-MTA for Hessian and frequency calculation of large molecular systems. For the test cases we have presented, the CG-MTA based Hessian is observed to be highly accurate when compared to its actual counterpart. The typical maximum error in any Hessian element is of the order of only $10^{-3}$ a.u. All the frequencies derived from the CG-MTA based Hessian are generally found to be correct up to 0.5%. In terms of memory and CPU requirements, a substantial savings is seen while comparing a CG-MTA calculation with the conventional one.

The test cases presented in this work clearly show that CG-MTA indeed enables calculations of the Hessian matrix and vibrational frequencies that are otherwise not possible on contemporary PC-based hardware without significant loss of accuracy. The Hessian and frequency calculations, without any symmetry constraints, for a large molecule such as $(\text{H}_2\text{BO}_3)_{40}$ with 3670 basis functions, to the best of our knowledge, is the largest reported one. In the future, this technique could be extended for application of Hessian for geometry optimization or locating transition states, etc. Although the current study is mainly restricted to the HF and DFT levels of theory, we have also shown its applicability to correlated theories (MP2). Further studies with the MP2 level of theory are currently under way in our laboratory.

**Note added in proof.** A related work on geometry optimization and vibrational spectra of large molecules by Hua et al. appeared after the present manuscript was accepted for publication. This work has discussed the vibrational frequencies along with IR and Raman intensities for 12-mer of glycine and 28-mer of water. A more detailed comparison of our work with that of Hua et al. will be made in a forthcoming publication.

### ACKNOWLEDGMENTS

Financial support from the Center for Development of Advanced Computing (C-DAC), Pune, Naval Research Board (NRB), New Delhi, and Council of Scientific and Industrial Research (CSIR), New Delhi is gratefully acknowledged. The authors are also thankful to Dr. Michael Schmidt from Iowa State University for useful suggestions.

---