JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 1383−1394

pubs.acs.org/JCTC

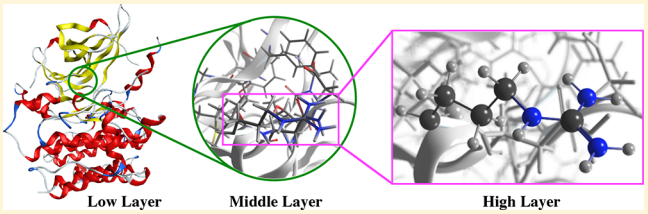
Assessment of Fragmentation Strategies for Large Proteins Using the Multilayer Molecules-in-Molecules Approach

Bishnu Thapa, Daniel Beckett, K. V. Jovan Jose, and Krishnan Raghavachari*

Department of Chemistry, Indiana University, Bloomington 47405, Indiana, United States

Supporting Information

ABSTRACT: We present a rigorous evaluation of the potential for the multilayer Molecules-in-Molecules (MIM) fragmentation method to be applied to large biomolecules. Density functional total energies of a test set of 8 peptides, sizes ranging from 107 to 721 atoms, were evaluated with MIM and compared to unfragmented energies to help develop a protocol for the treatment of large proteins. Fragmentation schemes involving subsystems of 4 to 5 covalently bonded fragments (tetramer or pentamer schemes) were tested with a single level of theory (MIM1) and produced errors on the order of 100 kcal/mol due to the relatively small size of the subsystems and the neglect of nonbonded interactions. Supplementing the two schemes with nonbonded dimer subsystems, formed from fragments within a specified cutoff distance (3.0 Å), nearly cut the MIM1 errors in half, leading us to employ these new schemes as starting points in multilayer calculations. When employing a DFT low level with a substantially smaller basis set (MIM2), the dimer-supplemented schemes produce errors below our target accuracy of 2 kcal/mol in the majority of cases. However, for the larger test systems, such as the 45 residue slice of a human protein kinase with over 10,000 basis functions in the high level, the low level calculation over the full molecule becomes the bottleneck for MIM2 calculations. To overcome any associated limitations, we explored, for the first time, 3-layer MIM methods (MIM3) with a distance-based medium level of fragmentation and dispersion-corrected semiempirical methods (e.g., PM6-D3) as the low level. A modestly sized cutoff distance in the medium level (3.0−3.5 Å), leading to subsystems of 30−50 atoms treated at the medium and low levels of theory, was able to match the low errors of the MIM2 calculations. These results allow us to develop a general prescription for 3-layer calculations wherein a much cheaper low level can be used, while fragment sizes in the high layer stay modest, allowing the MIM method to be applied to very large proteins in the future.



1. INTRODUCTION

Ab initio quantum mechanical (QM) calculations are being increasingly used to compute the structures and energies of large biomolecular systems, such as DNA or proteins. QM calculations on these species provide a much more accurate description of their physicochemical properties compared to the more routinely used force-field-based molecular mechanics (MM) methods.[1−8] In particular, many nontraditional interactions such as C−H···O hydrogen bonds or halogen-π interactions, while once thought to be weak, have been shown to be essential components in determining the structures of large biomolecules and require an accurate quantum mechanical treatment.[6−17] Unfortunately, the computational scaling of traditional quantum mechanical methods (ranging from N^4 to N^7 for the most widely used methods where N is a representation of system size) limits their applicability.[18] Even the most basic and computationally inexpensive methods such as Hartree−Fock (HF) and density functional theory (DFT) become impractical for large biological systems with thousands of atoms. Moreover, highly accurate, post-Hartree−Fock methods such as coupled−cluster (CC) methods are only applicable to molecules containing a few tens of atoms.[19−22]

The fragmentation approach is an efficient and attractive technique for lowering the steep scaling of conventional electronic structure methods to a linear (or near linear) scaling.[20−22] A range of fragmentation-based QM or QM/MM methods for large systems has been proposed and implemented over the past two decades by different research groups.[23−80] While all methods use a fragmentation-based determination of the energy, some apply a fragmentation scheme for the wave function also.[77−79] Relevant to the discussion at hand, Hua et al. have applied the generalized energy-based fragmentation (GEBF) to four proteins and a DNA decamer, containing 130 to 638 atoms, and achieved a mean absolute error (MAE) of 3.9 kcal/mol at the HF/6-31G* level of theory. However, the largest subsystem in this work had to reach 189 atoms to obtain viable results with the DNA decamer.[46] He and co-workers calculated the total energy of a benchmark set of 18 protein structures containing 242−1142 atoms using electrostatic embedded generalized molecular fractionation with conjugate caps (EE-GMFCC). For this 18 protein test set, a MAE of 2.4 kcal/mol was obtained for HF/6-31G* with a distance threshold of 4.0 Å and a maximum fragment size of 70

Received: November 29, 2017
Published: February 16, 2018

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

atoms. At the B3LYP/6-31G* level of theory, the MAE was calculated to be 5.4 kcal/mol with a maximum error of 7.6 kcal/mol for six proteins with between 218 and 608 atoms.[44] Recently, Liu and Herbert applied their pair−pair approximation to the generalized many-body expansion (GMBE) method to a similar set of 18 protein molecules as used by He and co-workers. The calculated MAE was 2.7 kcal/mol with a maximum deviation of 7.6 kcal/mol at the ωB97X-D/6-31G* level when using NPA embedding charges and subsystem sizes of ≤60 atoms, while the HF/6-31* MAE was even smaller.[80] For a comprehensive description of the current developments in fragmentation techniques, their advantages, and their application, we refer the reader to some recent reviews.[20−22] Although the terminologies and descriptions of the fragmentation-based methods may vary, they all share the same basic principle of dividing a large molecule or cluster into smaller nonoverlapping fragments (sometimes called "monomers"), assembling them into overlapping subsystems, evaluating them independently while taking care of any overcounting, and assembling properties for the full system, such as the electronic energy and its derivatives, from the subsystems.

The fragmentation approach is based on the assumption that most chemical properties are local in nature and only slightly influenced by groups far from the region of interest.[44,81,82] Hence, fragmentation methods employ some cutoff parameters (e.g., number of monomers, spatial distance) to obtain the subsystems that directly include only a limited number of interactions between the fragments, neglecting the longer range effects. Although the individual contributions from long-range, noncovalent interactions are expected to be rather small, their collective contributions can be substantial. This is particularly true for large and globular biomolecules, such as DNA and proteins, whose secondary structures are rich with nonbonded interactions. The deficiencies due to the neglect of long-range interactions within single layer fragmentation methods have long been recognized and addressed to some extent via systematically increasing the subsystem size or using multiple levels (layers) of theory.[61−73] Most fragmentation-based methods such as molecular fractionation with conjugate caps (MFCC), the fragment molecular orbital method (FMO), and the generalized many-body expansion (GMBE) routinely use a cutoff distance of ≥3.5 Å or 3−4 body expansion terms for truncation resulting in more than 50 atoms per subsystem to achieve a reasonable accuracy. However, in general, such an increase in the subsystem size (>50 atoms per fragment) significantly increases the associated computational cost. Additionally, many methods still inherit some systematic errors that are expected to increase with the size of the molecular systems; hence, they can be reliably used only for evaluating relative energies or properties. Considering these factors motivates us to investigate three layer models, wherein we keep the high level fragments as small as possible while relegating the larger fragments to the lower, and computationally less expensive levels of theory. This will be demonstrated by a convergence of the three layer results to the two layer results when the high level fragments are left at the same size, but the lowest level of theory is decreased to a semiempirical method.

In the present study, we have used our multilayer Molecules-in-Molecules (MIM) approach on a carefully selected test set of biomolecules specifically chosen to be challenging systems for fragmentation methods. The selected molecules represent the most common peptide secondary structures with many nonlocal interactions that must be accounted for at the high level of theory. The purpose of this study is to establish a standard protocol for the application of the MIM method to the study of various chemical properties of natural-sized biomolecules. A major goal of our study is to develop our protocol without having undue restrictions on the size of the molecules that can be handled. Specifically, we have purposefully kept the size of the region to be calculated with the high level of theory to be sufficiently small for our protocol to be applicable with a range of sophisticated theoretical methods. Additional layers are used with progressively more efficient computational methods applied to increasingly larger regions of the molecule. The use of multiple layers allows more interactions to be evaluated at a reasonable level of theory, contributing to the overall accuracy.

The accuracy of the multilayer MIM approach has been assessed by comparing the calculated MIM total energy to that of a full (unfragmented) calculation at the high level of theory. Our goal is to achieve a target of "chemical accuracy" (1−2 kcal/mol) in the total energy for the molecules in the test set. Common DFT methods such as B3LYP, B3LYP-D3BJ, and ωB97X-D with different basis sets, as well as dispersion-corrected semiempirical methods, are evaluated for their ability to perform as high, medium, or low levels of theory. A careful and systematic assessment of the critical factors and effects influencing the accuracy of the results has been performed to obtain the optimum protocol for obtaining the total energy. Though the current study is restricted to the calculation of the total energy of the system, other molecular properties (e.g., binding energies, NMR, pK_as, etc.) can be investigated using a similar approach. However, for molecules undergoing significant change in geometry (such as in tautomerization reactions or during molecular dynamics), care has to be exercised in the fragmentation schemes to avoid or to minimize the effects of discontinuities in the resulting potential energy surfaces, particularly if the makeup of the subsystems changes significantly. Work in these directions, including the use of appropriate smoothing functions, is currently ongoing in our group.

2. COMPUTATIONAL METHODS

The working principles behind the Molecules-in-Molecules (MIM) fragment-based method have been explained in previous publications by Raghavachari and co-workers.[73,74] The multilayer component of the method shares the same underlying philosophy as the ONIOM methodology developed by Morokuma and co-workers.[83−85] The fundamentals of MIM can be described in four steps: (1) initial fragmentation of the large molecule into nonoverlapping, small fragments, (2) formation of overlapping primary subsystems from the local interactions between fragments, (3) formation of the derivative subsystem to account for the overcounting from the overlapping regions via the inclusion-exclusion principle, and (4) evaluation of the large molecules' energy by summation of the independent energies of the individual subsystems, taking into account the signs for the energies of the derivative subsystems. The initial fragmentation step follows a general method of breaking only single bonds between non-hydrogen atoms. As previous work by Saha and Raghavachari[74] suggested that keeping the peptide C−N linkage together results in a smaller error in the calculated total energy, all peptide linkages are left unbroken and intact in this study.

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

Each primitive, nonoverlapping fragment ("monomer") initiates a primary subsystem. The various fragmentation schemes used in this study are illustrated in Figure 1. The

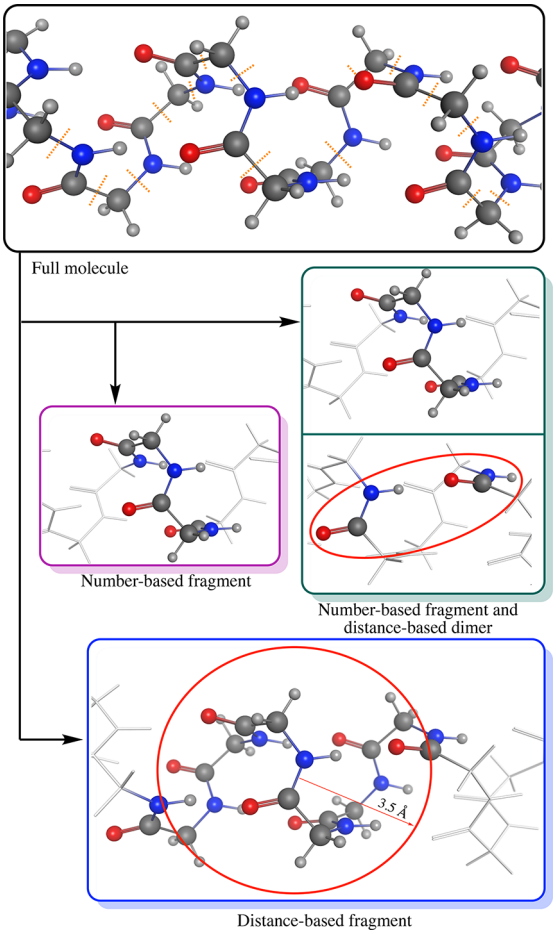


Figure 1. Graphical illustration of the various fragmentation schemes used in this study. Dashed lines across bonds in the full molecule indicate bonds to be broken during the initial fragmentation. See the Computational Methods section of the main text for more details.

formation of the primary subsystems has been carried out in three different ways: (**a**) number-based (η) fragmentation, where η denotes the number of covalently bonded monomers combined to form a primary subsystem, (**b**) distance-based fragmentation with a cutoff distance (r) where all monomers within r Å are collected together, and (**c**) number-based fragmentation augmented with nonbonded dimers based on a cutoff distance. The first two subsystem formation schemes have been explained and used in detail in previous studies from our group.[73−76] As most of the protein molecules contain a large number of nonbonded interactions (attractive and/or repulsive), a purely number-based fragmentation based on the connectivity could have larger errors (*vide infra*). To incorporate some of the important nonbonded interactions, we have introduced the new scheme (**c**). The goal of this third scheme is to supplement the number-based subsystems with key, nonbonded interactions by defining a new cutoff distance, d Å. Whereas the traditional number-based scheme (**a**) only forms subsystems from covalently bonded fragments, scheme (**c**) also allows dimer subsystems to be formed from the nonbonded interactions between fragments within d Å and can be thought of as a hybrid number- and distance-based method.

All truncated bonds in the primary and derivative subsystems are capped with link hydrogen atoms. Once the primary subsystems are formed, derivative subsystems are generated from the overlaps between primary subsystems. Finally, the independently calculated energies are summed according to the inclusion−exclusion principle to obtain the final total energy and other desired molecular properties:

$$|A_1 \cup \ldots \cup A_n| = \sum_i |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n-1} |A_1 \cap \ldots \cap A_n| \quad (1)$$

In this study, we have performed MIM calculations with 1, 2, and 3 layers (notations: MIM1, MIM2, and MIM3, respectively), each layer treated with a different level of theory. In MIM1, the primary subsystems are generated using a single parameter, and calculations are performed with a single level of theory. The energy expression for MIM1 is

$$E^{\mathrm{MIM1}} = E^{r}_{\mathrm{high}} \quad (2)$$

where "r" is an arbitrary parameter defining the size of the subsystem. It can be, for example, the number of monomers combined (in number−based fragmentation) or a distance cutoff (in distance-based fragmentation).

The two-layer MIM calculation (MIM2) uses two fragmentation parameters ($r \ll R$) and two levels of theory. The subsystems formed with the smaller parameter (r) are calculated with both the high and low levels of theory, and the low level is used on the subsystems generated using the larger parameter (R). R can be (and in this study, is always) the full molecule ($R = \infty$). The general MIM2 energy expression can be written in the same manner as the standard ONIOM extrapolation:[83]

$$E^{\mathrm{MIM2}} = E^{r}_{\mathrm{high}} - (E^{r}_{\mathrm{low}} - E^{R}_{\mathrm{low}}) \quad (3)$$

MIM3 involves three fragmentation parameters ($r < r' \ll R$) with three different levels of theory. The subsystems generated with the smallest parameter (r) are calculated with the high and medium levels of theory, subsystems generated with the intermediate parameter (r') are treated with the medium and low levels of theory, and finally, the energy of the fragments with the largest parameter (R) is calculated with the lowest level of theory. The size of subsystems in the middle layer should be larger than those in the high layer to allow some of the missing long-range interactions at the high level to be picked up by the medium level of theory, without necessarily increasing the overall computational cost. Energy calculation of the full molecule (when $R = \infty$) can be performed by using some computationally efficient level of theory such as semiempirical methods (e.g., PM3, PM6, etc.). The general energy expression for MIM3 can be written as

$$E^{\mathrm{MIM3}} = E^{r}_{\mathrm{high}} - (E^{r}_{\mathrm{med}} - E^{r'}_{\mathrm{med}}) - (E^{r'}_{\mathrm{low}} - E^{R}_{\mathrm{low}}) \quad (4)$$

In a similar way, more layers can be added to the MIM calculation (MIM4, MIM5, etc.) to extrapolate the total energy. It is important to note that the use of multiple layers with different levels of theory makes MIM a very efficient extrapolation method compared to other similar methods employing only a single layer of fragmentation, as including a low level calculation allows overall subsystem sizes to be

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

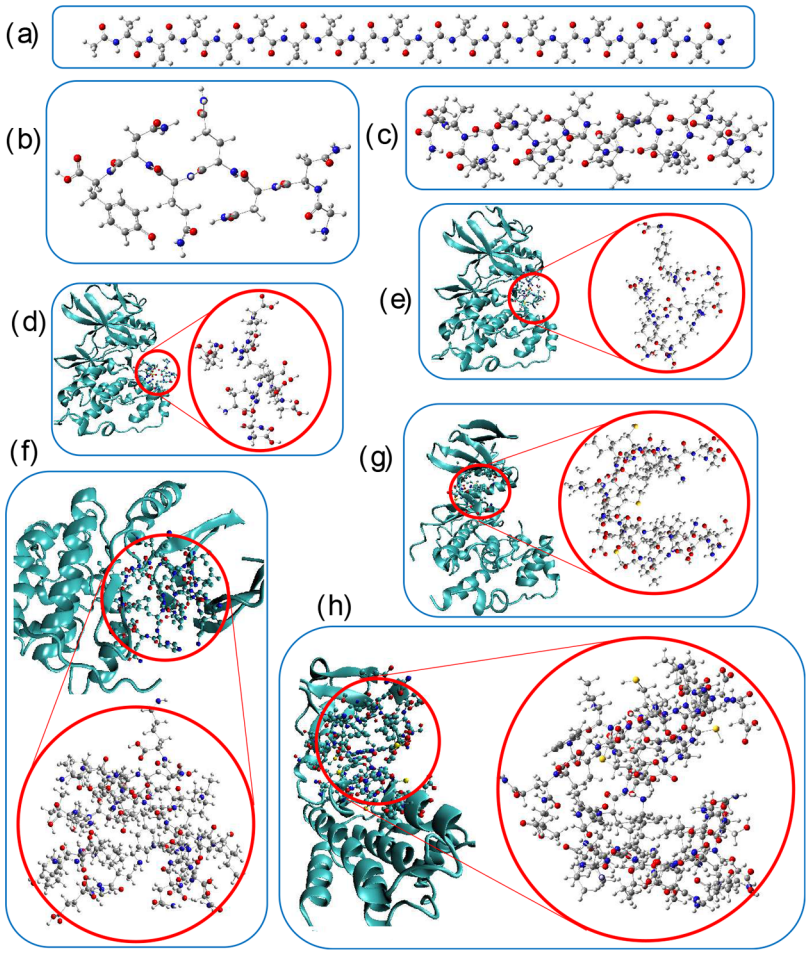


Figure 2. Molecules used in this study. Sections of molecules d−h shown in the red circle were used in this study. The different motifs shown above are (a) β-Acetyl(ala)$_{18}$-NH$_2$ (189 atoms, structure label: **β-strand**), (b) PDB ID: 1YJP (107 atoms, structure label: **1YJP**), (c) α-Acetyl(ala)$_{18}$-NH$_2$ (189 atoms, structure label: **α-helix**), (d) PDB ID: 3O17, structure shown in the red circle includes residues within 5 Å of the ligand SO4-801 (144 atoms, structure label: **3O17-a**), (e) PDB ID: 3O17, structure shown in the red circle includes residues within 5.0 Å of the ligand SO4-802 (190 atoms, structure label: **3O17-b**), (f) PDB ID: 1UKI, structure shown in the red circle includes residues within 7.0 Å of the ligand 537 (384 atoms, structure label: **1UKI**), (g) PDB ID: 2VTA, structure shown in the red circle includes residues within 8 Å of the ligand LZ1 (565 atoms, structure label: **2VTA**), and (g) PDB ID: 4QD6, structure shown in the red circle includes residues within 8.0 A of the ligand 30T (721 atoms, structure label: **4QD6**).

reduced. While the low level calculation on the full molecule is typically the longest single calculation in a multilayer MIM job, going to 3 or 4 levels of theory allows us to choose a low level efficient enough that this calculation remains tractable while avoiding a significant decrease in accuracy.

The single point energies (full and fragmented) of all test set molecules were calculated with a range of density functionals: B3LYP,[86−88] dispersion-corrected B3LYP-D3BJ (using Grimme's D3 dispersion correction[89] and Becke-Johnson damping[90]), B97-D3BJ,[91,92] and ωB97X-D[93] with 6-311++G(d,p), 6-31+G(d), and 6-31+G basis sets.[94−98] Additionally, three semiempirical methods (PM6,[99] PM6-D3, and PM6-D3H4[100]) were investigated to explore their effectiveness in evaluating long-range interactions inexpensively. All calculations were performed in the gas phase. The simplicity of the MIM method allows us to use different computational packages to evaluate the different component energies, as appropriate. Most calculations were performed using the Gaussian (G16) program suite.[101] PM6-D3H4 calculations were performed using the MOPAC semiempirical computational package.[102]

3. RESULTS AND DISCUSSION

Performance Assessment for DFT. The set of 8 biomolecules used in this study is shown in Figure 2. The total number of atoms ranges from 107 to 721 across the test set. The molecules listed in Figure 2 were selected such that they are chemically interesting while still being of a reasonable enough size that calculations could be performed on the unfragmented molecule to obtain reference values. Structures **1a**, **1b**, and **1c** were taken from the paper by Saha and Raghavachari.[74] In particular, **1a** and **1c** correspond to the **β-strand** and **α-helix** conformations of an alanine polypeptide (18-mer) chain and may be illustrative to assess the performance of MIM for describing conformational preferences of importance in processes such as protein folding. The other molecules are protein kinases, taken from crystal structures provided in the Protein Data Bank (PDB) (Figure 2 caption provides additional detail about these structures). VMD visualization software was used to select a region around a chosen ligand (residues within 5−8 Å radius of the ligand) at the active site of the protein molecule. Since the principal focus of this study is to calibrate the ability of MIM to reproduce the

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

Table 1. Deviation in MIM1 Total Energy at DFT/6-311++G(d,p)[a]

structure label	B3LYP				B3LYP-D3BJ				B97-D3BJ				ωB97X-D			
	N4	N5	N4D	N5D	N4	N5	N4D	N5D	N4	N5	N4D	N5D	N4	N5	N4D	N5D
β-strand	4.5	7.3	4.5	7.3	8.2	7.8	8.2	7.8	10.0	8.1	10.0	8.1	6.9	7.7	6.9	7.7
1YJP	119.9	117.0	27.6	30.4	178.5	159.8	42.0	43.0	161.2	141.9	36.2	37.9	181.7	162.6	48.6	48.2
α-helix	11.2	13.9	0.6	3.2	25.7	26.0	6.1	6.4	24.6	24.3	6.3	6.0	25.3	26.0	5.6	6.3
3O17-a	−27.6	−18.7	19.9	15.3	9.2	16.3	31.5	27.6	8.8	16.3	33.0	28.7	10.3	18.0	30.8	27.6
3O17-b	−1.2	−0.8	15.0	9.3	50.9	48.6	32.5	26.8	48.5	46.4	32.8	26.6	53.3	51.6	33.6	28.8
1UKI	−17.1	−21.5	43.3	37.8	109.9	97.0	97.9	86.0	102.9	88.9	98.8	85.8	113.6	102.8	99.7	90.2
2VTA	38.4	36.7	42.7	34.2	249.1	230.7	122.3	108.9	240.1	220.1	119.5	104.5	261.3	244.4	126.8	116.3
4QD6	90.3	80.8	56.6	51.3	396.6	363.3	178.6	163.6	376.1	342.1	173.9	156.7	408.6	377.5	185.8	173.8
mean absolute error	38.8	37.1	26.3	23.6	128.5	118.7	64.9	58.8	121.5	111.0	63.8	56.8	132.6	123.8	67.2	62.4

[a]Errors in kcal/mol compared to full calculation at DFT/6-311++G(d,p). Fragmentation schemes **N4**, **N5**, **N4D**, and **N5D** are as described in the Computational Methods section.

energies of complex protein structures, the ligand molecule itself along with the solvent molecules was removed from the active site, and hydrogen atoms were added to satisfy the valence properties of the atoms. The charged residues: aspartic acid (ASP), glutamic acid (GLU), lysine (LYS), histidine (HIS), and arginine (ARG) were neutralized. While MIM calculations can be performed in solvent with an implicit solvation model or by including explicit solvent molecules, we chose to limit the number of variables in this study by considering only the neutral, gas phase proteins, while still being sufficiently complex with a range of sizes and secondary interactions.[76] Any unphysical contacts among the added hydrogen atoms were removed by optimizing the hydrogen atoms using a semiempirical method, PM6-D3, while keeping the coordinates of heavy atoms fixed.

We have used four combinations of the number-based fragmentation schemes for generating primary subsystems in the high layer:

a) four monomers combined (η = 4) (**N4**, henceforth),

b) five monomers combined (η = 5) (**N5**, henceforth),

c) four monomers combined (η = 4) plus distance-based nonbonded dimers (**N4D**, henceforth),

d) five monomers combined (η = 5) plus distance-based nonbonded dimers (**N5D**, henceforth).

The subsystems generated by combining 4 and 5 monomers (**N4** and **N5**) follow the connectivity along the chain. Distance-based, nonbonded dimers were included by using a cutoff distance (d Å). Test calculations showed that a cutoff distance of 3.0 Å is an ideal value for including some of the important local, nonbonded interactions. In particular, both strong and weak hydrogen bonding interactions will be included using this cutoff distance in the high layer, contributing to the higher overall accuracy of MIM for these systems. Thus, for schemes **N4D** and **N5D**, nonbonded dimers within 3.0 Å were included. For two-layer MIM (MIM2) studies, all the low layer calculations were performed on the full molecule (R = ∞). For three-layer MIM (MIM3) studies, a distance-based scheme (cutoff radius of 2.5, 3.0, and 3.5 Å) was used to obtain the subsystems for the middle layer. Since the size of the subsystems with 3.5 Å distance is already quite large (containing about 40−50 atoms), distance parameters larger than 3.5 Å were not explored. Nevertheless, ring stacking interactions, if present, will be included at this cutoff distance, though they were not prevalent in our test systems. Illustrations of the fragmentation schemes used in this study are shown in Figure 1.

Table 1 presents the errors in the calculated total energies for the individual molecules at the MIM1 level. Four different DFT functionals (B3LYP, B3LYP-D3BJ, B97-D3BJ, and ωB97X-D) in conjunction with the 6-311++G(d,p) basis set were used in the MIM1 calculations. The errors in the calculated total energies are quite significant for the purely number-based fragmentation schemes (**N4** and **N5**). A closer inspection of the energy deviations reveals that the error is smaller for less crowded molecules such as ***β*-strand** and **1YJP** and grows rapidly as the molecules become larger and bulkier, being as large as 408 kcal/mol for **4QD6** with ωB97X-D. When the nonbonded interactions were included via the distance-based dimers (fragmentation schemes **N4D** and **N5D**) with a cutoff distance of 3.0 Å, the error is reduced substantially (by as much as 200 kcal/mol in the case of **4QD6**). Interestingly, the deviation in MIM1 total energy is larger for the functionals including dispersion corrections (B3LYP-D3BJ, B97-D3BJ, and ωB97X-D) compared to the one without (B3LYP). This suggests that part of the large deviation observed in the MIM1 total energy for the bulkier molecules could be due to the failure of MIM1 to completely capture the interfragment, nonbonded interactions. In this context, it is important to remember that the error in MIM1 could be significantly reduced by increasing the subsystem size as has been done by many other fragment-based methods, but that would also increase the computational cost substantially. Instead, as mentioned earlier, we purposefully use a relatively small subsystem size to explore the effects of multilayer calculations, while still keeping the high layer calculations computationally tractable even for correlated levels of theory.

Tables 2 and 3 summarize the deviations in the total energies calculated using two-layer MIM (MIM2) fragmentation. As in MIM1, the four DFT functionals (B3LYP, B3LYP-D3BJ, B97-D3BJ, and ωB97X-D) with the 6-311++G(d,p) basis set were used as the high levels of theory, and the same DFT functionals with 6-31+G and 6-31+G(d) basis sets were employed as the low levels of theory. The low layer for all MIM2 calculations was the full molecule (R = ∞). The benefit of including the long-range interactions missing in the first layer via the second layer can immediately be seen across all methods. The mean absolute error has been reduced by a significant amount (on average by about 60 kcal/mol) compared to MIM1. The error is especially small (less than 3 kcal/mol, close to our target

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

Table 2. Deviation in MIM2 Total Energy at DFT/6-311++G(d,p):DFT/6-31+G[a]

	B3LYP				B3LYP-D3BJ				B97-D3BJ				ωB97X-D			
structure label	N4	N5	N4D	N5D	N4	N5	N4D	N5D	N4	N5	N4D	N5D	N4	N5	N4D	N5D
β-strand	−0.5	−0.7	−0.5	−0.7	−0.5	−0.7	−0.5	−0.7	−0.8	−0.8	−0.8	−0.8	−0.4	−0.9	−0.4	−0.9
1YJP	−22.1	−22.2	−7.9	−7.5	−22.1	−22.2	−7.9	−7.5	−19.8	−19.8	−7.5	−7.1	−19.2	−20.4	−9.6	−8.9
α-helix	−3.1	−2.9	−1.2	−1.0	−3.1	−2.9	−1.2	−1.0	−2.9	−2.6	−1.3	−1.1	−2.6	−2.4	−1.4	−1.2
3O17-a	−7.8	−6.7	−4.1	−3.9	−7.8	−6.7	−4.1	−3.9	−6.7	−5.8	−4.3	−4.1	−6.4	−5.3	−4.2	−4.0
3O17-b	−9.8	−9.5	−3.9	−4.0	−9.8	−9.5	−3.9	−4.0	−8.1	−7.9	−3.9	−3.9	−7.2	−7.0	−4.1	−4.2
1UKI	−21.0	−19.9	−13.7	−12.7	−21.0	−19.9	−13.7	−12.7	−17.1	−16.1	−13.8	−12.8	−15.1	−14.2	−14.3	−13.4
2VTA	−36.3	−35.0	−18.7	−18.2	−36.3	−35.0	−18.7	−18.2	−30.5	−29.4	−19.2	−18.5	−26.8	−26.0	−19.8	−19.4
4QD6	−67.9	−64.2	−33.3	−30.8	−67.9	−64.2	−33.3	−30.8	−59.0	−55.7	−33.6	−31.2	−55.9	−52.8	−36.0	−33.4
mean absolute error	21.0	20.2	10.4	9.9	21.0	20.2	10.4	9.9	18.1	17.3	10.5	9.9	16.7	16.1	11.2	10.7

[a]Errors in kcal/mol compared to full calculation at DFT/6-311++G(d,p). Fragmentation schemes **N4**, **N5**, **N4D**, and **N5D** are as described in the Computational Methods section.

accuracy of 2 kcal/mol) for the polypeptides with less compact structures such as ***β*-strand** and ***α*-helix**. However, the error is quite large (more than 30 kcal/mol) for other dense molecules when 6-31+G is the lower level basis set (Table 2). This implies the fragment−fragment long-range interactions are not fully recovered even when the 6-31+G basis set is used on the entire molecule. When the slightly larger basis set 6-31+G(d) was used as the lower level theory (Table 3), the mean absolute error in MIM2 energy is dramatically reduced to less than 3 kcal/mol and is consistent for all of the DFT methods used. This further corroborates that the 6-31+G basis set is not sufficient enough to capture all of the long-range interactions that are missing in the high layer. When we inspected the deviation in the MIM2 total energy with the 6-31+G(d) basis set, we noticed an interesting pattern. The deviations in MIM2 energy for **N4D** and **N5D** are higher compared to the purely number-based fragmentations schemes **N4** and **N5** by up to 4 kcal/mol. Surprisingly, it is found to be true for almost all of the methods used. To understand this seemingly inconsistent behavior, we performed a few test calculations using a larger basis set, 6-31+G(d,p), as the low level of theory, adding polarization functions on hydrogen atoms. With the 6-31+G(d,p) basis set, the order of performance of the fragmentation schemes **N4**, **N5**, **N4D**, and **N5D** is again consistent as in the case of MIM1 (errors: **N4** > **N5** > **N4D** > **N5D**). This strongly suggests the unsystematic trend seen in the deviation of the MIM2 energy is due to the slight imbalance in the treatment of nonbonded interactions, specifically hydrogen bonds, by the two basis sets 6-311++G(d,p) and 6-31+G(d) used. Nevertheless, it can be concluded that the use of the second layer is very important to include the interactions missing in the first layer, in agreement with the previous study by Saha and Raghavachari.[74]

Though calculations on the unfragmented molecule at a lower level of theory in the second layer of MIM2 significantly increase the accuracy, such full calculations for molecules with more than 1000 atoms can be prohibitively expensive, even for DFT with modest basis sets. Most chemically interesting biomolecules contain several hundred to several thousand atoms. To make quantum chemical calculations practical for studying such large biomolecules, we have performed a careful benchmark study of the performance of the three-layer MIM3 model for our test systems. In MIM3, we maintained the levels of theory and the subsystem generation schemes in the high layer to be the same as in MIM1 and MIM2 (i.e., DFT with 6-311++G(d,p) and four different types of subsystems). As mentioned earlier, a distance-based scheme (cutoff radius of 2.5, 3.0, and 3.5 A) was used to obtain the subsystems for the middle layer. The middle layer calculations were performed at the DFT level with two basis sets (6-31+G and 6-31+G(d)), same as the ones used in the lower layer of MIM2. The same DFT functional was used in high and middle layer calculations to ensure the calculated error is not due to the difference in density functionals. In the low layer, full calculations were performed using three semiempirical methods (PM6, PM6-D3, and PM6-D3H4). Since there are 4 variations (4 different subsystems) in the high layer, 6 variations (3 different subsystems with 2 different basis sets) in the middle layer, and 3 variations (3 different semiempirical methods) in the low layer, in total, 72 combinations are possible for each of the four DFT methods.

A summary of the mean absolute error (MAE) and maximum error calculated for various methods with MIM3 is

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

Table 3. Deviation in MIM2 Total Energy at DFT/6-311++G(d,p):DFT/6-31+G(d)[a]

	B3LYP				B3LYP-D3BJ				B97-D3BJ				ωB97X-D			
structure label	N4	N5	N4D	N5D	N4	N5	N4D	N5D	N4	N5	N4D	N5D	N4	N5	N4D	N5D
β-strand	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.6	0.1	0.6	0.1
1YJP	−1.3	−0.8	−1.4	−0.8	−1.3	−0.8	−1.4	−0.8	−0.3	−0.1	−1.5	−1.0	0.5	0.5	−3.5	−2.6
α-helix	−0.2	−0.1	−0.3	−0.2	−0.2	−0.1	−0.3	−0.2	0.0	0.0	−0.3	−0.2	0.4	0.5	−0.5	−0.3
3O17-a	−0.8	−0.5	−1.3	−1.0	−0.8	−0.5	−1.3	−1.0	−0.2	0.1	−1.3	−1.0	0.7	0.9	−1.1	−0.8
3O17-b	−1.1	−1.3	−0.8	−0.8	−1.1	−1.3	−0.8	−0.8	−0.1	−0.3	−0.7	−0.8	1.3	1.1	−0.8	−0.9
1UKI	1.7	1.5	−1.7	−1.9	1.7	1.5	−1.7	−1.9	4.0	3.5	−1.8	−2.2	6.5	6.0	−1.9	−2.3
2VTA	−0.4	−0.2	−3.4	−3.3	−0.3	−0.1	−3.4	−3.3	3.1	3.0	−3.6	−3.7	7.3	7.1	−4.8	−4.8
4QD6	−3.0	−3.0	−7.4	−7.0	−3.0	−3.1	−4.6	−4.5	1.8	1.2	−5.3	−4.4	7.1	6.5	−8.1	−7.5
mean absolute error	1.2	0.9	2.1	1.9	1.2	0.9	1.8	1.6	1.3	1.1	1.9	1.7	3.0	2.8	2.7	2.4

[a]Errors in kcal/mol compared to full calculation at DFT/6-311++G(d,p). Fragmentation schemes **N4**, **N5**, **N4D**, and **N5D** are as described in the Computational Methods section.

Table 4. Mean Absolute Error and Maximum Deviations in MIM3 Total Energy at a DFT/6-311++G(d,p):DFT/6-31+G(d):Semiempirical Level of Theory[a]

High[a] (Theory) →			B3LYP/6-311++G(d,p)			B3LYP-D3BJ/6-311++G(d,p)			B97-D3BJ/6-311++G(d,p)			ωB97X-D/6-311++G(d,p)		
(Frag) ↓	*Middle*[a] →		*B3LYP/6-31+G(d)*			*B3LYP-D3BJ/6-31+G(d)*			*B97-D3BJ/6-31+G(d)*			*ωB97X-D/6-31+G(d)*		
	↓	Low[a]	PM6	PM6-D3	PM6-D3H4	PM6	PM6-D3	PM6-D3H4	PM6	PM6-D3	PM6-D3H4	PM6	PM6-D3	PM6-D3H4
N4	*3.0 Å*	Full (∞)	2.1 (7.9)	8.1 (27.5)	8.2 (27.4)	8.5 (19.5)	**1.9 (5.7)**	**1.8 (5.2)**	11.1 (26.0)	4.4 (9.2)	4.3 (8.7)	12.5 (30.5)	5.8 (11.2)	5.7 (11.1)
	3.5 Å	Full (∞)	**1.8 (5.8)**	4.3 (16.7)	4.2 (16.3)	4.7 (10.0)	**1.7 (4.7)**	**1.3 (4.8)**	6.7 (14.9)	3.2 (7.2)	2.9 (7.3)	8.5 (19.7)	5.0 (9.7)	4.8 (9.8)
N5	*3.0 Å*	Full (∞)	2.1 (7.9)	8.1 (27.5)	8.2 (27.4)	8.6 (19.5)	**1.9 (5.5)**	**1.8 (5.2)**	10.9 (25.4)	4.2 (8.8)	4.2 (8.2)	12.3 (29.9)	5.6 (11.0)	5.5 (10.9)
	3.5 Å	Ful (∞)l	**1.7 (5.8)**	4.2 (16.7)	4.3 (16.3)	4.7 (9.7)	**1.7 (4.5)**	**1.4 (4.5)**	6.5 (14.3)	3.0 (6.8)	2.8 (6.8)	8.3 (19.1)	4.8 (9.4)	4.6 (9.2)
N4D	*3.0 Å*	Full (∞)	3.1 (13.3)	9.5 (31.5)	9.6 (31.8)	7.5 (17.9)	**1.4 (2.3)**	**1.3 (2.0)**	8.2 (18.8)	**1.7 (3.5)**	**1.6 (2.9)**	6.9 (15.4)	**1.5 (4.2)**	**1.4 (4.1)**
	3.5 Å	Full (∞)	2.5 (10.2)	5.6 (21.1)	5.6 (20.7)	3.6 (7.2)	**1.5 (3.7)**	**1.4 (3.3)**	3.8 (7.7)	**1.4 (3.2)**	**1.3 (2.8)**	3.0 (6.5)	**1.7 (6.3)**	**1.6 (5.9)**
N5D	*3.0 Å*	Full (∞)	3.0 (11.8)	9.4 (31.4)	9.5 (31.3)	7.5 (18.1)	**1.4 (2.2)**	**1.3 (2.3)**	8.3 (19.8)	**1.6 (3.1)**	**1.6 (2.6)**	7.0 (16.0)	**1.5 (3.6)**	**1.4 (3.5)**
	3.5 Å	Full (∞)	2.4 (9.7)	5.6 (20.6)	5.6 (20.2)	3.6 (7.4)	**1.5 (3.5)**	**1.4 (3.1)**	3.9 (8.7)	**1.3 (2.2)**	**1.2 (2.1)**	3.1 (6.1)	**1.6 (5.7)**	**1.6 (5.3)**

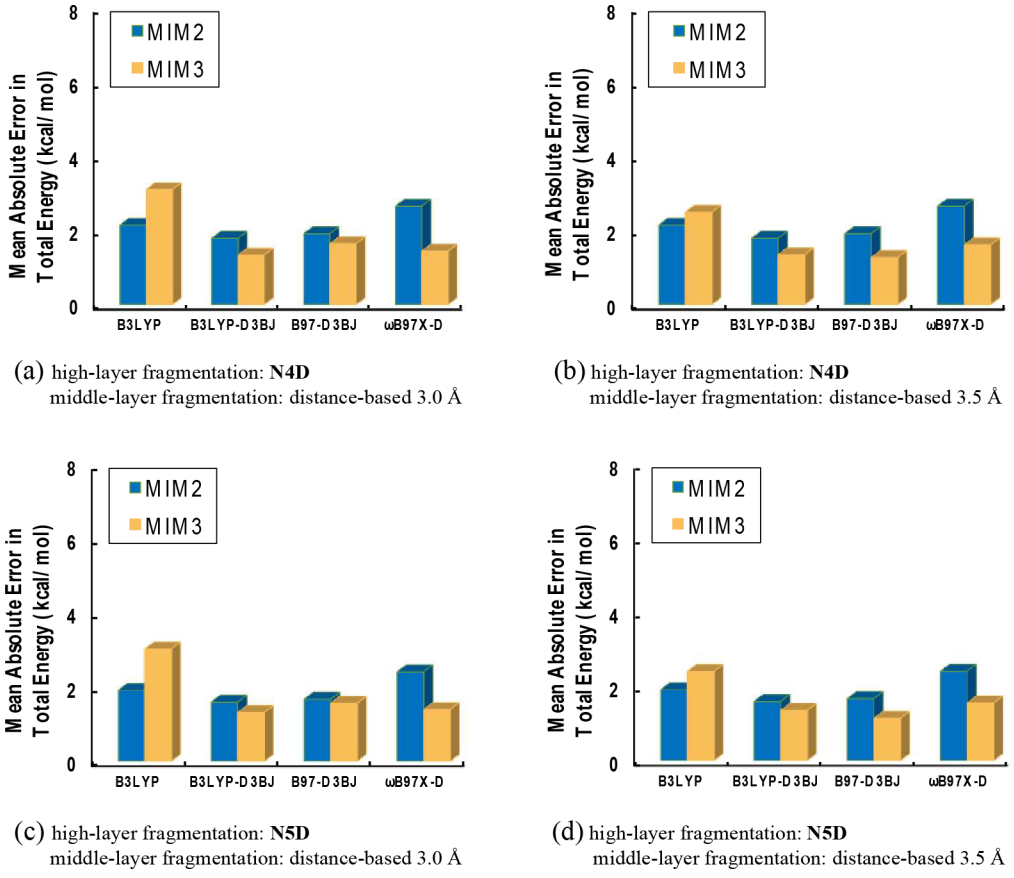
[a]The horizontal axis represents the level of theory used, and the vertical axis represents the fragmentation scheme used in that particular layer. Maximum errors are shown in parentheses; errors smaller than 2 kcal/mol are shown in blue and bolded. Errors in kcal/mol compared to full calculation at DFT/6-311++G(d,p). Fragmentation schemes **N4**, **N5**, **N4D**, and **N5D** are as described in the Computational Methods section.

given in Table 4, with more information available in the Supporting Information in Tables S1 and S2. As in the case of MIM2, the mean absolute errors for MIM3 with DFT/6-31+G as a middle layer are quite significant, ranging from 3 to 70 kcal/mol (SI Table S1). Results are better for **N4D** and **N5D** fragmentation schemes compared to **N4** and **N5**. However, no systematic improvement in the calculated total energy was observed despite several combinations of the fragmentation parameters and semiempirical methods used. Overall, the 6-31+G basis set is found to be insufficient to achieve a target accuracy of better than 2 kcal/mol when coupled with a high level basis set containing polarization functions as 6-311++G(d,p) does.

When the 6-31+G(d) basis set is used as the intermediate level of theory (Tables 4 and S2), some interesting observations can be made. As in MIM1 and MIM2, **N4** and **N5** fragmentation schemes in the high layer of MIM3 resulted in larger errors compared to **N4D** and **N5D**. This suggests that some of the local, nonbonded interactions need to be treated at the high level of theory. As expected, increasing the size of the middle layer (via increasing cutoff distance) increases the accuracy of the MIM3 energy. The MAE is reduced by more than 10 kcal/mol when the cutoff distance is increased from 2.5 to 3.0 Å in the middle layer. When the cutoff distance is increased from 3.0 to 3.5 Å, there is only a modest change in the MIM3 error, implying that we are close to convergence with respect to the size of the middle layer. B3LYP without dispersion corrections performs better with PM6 compared to the other semiempirical methods with dispersion corrections (PM6-D3 or PM6-D3H4). This is consistent with our expectation that the basic components of each level should be similar in multilayer calculations to yield the best-performing combinations. For B3LYP, the error in the calculated total energy is found to be smaller for less bulky molecules. The dispersion-corrected DFT functionals, B3LYP-D3BJ, B97-D3BJ and ωB97X-D, perform very well (MAE < 2 kcal/mol) with the dispersion-corrected PM6-D3 and PM6-D3H4 semiempirical methods. The best performance was obtained with B3LYP-D3BJ and B97-D3BJ functionals. Using **N4D** or **N5D** fragmentation schemes in the high layer, coupled with a distance-based fragmentation (3.0 or 3.5 Å distance cutoff) in the middle layer, the calculated mean absolute error is less than 1.5 kcal/mol.

Overall, several important conclusions can be drawn from our investigations. In the MIM1 calculations, smaller errors are found with the number-based fragments augmented with distance-based dimers, i.e., **N4D** and **N5D**. However, the errors are too large for MIM1 to be a useful method for calculating the energies of large molecules. For MIM2 calculations, DFT/6-31+G(d) as the low level of theory is a

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394



(a) high-layer fragmentation: **N4D**
middle-layer fragmentation: distance-based 3.0 Å

(b) high-layer fragmentation: **N4D**
middle-layer fragmentation: distance-based 3.5 Å

(c) high-layer fragmentation: **N5D**
middle-layer fragmentation: distance-based 3.0 Å

(d) high-layer fragmentation: **N5D**
middle-layer fragmentation: distance-based 3.5 Å

Figure 3. Graphical representation of mean absolute error (MAE) in total energy over the full test set, calculated using MIM2 (blue) and MIM3 (orange) fragment methods compared to the full calculation at the DFT/6-311++G(d,p) basis set. The high level basis set is 6-311++G(d,p). DFT/6-31+G(d) is used as the low level of theory in MIM2 and the intermediate level of theory in MIM3. In MIM3, PM6 is used as the low level of theory for B3LYP and PM6-D3H4 is used for the dispersion-corrected functionals.

straightforward choice, since the high level is also likely to contain multiple sets of diffuse and polarization functions, an important prescription for nonbonded interactions. For MIM3 calculations, the sizes of the middle layer fragments and basis sets are the two most important parameters. It is obvious that one would prefer to have a fragmentation scheme involving subsystems that are not large enough to become the computational bottleneck. Additionally, it is also important not to have subsystems so small that the important long-range interactions are not included at the high level. From the results presented above, subsystems generated with a distance-based parameter with a cutoff radius of 3.0 Å or higher appear to be ideal for the middle layer. Since the number of atoms per subsystem with 3.0 Å distance cutoff is in the range of 30−40, DFT methods with slightly larger basis sets (double-ζ or triple-ζ with polarization functions on both the hydrogen and heavy atoms, for example) can still be affordable without placing unreasonable demands on the computational cost. Furthermore, since the dispersion-corrected DFT methods resulted in the smallest MAE with PM6-D3 and PM6-D3H4, the latter methods can reliably be used to perform the full calculation in the third layer. Figure 3 summarizes the improvement in the total energy for some of the best-performing combinations of the DFT functionals and fragmentation schemes.

The computational rate limiting steps in the different MIM models result from the energy evaluations involving the largest components, viz. primary subsystems. In general, the number of primary subsystems grows linearly with the size of the molecule, while the size of the average subsystem in all the models described in this paper is independent of the size of the parent molecule. Thus, MIM1 scales asymptotically linearly with the size of the large molecule. In MIM2, the high level calculations scale as in MIM1, while the low level calculation on the unfragmented molecule will become rate limiting for the larger molecules. However, in MIM3, *both high level and medium level calculations are independent of the size of the parent molecule*. In addition, since we are using very inexpensive methods such as PM6-D3 as the low level calculation on the whole molecule (to evaluate the long-range effects), the overall scaling in MIM3 is expected to be close to linear for the case of large molecules. In addition, we have shown that MIM3 reaches nearly the same accuracy attained by MIM2, making MIM3 the method of choice for performing accurate calculations on very large molecules.

4. CONCLUSIONS

In this study, we have explored the various fragmentation schemes within our multilayer Molecules-in-Molecules (MIM) fragmentation method with the goal of finding the best fragmentation schemes to accurately predict the total energy and other properties of large biomolecules containing several hundreds of atoms. We have assessed the performance of the fragmentation method using four DFT functionals with the 6-311++G(d,p) basis set. As discussed in the previous publications from our group, two-layer MIM (MIM2) performs substantially better than MIM1 to predict the total energy of

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

the unfragmented molecule with high accuracy. An accuracy of less than 2 kcal/mol in the MIM2 total energy was routinely achieved for the calibration systems using DFT/6-31+G(d) as the low layer, independent of the fragmentation scheme used in the high layer. Since the MIM2 calculations could still be quite expensive for biomolecules with several hundreds of atoms, we have, for the first time, employed three-layer MIM calculations (MIM3) where the full calculation on the unfragmented molecule can be carried out at the semiempirical level. Rigorous calibration of the various fragmentation schemes, different combinations of basis sets, and semiempirical methods allowed us to find the combinations of best performing fragmentation schemes and computational methods. For compact, 3-dimensional molecules with many nonbonded interactions, the 6-31+G basis set (as the middle layer for MIM3 or the low layer for MIM2) performed poorly. For MIM3, a cutoff distance of 2.5 Å used for the middle layer fragments gave larger errors suggesting that some of the strong, nonbonded interactions beyond the 2.5 Å radius are not properly accounted for by the semiempirical methods. The fragments generated with the cutoff distance of 3.0 and 3.5 Å in the middle layer of MIM3 resulted in the smallest errors, with only a small difference in performance between the two distances when the levels of theory are properly paired. The mean absolute errors for B3LYP with any of the combinations of fragmentations schemes and QM methods do not achieve the target accuracy of <2 kcal/mol. Dispersion-corrected DFT (B3LYP-D3BJ, B97-D3BJ, and ωB97X-D) methods pair nicely with semiempirical methods, PM6-D3 and PM6-D3H4. For B3LYP-D3BJ and B97-D3BJ with MIM3, the best results (MAE < 2 kcal/mol) are obtained with number-based fragments augmented with distance-based dimers (i.e., **N4D** and **N5D**) in the high layer, distance-based subsystems with a cutoff radius of 3.0 and 3.5 Å with 6-31+G(d) basis set in the middle layer, and semiempirical methods (PM6-D3 and PM6-D3H4) as a low layer. These results suggest that through increasing the number of layers in a given calculation and paying close attention to the compatibility between different levels of theory and fragmentation schemes, the MIM fragmentation method can be used to study biomolecules with several hundred (even several thousand) atoms.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b01198.

Mean absolute error (MAE) and maximum error calculated for various methods with MIM3 and DFT total energies of unfragmented molecules (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kraghava@indiana.edu.

ORCID

Bishnu Thapa: 0000-0003-3521-1062

Krishnan Raghavachari: 0000-0003-3275-1426

Funding

This work was supported by Eli Lilly and Company through the Lilly Research Award Program at Indiana University. The methods used in this project were developed based on the support from the National Science Foundation, grant CHE-1665427 at Indiana University.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Jon Erickson for stimulating discussions.

■ REFERENCES

(1) Brorsen, K.; Fedorov, D. G. Fully analytic energy gradient in the fragment molecular orbital method. *J. Chem. Phys.* **2011**, *134* (12), 124115.

(2) Car, R.; Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **1985**, *55* (22), 2471–2474.

(3) Gao, J. Toward a molecular orbital derived empirical potential for liquid simulations. *J. Phys. Chem. B* **1997**, *101* (4), 657–663.

(4) Xie, W.; Gao, J. Design of a next generation force field: The X-POL potential. *J. Chem. Theory Comput.* **2007**, *3* (6), 1890–1900.

(5) Xie, W.; Orozco, M.; Truhlar, D. G.; Gao, J. X-POL potential: An electronic structure-based force field for molecular dynamics simulation of a solvated protein in water. *J. Chem. Theory Comput.* **2009**, *5* (3), 459–467.

(6) Peters, M.; Raha, K.; Merz, K., Jr. Quantum mechanics in structure-based drug design. *Curr. Opin. Drug. Discovery Devel.* **2006**, *9* (3), 370–379.

(7) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M. The role of quantum mechanics in structure-based drug design. *Drug Discovery Today* **2007**, *12* (17), 725–731.

(8) Heifetz, A.; Chudyk, E. I.; Gleave, L.; Aldeghi, M.; Cherezov, V.; Fedorov, D. G.; Biggin, P. C.; Bodkin, M. J. The fragment molecular orbital method reveals new insight into the chemical nature of gpcr–ligand interactions. *J. Chem. Inf. Model.* **2016**, *56* (1), 159–172.

(9) Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010**, *53* (14), 5061–5084.

(10) Tong, Y.; Mei, Y.; Li, Y. L.; Ji, C. G.; Zhang, J. Z. Electrostatic polarization makes a substantial contribution to the free energy of avidin–biotin binding. *J. Am. Chem. Soc.* **2010**, *132* (14), 5137–5142.

(11) Beratan, D. N.; Liu, C.; Migliore, A.; Polizzi, N. F.; Skourtis, S. S.; Zhang, P.; Zhang, Y. Charge transfer in dynamical biosystems, or the treachery of (static) images. *Acc. Chem. Res.* **2015**, *48* (2), 474–481.

(12) Ozawa, T.; Okazaki, K.; Kitaura, K. CH/π hydrogen bonds play a role in ligand recognition and equilibrium between active and inactive states of the β2 adrenergic receptor: An ab initio fragment molecular orbital (FMO) study. *Bioorg. Med. Chem.* **2011**, *19* (17), 5231–5237.

(13) Fedorov, D. G.; Nagata, T.; Kitaura, K. Exploring chemistry with the fragment molecular orbital method. *Phys. Chem. Chem. Phys.* **2012**, *14* (21), 7562–7577.

(14) Gallivan, J. P.; Dougherty, D. A. Cation-π interactions in structural biology. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (17), 9459–9464.

(15) Heifetz, A.; Trani, G.; Aldeghi, M.; MacKinnon, C. H.; McEwan, P. A.; Brookfield, F. A.; Chudyk, E. I.; Bodkin, M.; Pei, Z.; Burch, J. D.; Ortwine, D. F. Fragment molecular orbital method applied to lead optimization of novel interleukin-2 inducible T-cell kinase (ITK) inhibitors. *J. Med. Chem.* **2016**, *59* (9), 4352–4363.

(16) Johnston, R. C.; Cheong, P. H.-Y. C–H···O non-classical hydrogen bonding in the stereomechanics of organic transformations: Theory and recognition. *Org. Biomol. Chem.* **2013**, *11* (31), 5057–5064.

(17) Lu, Y. X.; Zou, J. W.; Wang, Y. H.; Yu, Q. S. Substituent effects on noncovalent halogen/π interactions: Theoretical study. *Int. J. Quantum Chem.* **2007**, *107* (6), 1479–1486.

(18) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry: Introduction to advanced electronic structure theory*; Courier Corporation: 2012.

(19) Phipps, M. J.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Energy decomposition analysis approaches and their evaluation on prototypical protein−drug interaction patterns. *Chem. Soc. Rev.* **2015**, *44* (10), 3177−3211.
(20) Collins, M. A.; Bettens, R. P. A. Energy-based molecular fragmentation methods. *Chem. Rev.* **2015**, *115* (12), 5607−5642.
(21) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.* **2012**, *112* (1), 632−672.
(22) Raghavachari, K.; Saha, A. Accurate composite and fragment-based quantum chemical models for large molecules. *Chem. Rev.* **2015**, *115* (12), 5643−5677.
(23) Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137* (6), 064113.
(24) Saha, A.; Raghavachari, K. Dimers of dimers (DOD): A new fragment-based method applied to large water clusters. *J. Chem. Theory Comput.* **2014**, *10* (1), 58−67.
(25) Nanda, K. D.; Beran, G. J. Prediction of organic molecular crystal geometries from MP2-level fragment quantum mechanical/molecular mechanical calculations. *J. Chem. Phys.* **2012**, *137* (17), 174106.
(26) Wen, S.; Nanda, K.; Huang, Y.; Beran, G. J. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. *Phys. Chem. Chem. Phys.* **2012**, *14* (21), 7578−7590.
(27) Isegawa, M.; Truhlar, D. G. Valence excitation energies of alkenes, carbonyl compounds, and azabenzenes by time-dependent density functional theory: Linear response of the ground state compared to collinear and noncollinear spin-flip tddft with the tamm-dancoff approximation. *J. Chem. Phys.* **2013**, *138* (13), 134111.
(28) Leverentz, H. R.; Truhlar, D. G. Electrostatically embedded many-body approximation for systems of water, ammonia, and sulfuric acid and the dependence of its performance on embedding charges. *J. Chem. Theory Comput.* **2009**, *5* (6), 1573−1584.
(29) Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. Evaluation of the electrostatically embedded many-body expansion and the electrostatically embedded many-body expansion of the correlation energy by application to low-lying water hexamers. *J. Chem. Theory Comput.* **2008**, *4* (1), 33−41.
(30) Dahlke, E. E.; Truhlar, D. G. Electrostatically embedded many-body expansion for large systems, with applications to water clusters. *J. Chem. Theory Comput.* **2007**, *3* (1), 46−53.
(31) Huang, L.; Massa, L.; Karle, J. The kernel energy method: Application to a tRNA. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (5), 1233−1237.
(32) Huang, L.; Massa, L.; Karle, J. Kernel energy method illustrated with peptides. *Int. J. Quantum Chem.* **2005**, *103* (6), 808−817.
(33) Tan, H.-J.; Bettens, R. P. Ab initio NMR chemical-shift calculations based on the combined fragmentation method. *Phys. Chem. Chem. Phys.* **2013**, *15* (20), 7541−7547.
(34) Frankcombe, T. J.; Collins, M. A. Potential energy surfaces for gas-surface reactions. *Phys. Chem. Chem. Phys.* **2011**, *13* (18), 8379−8391.
(35) Le, H.-A.; Tan, H.-J.; Ouyang, J. F.; Bettens, R. P. Combined fragmentation method: A simple method for fragmentation of large molecules. *J. Chem. Theory Comput.* **2012**, *8* (2), 469−478.
(36) Collins, M. A. Systematic fragmentation of large molecules by annihilation. *Phys. Chem. Chem. Phys.* **2012**, *14* (21), 7744−7751.
(37) He, X.; Zhang, J. Z. H. A new method for direct calculation of total energy of protein. *J. Chem. Phys.* **2005**, *122* (3), 031103.
(38) Yeole, S. D.; Gadre, S. R. Molecular cluster building algorithm: Electrostatic guidelines and molecular tailoring approach. *J. Chem. Phys.* **2011**, *134* (8), 084111.
(39) Collins, M. A. Molecular potential energy surfaces constructed from interpolation of systematic fragment surfaces. *J. Chem. Phys.* **2007**, *127* (2), 024104.
(40) Collins, M. A.; Deev, V. A. Accuracy and efficiency of electronic energies from systematic molecular fragmentation. *J. Chem. Phys.* **2006**, *125* (10), 104104.
(41) Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. The combined fragmentation and systematic molecular fragmentation methods. *Acc. Chem. Res.* **2014**, *47* (9), 2776−2785.
(42) Chen, X.; Zhang, D.; Zhang, J. Fractionation of peptide with disulfide bond for quantum mechanical calculation of interaction energy with molecules. *J. Chem. Phys.* **2004**, *120* (2), 839−844.
(43) Gao, A. M.; Zhang, D. W.; Zhang, J. Z.; Zhang, Y. An efficient linear scaling method for ab initio calculation of electron density of proteins. *Chem. Phys. Lett.* **2004**, *394* (4), 293−297.
(44) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. Fragment quantum mechanical calculation of proteins and its applications. *Acc. Chem. Res.* **2014**, *47* (9), 2748−2757.
(45) Zhang, D. W.; Zhang, J. Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein−molecule interaction energy. *J. Chem. Phys.* **2003**, *119* (7), 3599−3605.
(46) Hua, S.; Hua, W.; Li, S. An efficient implementation of the generalized energy-based fragmentation approach for general large molecules. *J. Phys. Chem. A* **2010**, *114* (31), 8126−8134.
(47) Hua, S.; Li, W.; Li, S. The generalized energy-based fragmentation approach with an improved fragmentation scheme: Benchmark results and illustrative applications. *ChemPhysChem* **2013**, *14* (1), 108−115.
(48) Wang, K.; Li, W.; Li, S. Generalized energy-based fragmentation CCSD(T)-F12a method and application to the relative energies of water clusters $(H_2O)_{20}$. *J. Chem. Theory Comput.* **2014**, *10* (4), 1546−1553.
(49) Li, S.; Li, W.; Ma, J. Generalized energy-based fragmentation approach and its applications to macromolecules and molecular aggregates. *Acc. Chem. Res.* **2014**, *47* (9), 2712−2720.
(50) Li, W.; Li, S.; Jiang, Y. Generalized energy-based fragmentation approach for computing the ground-state energies and properties of large molecules. *J. Phys. Chem. A* **2007**, *111* (11), 2193−2199.
(51) Hua, W.; Fang, T.; Li, W.; Yu, J.-G.; Li, S. Geometry optimizations and vibrational spectra of large molecules from a generalized energy-based fragmentation approach. *J. Phys. Chem. A* **2008**, *112* (43), 10864−10872.
(52) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular tailoring approach for geometry optimization of large molecules: Energy evaluation and parallelization strategies. *J. Chem. Phys.* **2006**, *125* (10), 104109.
(53) Rahalkar, A. P.; Katouda, M.; Gadre, S. R.; Nagase, S. Molecular tailoring approach in conjunction with MP2 and RI-MP2 codes: A comparison with fragment molecular orbital method. *J. Comput. Chem.* **2010**, *31* (13), 2405−2418.
(54) Sahu, N.; Gadre, S. R. Molecular tailoring approach: A route for ab initio treatment of large clusters. *Acc. Chem. Res.* **2014**, *47* (9), 2739−2747.
(55) Rahalkar, A. P.; Ganesh, V.; Gadre, S. R. Enabling ab initio Hessian and frequency calculations of large molecules. *J. Chem. Phys.* **2008**, *129* (23), 234101.
(56) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. Molecular tailoring approach for simulation of electrostatic properties. *J. Phys. Chem.* **1994**, *98* (37), 9165−9169.
(57) Ishida, T.; Fedorov, D. G.; Kitaura, K. All electron quantum chemical calculation of the entire enzyme system confirms a collective catalytic device in the chorismate mutase reaction. *J. Phys. Chem. B* **2006**, *110* (3), 1457−1463.
(58) Fedorov, D. G.; Kitaura, K. Use of an auxiliary basis set to describe the polarization in the fragment molecular orbital method. *Chem. Phys. Lett.* **2014**, *597*, 99−105.
(59) Pruitt, S. R.; Bertoni, C.; Brorsen, K. R.; Gordon, M. S. Efficient and accurate fragmentation methods. *Acc. Chem. Res.* **2014**, *47* (9), 2786−2794.
(60) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. Fragment molecular orbital method: An approximate computational method for large molecules. *Chem. Phys. Lett.* **1999**, *313* (3), 701−706.
(61) Hopkins, B. W.; Tschumper, G. S. A multicentered approach to integrated QM/QM calculations. Applications to multiply hydrogen bonded systems. *J. Comput. Chem.* **2003**, *24* (13), 1563−1568.

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

(62) Hopkins, B. W.; Tschumper, G. S. Multicentred QM/QM methods for overlapping model systems. *Mol. Phys.* **2005**, *103* (2−3), 309−315.
(63) Hopkins, B. W.; Tschumper, G. S. Integrated quantum mechanical approaches for extended π systems: Multicentered QM/QM studies of the cyanogen and diacetylene trimers. *Chem. Phys. Lett.* **2005**, *407* (4), 362−367.
(64) Řezáč, J.; Salahub, D. R. Multilevel fragment-based approach (MFBA): A novel hybrid computational method for the study of large molecules. *J. Chem. Theory Comput.* **2010**, *6* (1), 91−99.
(65) Dahlke, E. E.; Truhlar, D. G. Electrostatically embedded many-body correlation energy, with applications to the calculation of accurate second-order Møller−Plesset perturbation theory energies for large water clusters. *J. Chem. Theory Comput.* **2007**, *3* (4), 1342−1348.
(66) Beran, G. J. Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. *J. Chem. Phys.* **2009**, *130* (16), 164115.
(67) Elsohly, A. M.; Shaw, C. L.; Guice, M. E.; Smith, B. D.; Tschumper, G. S. Analytic gradients for the multicentred integrated QM:QM method for weakly bound clusters: Efficient and accurate 2-body:many-body geometry optimizations. *Mol. Phys.* **2007**, *105* (19−22), 2777−2782.
(68) Nakano, T.; Mochizuki, Y.; Yamashita, K.; Watanabe, C.; Fukuzawa, K.; Segawa, K.; Okiyama, Y.; Tsukamoto, T.; Tanaka, S. Development of the four-body corrected fragment molecular orbital (FMO4) method. *Chem. Phys. Lett.* **2012**, *523*, 128−133.
(69) Mayhall, N. J.; Raghavachari, K. Many-overlapping-body (MOB) expansion: A generalized many body expansion for non-disjoint monomers in molecular fragmentation calculations of covalent molecules. *J. Chem. Theory Comput.* **2012**, *8* (8), 2669−2675.
(70) Tschumper, G. S. Multicentered integrated QM:QM methods for weakly bound clusters: An efficient and accurate 2-body:many-body treatment of hydrogen bonding and van der Waals interactions. *Chem. Phys. Lett.* **2006**, *427* (1), 185−191.
(71) Beran, G. J.; Nanda, K. Predicting organic crystal lattice energies with chemical accuracy. *J. Phys. Chem. Lett.* **2010**, *1* (24), 3480−3487.
(72) Liu, K.-Y.; Herbert, J. M. Understanding the many-body expansion for large systems. III. Critical role of four-body terms, counterpoise corrections, and cutoffs. *J. Chem. Phys.* **2017**, *147* (16), 161729.
(73) Mayhall, N. J.; Raghavachari, K. Molecules-in-Molecules: An extrapolated fragment-based approach for accurate calculations on large molecules and materials. *J. Chem. Theory Comput.* **2011**, *7* (5), 1336−1343.
(74) Saha, A.; Raghavachari, K. Analysis of different fragmentation strategies on a variety of large peptides: Implementation of a low level of theory in fragment-based methods can be a crucial factor. *J. Chem. Theory Comput.* **2015**, *11* (5), 2012−2023.
(75) Jose, K. V. J.; Raghavachari, K. Evaluation of energy gradients and infrared vibrational spectra through Molecules-in-Molecules fragment-based approach. *J. Chem. Theory Comput.* **2015**, *11* (3), 950−961.
(76) Jose, K. V. J.; Raghavachari, K. Fragment-based approach for the evaluation of NMR chemical shifts for large biomolecules incorporating the effects of the solvent environment. *J. Chem. Theory Comput.* **2017**, *13* (3), 1147−1158.
(77) Wu, F.; Liu, W.; Zhang, Y.; Li, Z. Linear-scaling time-dependent density functional theory based on the idea of "from fragments to molecule. *J. Chem. Theory Comput.* **2011**, *7* (11), 3643−3660.
(78) Li, Z.; Li, H.; Suo, B.; Liu, W. Localization of molecular orbitals: From fragments to molecule. *Acc. Chem. Res.* **2014**, *47* (9), 2758−2767.
(79) Li, H.; Liu, W.; Suo, B. Localization of open-shell molecular orbitals via least change from fragments to molecule. *J. Chem. Phys.* **2017**, *146* (10), 104104.
(80) Liu, J.; Herbert, J. M. Pair−pair approximation to the generalized many-body expansion: An alternative to the four-body expansion for ab initio prediction of protein energetics via molecular fragmentation. *J. Chem. Theory Comput.* **2016**, *12* (2), 572−584.
(81) Kohn, W. Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.* **1996**, *76* (17), 3168−3171.
(82) Prodan, E.; Kohn, W. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (33), 11635−11638.
(83) Humbel, S.; Sieber, S.; Morokuma, K. The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN_2 reaction: RCl + Cl^-. *J. Chem. Phys.* **1996**, *105* (5), 1959−1967.
(84) Svensson, M.; Humbel, S.; Froese, R. D.; Matsubara, T.; Sieber, S.; Morokuma, K. Oniom: A multilayered integrated MO + MM method for geometry optimizations and single point energy predictions. A test for Diels−Alder reactions and $Pt(P(t\text{-}bu)_3)_2 + H_2$ oxidative addition. *J. Phys. Chem.* **1996**, *100* (50), 19357−19363.
(85) Vreven, T.; Morokuma, K. On the application of the IMOMO (integrated molecular orbital + molecular orbital) method. *J. Comput. Chem.* **2000**, *21* (16), 1419−1432.
(86) Becke, A. D. Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J. Chem. Phys.* **1992**, *96* (3), 2155−2160.
(87) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37* (2), 785.
(88) Stephens, P.; Devlin, F.; Chabalowski, C.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98* (45), 11623−11627.
(89) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.
(90) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32* (7), 1456−1465.
(91) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *J. Chem. Phys.* **1997**, *107* (20), 8554−8560.
(92) Schmider, H. L.; Becke, A. D. Optimized density functionals from the extended g2 test set. *J. Chem. Phys.* **1998**, *108* (23), 9624−9631.
(93) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom−atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10* (44), 6615−6620.
(94) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li-F. *J. Comput. Chem.* **1983**, *4* (3), 294−301.
(95) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54* (2), 724−728.
(96) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77* (7), 3654−3665.
(97) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoret. Chim. Acta* **1973**, *28* (3), 213−222.
(98) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257−2261.
(99) Stewart, J. J. Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13* (12), 1173−1213.
(100) Řezáč, J.; Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **2012**, *8* (1), 141−151.

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394

(101) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16, Revision-a03*; Gaussian, Inc.: Wallingford, CT, 2016.

(102) Stewart, J. P. *MOPAC2016*; Stewart Computational Chemistry: Colorado Springs, CO, USA, 2016. http://openmopac.net/ (accessed Feb 8, 2018).

DOI: 10.1021/acs.jctc.7b01198
J. Chem. Theory Comput. 2018, 14, 1383−1394