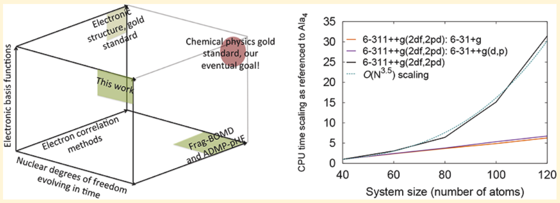
JCTC Journal of Chemical Theory and Computation
Article
Cite This: J. Chem. Theory Comput. 2018, 14, 5535−5552
pubs.acs.org/JCTC

Efficiently Capturing Weak Interactions in ab Initio Molecular Dynamics with on-the-Fly Basis Set Extrapolation

Timothy C. Ricard and Srinivasan S. Iyengar*

Department of Chemistry and Department of Physics, Indiana University, 800 E. Kirkwood Avenue, Bloomington, Indiana 47405, United States

ABSTRACT: Weak interactions have a critical role in accurately portraying conformational change. However, the computational study of these often requires large basis electronic structure calculations that are generally cost-prohibitive within ab initio molecular dynamics. Here, we present a new approach to efficiently obtain AIMD trajectories in agreement with large, triple-ζ, polarized valence basis functions, at much reduced computational cost. For example, it follows from our studies that AIMD trajectories can indeed be constructed in agreement with basis sets such as 6-311++G(2df,2pd) with computational effort commensurate with those from much smaller basis sets such as 6-31+G(d), for polypeptide systems with 100+ atoms. The method is based on molecular fragmentation and allows a range-specified repartitioning of intramolecular (and potentially intermolecular) interactions where noncovalent interactions are selectively assembled using a piece-wise reconstruction based on a set-theoretic inclusion−exclusion principle generalization of ONIOM. Through a simplex decomposition of molecular systems the approach efficiently provides the necessary many-body interactions to faithfully represent noncovalent interactions at the large basis limit. Conformational stabilization energies are provided at close to the complete-basis limit at much reduced cost, and similarly AIMD trajectories (both Born−Oppenheimer and Car−Parrinello-type) are obtained in agreement with very large basis set sizes, in an extremely efficient and accurate manner. The method is demonstrated through simulations on polypeptide fragments of a variety of sizes.



I. INTRODUCTION

Classical molecular dynamics has proved to be a major workhorse in the study of complex chemical and biochemical problems.[1−9] But, the use of experimentally parametrized force fields in classical MD deeply limits its applications to mostly nonreactive, equilibrium systems. Exceptions include the empirical valence bond theory (EVB)[10−13] and reactive-force-fields such as ReaxFF[14] where potentials are tailored for specific reactive applications, and it is in the sense of providing a general protocol for reactive systems that ab initio molecular dynamics (AIMD)[15−22] has had the greatest impact. With the application of density functional theory to quantum chemistry[23,24] moderate sized systems have been readily studied with AIMD.[25−31] Here, the electronic structure calculations are performed at every time step, and this greatly limits the routine use of AIMD for complex chemical problems, with DFT being the only practical and affordable choice. However, despite great progress, several challenges remain for use of DFT methods.[23,32−34] In this regard, we have recently developed new methods that employ molecular fragmentation[35−37] and geometric networks[38] to perform AIMD calculations with MP2[35,36,38] and CCSD[37] accuracy at DFT cost.[37,38]

While these methods may influence computational chemical modeling of complex problems, it is also clear from Figure 1a that accurate quantum chemical and AIMD calculations are only possible if we also simultaneously consider basis set size effects. For example, hybrid functional DFT methods formally scale as $\mathcal{O}(N^4)$, with the size of the electronic basis set, N, and this scaling is effectively reduced in larger systems to $\mathcal{O}(N^{3.5})$ due to reuse of two-electron integrals;[39] this critically affects the size of basis sets that can be routinely employed in large-scale computational simulations. Indeed, as noted in Figure 1, the gold standard for electronic structure theory resides on the two-dimensional space of basis set size and electron-correlation methods, with the correct answer on the top-right corner of Figure 1a. Ab initio molecular dynamics methods may need this accuracy to be computed in a dynamical fashion, as represented in Figure 1b, which enormously complicates the problem due to the sheer number of such calculations to be performed. In this paper we extend our fragment-based dynamics methods,[35−38] both extended Lagrangian[21,22,36,37] and Born−Oppenheimer[17,18,35,37,38] versions, to provide accurate dynamics in the larger basis-set limit, at much reduced computational cost. *In fact, a critical hallmark of the study here is that we are able to perform AIMD trajectory calculations with accuracy comparable to large triple-ζ basis functions that include polarization and diffuse functions using much smaller basis functions deeply reducing computational cost and scaling.* We utilize a graph-theoretic adaptation[38] of the generalization to the well-known ONIOM[40] method that uses the set-theoretic inclusion−exclusion principle[35−37] to construct AIMD trajectories that are accurate in the large basis set limit. With this work we advance a larger goal of combining basis set fragment-based extrapolation with electronic structure fragment-based extrapolations to obtain greater AIMD accuracy at reduced computational costs. This compound extrapolation will, in the future, be used to propagate

Received: August 2, 2018
Published: October 18, 2018

DOI: 10.1021/acs.jctc.8b00803
J. Chem. Theory Comput. 2018, 14, 5535−5552

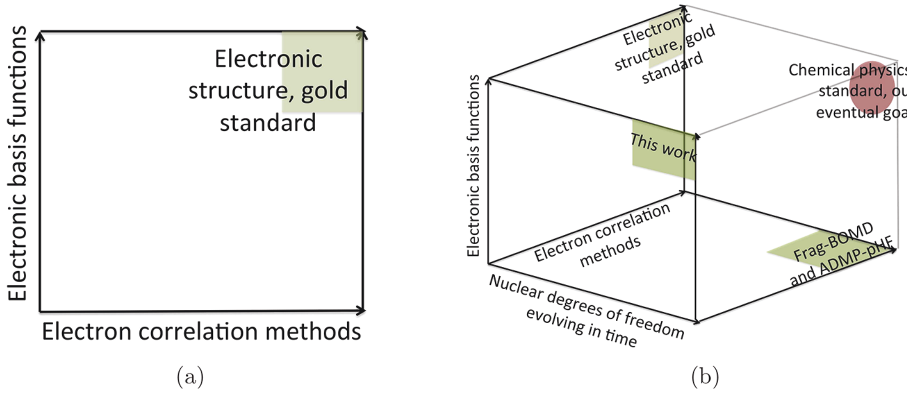
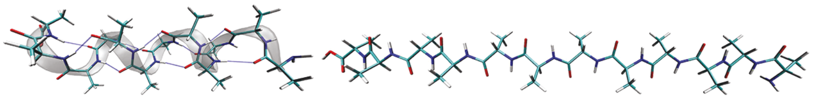

Electronic basis functions
Electronic structure, gold standard
Electron correlation methods
(a)
Electronic structure, gold standard
Chemical physics gold standard, our eventual goal!
Electronic basis functions
This work
Electron correlation methods
Frag-BOMD and ADMP-pHF
Nuclear degrees of freedom evolving in time
(b)

Figure 1. Gold standard for electronic structure is shown in part (a) which becomes critical for systems with weak interactions where both large basis sets and electron-correlation are needed to obtain accurate results. Part (b) adds nuclear degrees of freedom to the problem, where our previous work (Frag-BOMD and ADMP-pHF)[35−38] and this publication are depicted. The red ellipse represents a future goal.



(a)3_{10} helix ($\phi = -60^o$, $\psi = -20^o$) (b)β-strand ($\phi = -160^o$, $\psi = 160^o$)

Figure 2. 3_{10} helical (a) and β-strand (b) conformers for Ala_{12}. In part (a), the stabilizing ($i \rightarrow i + 3$) hydrogen bonds are shown.

the nuclear degrees of freedom forward, the overall vision presented in Figure 1b.

This paper is organized as follows: In Section II we examine the effects of basis set choice in obtaining conformational stabilization energy between various conformers of polyalanine. In Section III we discuss our approach for "on-the-fly" basis set extrapolation, which is then benchmarked in Section IV to yield accurate stabilization energies at much cheaper cost as compared to the larger basis set limit. We also discuss our dynamical simulations in Section V, and in Section VI we present our conclusions.

II. BASIS SET CONTRIBUTIONS TO WEAK INTERACTIONS THAT DEFINE CONFORMATIONAL CHANGE

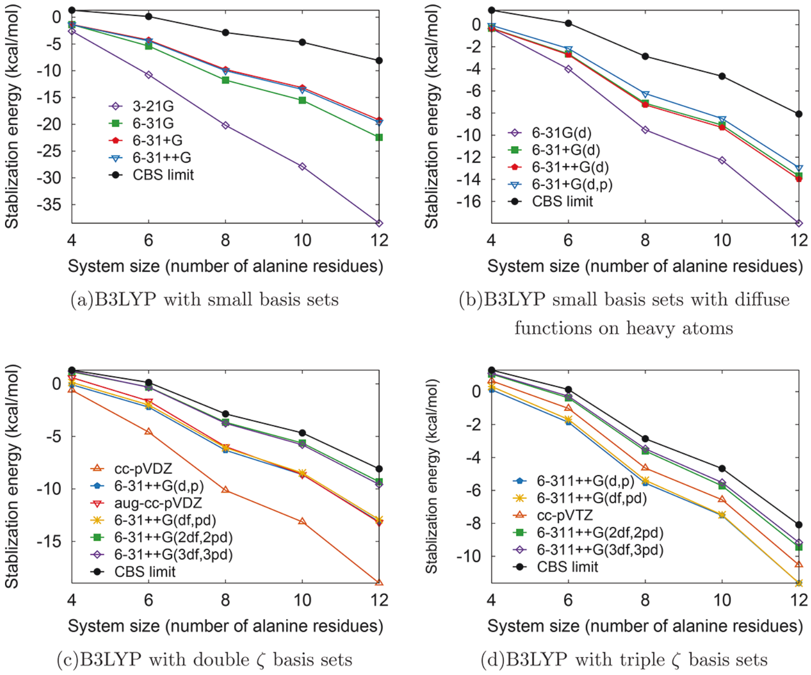
In this section, we gauge the effect of basis set composition on conformational stabilization energy. We conduct a survey of B3LYP level conformational stabilization energies for a range of polyalanine systems (Ala_4–Ala_{12}) using a set of Gaussian basis functions. The starting structures were obtained from optimization at the B3LYP/6-31++G(d,p) level of theory to obtain the β-strand and 3_{10} helical structures shown in Figure 2 for Ala_{12}. The β-strand is characterized by peptide dihedral angles, $\phi = -160°$ and $\psi = 160°$, and represents one strand of a β-sheet, while the 3_{10} helix is characterized by $\phi = -60°$ and $\psi = -20°$ with hydrogen bonding between amino acids indexed ($i \rightarrow i + 3$). The conformational stabilization energy between these structures is defined here as the difference in electronic energies for the 3_{10} helical system and the β-strand system. We approximated the stabilization energy at the complete basis set limit through the exponential extrapolation scheme[41] of Dunning style basis sets.[42] Figure 3a shows the conformational stabilization energy calculated with B3LYP on smaller bases, many of which are commonly used for dynamics and structural calculations in biochemical systems.[43,44] The inclusion of polarized and diffuse functions significantly contributes to the stability in these structures; the addition of polarized or diffuse functions on the heavy atoms marked an improvement of 3–5 kcal/mol toward the complete basis set limit. In general, all of the basis sets in Figure 3a show great discrepancy with respect to the CBS limit.

Next we considered fully polarized double-ζ (Figures 3b,c) and triple-ζ (Figure 3d) bases with increasing number of diffuse functions that are generally cost-prohibitive to utilize for AIMD simulations of biochemically relevant problems. Here, as we add diffuse functions to the Pople style Gaussian basis sets,[46] we observe monotonic convergence toward the CBS limit with the exception of a small deviation where 6-31++G(2df,2pd) is marginally closer to the CBS limit as compared to 6-31++G-(3df,3pd) (by a few tenths of a kcal/mol). From this analysis we would consider double and triple ζ basis sets with (2df,2pd) or (3df,3pd) diffuse functions as fair approximations to the CBS limits shown here. But these calculations present a steep computational cost for AIMD. Density functional methods, such as the B3LYP functional that we employ here, formally scale as $O(N^4)$ and hence a choice of more exhaustive basis would significantly hinder the feasibility as system size grows.

We also conducted studies similar to those in Figure 3, but with dispersion corrected[45] B3LYP, and observed similar basis set dependent trends as those in Figure 3. The associated results are summarized in Figure 4 using the B3LYP-D3 functional[45] with the Grimme dispersion correction with additional details in Appendix A (see Figures A-1 and A-2). Inclusion of dispersion corrections significantly increase the stability of the helical conformations relative to the β-strand. In the next section, to retain the quality afforded by use of a larger basis set, but with much reduced computational expense, we propose the use of molecular fragmentation to obtain AIMD trajectories and conformational stabilization energies.

III. GRAPH-THEORETIC AND SET-THEORETIC GENERALIZATIONS TO ONIOM FOR BASIS SET EXTRAPOLATION AND AB INITIO MOLECULAR DYNAMICS

When nonorthogonal atom-centered Gaussian basis sets are used, the basis functions localized on one atom are



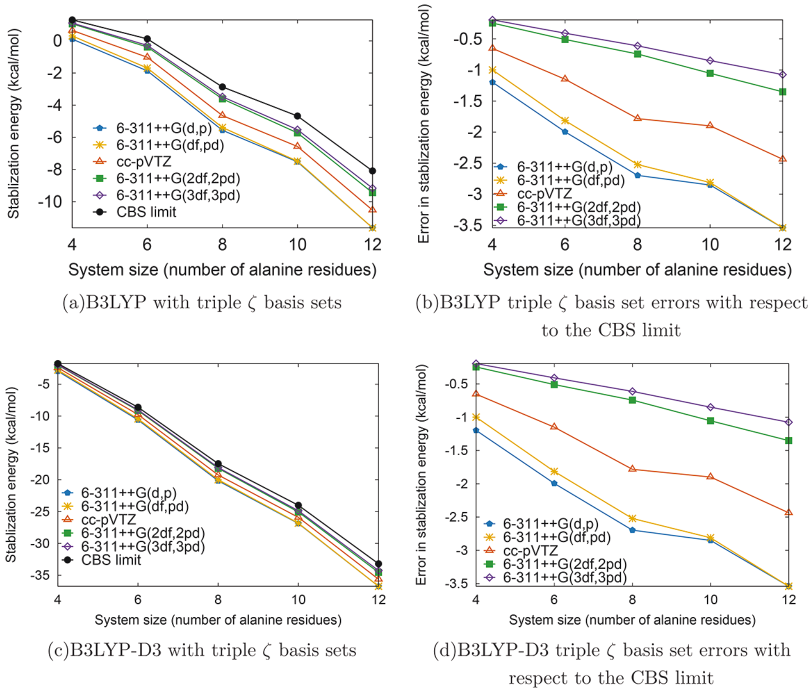
(a)B3LYP with small basis sets

(b)B3LYP small basis sets with diffuse functions on heavy atoms

(c)B3LYP with double ζ basis sets

(d)B3LYP with triple ζ basis sets

Figure 3. Stabilization energy dependence on the choice of basis. The energy differences between the 3_{10} helical and β-strand conformers are shown for a range of polyalanine systems with increasing basis set coverage. Parts (a) and (b) show the smaller basis sets commonly used in electronic structure calculations and AIMD. Part (c) shows the stabilization energy for double ζ basis sets with increasing number of diffuse functions. Part (d) shows triple-ζ basis sets with polarization functions and increasing number of diffuse functions. The basis sets, 6-311++G(2df,2pd) and 6-311++G(3df,3pd), are treated as good target approximations for the remaining part of the paper. The errors for these with respect to CBS limit are presented in Figure 4 with similar results for the B3LYP-D3 functional.[45] All CBS extrapolations were constructed using the cc-pVDZ, cc-pVTZ, and cc-pVQZ basis sets through the exponential extrapolation scheme[41] of Dunning style basis sets.[42]



(a)B3LYP with triple ζ basis sets

(b)B3LYP triple ζ basis set errors with respect to the CBS limit

(c)B3LYP-D3 with triple ζ basis sets

(d)B3LYP-D3 triple ζ basis set errors with respect to the CBS limit

Figure 4. Errors in stabilization energy of the 3_{10} helical structure with respect to the β-strand compared to similar stabilization energy from triple ζ basis set calculations. Parts c and d include Grimme's dispersion corrections[45] while parts a and b do not include these corrections. Parts (b) and (d) are, however, very similar because the Grimme corrections[45] only depend on nuclear positions and not on the electronic structure. Hence the effect of these corrections is much reduced when differences are computed with respect to the CBS limit. More details can be found in Appendix A.

supplemented by contributions from basis functions that are localized on neighboring atom centers while representing the local region. As a result, one encounters the so-called "electron density leakage" problem where electron density is said to "leak" from the basis functions that belong to one atom to the basis functions on neighboring atoms.[47] This leakage leads to

DOI: 10.1021/acs.jctc.8b00803

an extension of the basis function space used to represent molecular wave functions. Basis set extension is beneficial as it expands the functional space for each atom. But from this additional expansion of the available basis set functional space arises the weakness that the availability of this expanded space is dependent on the relative atomic positions which leads to the well-known basis set superposition error (BSSE)[47−49] for reaction energies and in dynamics calculations.[50] These errors arise from the difference in the effective basis function space available between the composite system (e.g., A-B), the constituents considered separately (A and B), or for some intermediate separation between A and B, which may be encountered during dynamics calculations or during a reaction path calculations. In the composite system the vector space available to A is expanded from the basis sets centered on B and vice versa. But when the distance between A and B changes the available extension to the vector space for both systems are perturbed. This leads to changes to the quality of the effective basis set. Although BSSE is often associated with interaction energies, this concern appears in other studies as well. In ab initio molecular dynamics,[50] each time step modifies the distance between atoms which would in turn present a different linear vector space for determination of the electronic wave function at every instant in time during dynamics.[50] Such perturbations lead to artifacts in the potential energy surface sampled by the system during dynamics by introducing additional oscillations that represent in the electronic energies and wave functions that bear the signature of nuclear motion. However, as the quality and size of the basis set is increased, these artifacts disappear and hence larger basis sets are more desirable for electronic structure and "on-the-fly" dynamics, especially when weak interactions are involved.

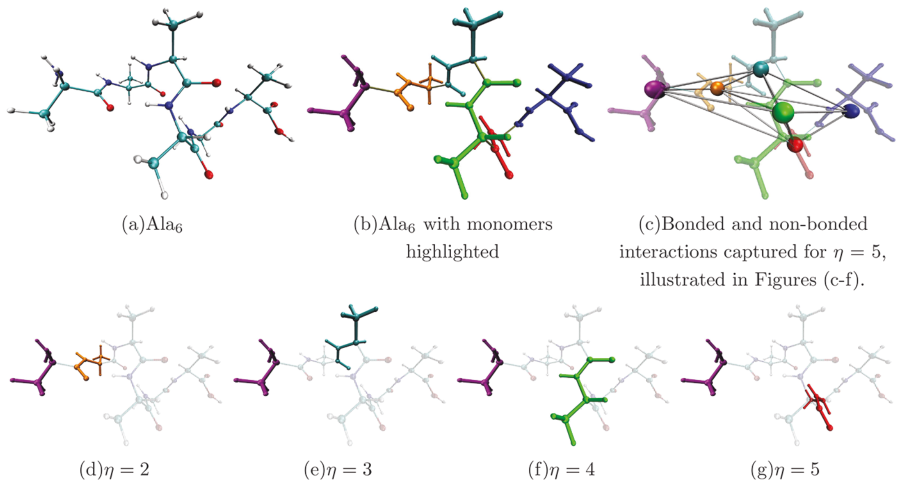
As a result of the above discussion and the improved results found in Section II with increasing quality and size of basis sets, we propose to adapt the principle of inclusion exclusion generalization of ONIOM[35−37] and the graph-theoretic/geometric analogue of the same,[38] to enhance the quality of basis functions used in AIMD at reduced computational cost. Although the formalism here is derived from ONIOM,[58] it also has close connections to other methods including the Multicentered QM:QM formalism,[59,60] the molecular tailoring approach (MTA),[61,62] the ONIOM-XO method,[63] and the molecules-in-molecules (MIM) methodology.[64−67] Indeed there are several other fragmentation methods[68−73] available, but the approaches in refs 35−38, 59, and 62−66 include long-range electronic effects through a full-system low level calculation, much in the same vein as the ONIOM[58] method. Furthermore, some of the fragmentation methods have been used for basis set extrapolations[60,62,74] and for AIMD.[35−38,75−81] We have noted in ref 38 that the approach studied here is also closely related to many-body expansions[82−84] and double many-body expansions.[82] At this point, it is also critical to note that other complementary approaches for basis set extrapolation include the dual basis methods[85] and multistep basis set partitioning[86] scheme.

In the method discussed in this publication, we partition our systems into orthogonal units, which will be referred to as monomer units. In ref 38, these monomer units are referred to as *CG-nodes* in a graph, since in a sense these units together represent a "coarse-grained" form of the system. These monomers are then connected to obtain dimer units that are represented as *edges* in a graph or a geometric network, the elements of which are a union of the elements within the two connected monomer units. In ref 38 we have used two edge-construction techniques. (a) In one case, the Delaunay triangulation method[87−92] was used, which allows for an orthogonal partitioning, or "tiling", of the molecular space. Delaunay triangulation is a dual-representation of Voronoi diagrams[87,88] and has been employed in a variety of other applications.[93−95] In ref 38, Delaunay triangulation provides the edges that connect monomers to obtain a simplicial complex[96] or a connected graph that depicts the molecular framework in a coarse-grained fashion. These edges are then used to construct dimer fragments, and potential higher order fragments to construct a molecular fragmentation procedure. (b) In the second approach introduced in ref 38, we have the flexibility to include all possible many-body interactions inside a local neighborhood which does not lead to an orthogonal partitioning (as in Delaunay triangulation) but is found to be numerically superior in ref 38. In essence, we define a local (chemically connected and spatial) neighborhood represented by a parameter, η, over which all possible edges (and hence second-order many-body interactions) are included, a pictorial illustration for which is provided in Appendix B. However, within this approach, higher-order interactions may also be included and these are for example represented through the contributions from *faces* and *tetrahedrons*, or simplexes of ranks-0 (*nodes*), rank-1 (*edges*), rank-2 (*faces*), and rank-3 (*tetrahedrons*). In principle, this manner of defining fragments will allow us to systematically and adaptively obtain a many-body expansion and create an isomorphism between fragment definitions and the geometric coarse-graining algorithm described in ref 38.

The parameter η, discussed above,[38] refers to the spatial extent over which many-body interactions are included. For example, for $\eta = 2$, interaction between adjacent monomer units are included. When $\eta = 3$ additional interactions are considered between the monomer units (*nodes*) that are part of *edges* that intersect at one *node*, and so on. (see Appendix B and also Figure 5 for an illustration of the interactions included from the parameter η). Below we present the energy expression from both set-theoretic and graph-theoretic decompositions. Using the PIE-ONIOM scheme from ref 35, we may construct a fragment-based treatment of a partitioned system where the full system is treated with a smaller basis ($N_{B,S}$), while each fragment (such as the edges and nodes in Appendix B) is considered with a larger ($N_{B,L}$) and smaller ($N_{B,S}$) basis:

$$\begin{aligned} E^{PIE-ONIOM} &= E^{N_{B,S}} + \sum_{i=1}^{n} \mathcal{S}(i) - \sum_{1\le i<j\le n} \mathcal{S}(i \cap j) \\ &\quad + \sum_{1\le i<j<k\le n} \mathcal{S}(i \cap j \cap k) - \cdots \\ &\quad + (-1)^{n-1} \sum \mathcal{S}(1 \cap \cdots \cap n) \end{aligned} \tag{1}$$

Here, $E^{N_{B,S}}$ is the energy for the full system with a smaller basis set for any level of electronic structure theory. The indices i, j, k, $\cdots$, n are the dimer units represented as ellipses and as edges in Appendix B and are referred to as primary fragments. The overlapping fragments are formed by the intersection ($\cap$) of the primary fragments and lead to the nodes in Appendix B. In this publication, we use the monomer units to represent single amino acid units, but this is not a hard and fast requirement; furthermore higher-order many-body interactions can be easily included as will be seen below. (More precisely, the monomer units are chosen as CHR−NH−CO peptide units to retain the partial double bond character[97] of the peptide bond.


(a)Ala_6

(b)Ala_6 with monomers highlighted

(c)Bonded and non-bonded interactions captured for $\eta = 5$, illustrated in Figures (c-f).

(d)$\eta = 2$

(e)$\eta = 3$

(f)$\eta = 4$

(g)$\eta = 5$

Figure 5. An illustration of the parameter η for a 3_{10} helical Ala_6 structure. Monomer units (or *nodes*) are shown in part (b), and these were used to form dimer fragments according to the geometric network in part (c). Examples for the various interactions captured are shown in parts (d)−(g). However, note that these interactions are cumulative in the sense that η = 4 captures interactions all the way from η = 1 through η = 4.

But these details are considered in Section IV.A.) The terms $\mathcal{S}(\cdots)$ are corrections to the fragment energy obtained as in refs 35−37, in the spirit of ONIOM,[40] according to

$$\mathcal{S}(i) = E^{N_{B,L}}(i) - E^{N_{B,S}}(i) \tag{2}$$

Alternatively, the graph-theoretic method leads to the expression:

$$E^{graph\text{-}theoretic} = E^{N_{B,S}} + \sum_{r=1}^{R} (-1)^r \left\{ \sum_{\alpha} \mathcal{S}(\alpha, r) \left[\sum_{m=r}^{R} (-1)^m p_\alpha^{r,m} \right] \right\} \tag{3}$$

where the quantity "α" in the second term in eq 3 represents a summation index over all rank-r simplexes embedded within the geometric network decomposition shown in Appendix B, with R being the largest rank simplex considered for electronic structure treatment. The quantity $\mathcal{S}(\alpha, r)$ is a difference in energy, analogous to $\mathcal{S}(i)$ in eq 2, and defined as

$$\mathcal{S}(\alpha, r) = E^{N_{B,L}}(\alpha, r) - E^{N_{B,S}}(\alpha, r) \tag{4}$$

Thus, $\mathcal{S}(\alpha, r)$ is the difference in energy between the larger and smaller basis set treatment of the αth simplex of rank r. The second term, in eq 3, is the summation of the energy contributions of progressively higher rank simplexes. As in eq 1, $E^{N_{B,S}}$ is the energy for the full system with a smaller basis set. The square bracketed term contains the overcounting correction from ref 38 for these simplexes, where $p_\alpha^{r,m}$ is the number of times the αth rank-r simplex appears in simplices of rank-m ($R \geq m \geq r$). If only edge (dimer) interactions are considered (as is done in numerical tests in this paper), R = 2 and eq 3 is truncated at the *edge*-contribution level; hence,

$$E_{R=2}^{graph\text{-}theoretic} = E^{N_{B,S}} + \sum_{\alpha} \mathcal{S}(\alpha, 2) - \sum_{\alpha} \mathcal{S}(\alpha, 1)[p_\alpha^{1,2} - p_\alpha^{1,1}] \tag{5}$$

where $p_\alpha^{1,2}$ is the number of times node α appears in an edge and $p_\alpha^{1,1}$ is 1 as each node is unique.

While eqs 3 and 5 are complementary ways of performing the computation in eq 1, eqs 3 and 5 provide a numerical approach that is much more efficient as discussed in ref 38. Equation 1 is closely related to previous fragmentation methods, MIM[65] and MTA,[61] but places fragmentation within the context of the well-known ONIOM approach. But the starting point for eq 3 shows a much deeper connection to many body expansions,[82] where N-body terms can be selectively added to the expression in a systematic fashion within a graphical representation.

In summary, there are essentially two dimensions where accuracy can be systematically improved and fragmentation can grow. (a) For a given truncation order of the generalized graph-theoretic version in eq 3, η is an expansion coefficient that allows us to spatially "tune-in" many-body interactions. In this paper, we exhaustively study all interactions up to η = N, that is the maximum number of two-body large basis corrections (including long-range interactions) to the small basis calculation. (b) The other dimension where fragmentation may be improved is through introducing energy contributions from *faces*, *tetrahedrons* and other embedded *simplexes* within the geometric network that depicts the molecule (see Appendix B) as outlined in eq 3. This will allow including three, four and higher number of monomers that are again required to be expanded using a generalized form of η, but we do not benchmark this generalization in this paper and find that the η-dependent edge interactions sufficiently describe the systems considered here.

For the remaining part of this paper, we employ the network decomposition in eq 5 to construct on-the-fly basis set extrapolation.

IV. BASIS SET EXTRAPOLATION FOR POLYALANINE SYSTEMS USING EQUATION 5

In Section IV.A we present a brief description of monomer (or node) selection and associated fragmentation. Figure 5 provides a complementary illustration that highlights the extent to which two-body basis-set contributions are captured. The fragmentation ideas in Section IV.A complement the corresponding discussion in ref 38 and hence the description is brief. In Section IV.B, we probe the accuracy in computing DFT level electronic conformational stabilization energies in close agreement with large basis sets using the formalism discussed in Section III. We consider polypeptide chains of lengths in the range (Ala_4−Ala_{12}) and evaluate the stabilization energies between helical and straight chain conformations as done in Section II but now with basis-set extrapolation (described in eqs 1, 3, and 5). The efficiency gain from this method is discussed in Section IV.C. Helical stability[98] in polyalanine relies on nonbonding interactions,[99−101] which are prone to basis set superposition error,[50,102,103] and is hence chosen as part of our studies.

We introduce an extended Lagrangian generalization to basis-set-extrapolated-AIMD using the energy functional in eq 5, and this done in Section V. We evaluate the associated quality, accuracy, and efficiency of AIMD (Born−Oppenheimer and extended-Lagrangian) trajectories in Section VI. *It follows from our results in Sections IV.C and VI that AIMD trajectories can indeed be constructed in agreement with larger basis set calculations (such as 6-311++G(2df,2pd)) with computational expense commensurate with those from much smaller basis sets (such as 6-31+G(d)).*

IV.A. Fragmentation Protocols for Polyalanine Systems. Here we consider the conformational stability of Ala_N between its linear and helical conformations. Details of these structures were discussed in Section II and are illustrated in Figure 2. The comparison of accuracy of conformational stability between the linear and 3_{10} helical structures allows us to gauge the accuracy of the methods discussed in Section III toward potentially capturing *dynamical* nonbonded interactions.

Before proceeding with our evaluation of the method for the extrapolation of basis sets, the monomers would need to be defined. Here we chose to partition the polyalanine structures by breaking the bond connecting the C_α atom and the carboxyl carbon atom, thus forming monomers of the kind, CHR-NH−CO. The carbon−carbon bond is broken instead of the peptide bond, NH−CO, as the latter is known to have a partial bond character.[97] When these monomers are defined, bonds are broken creating dangling valencies. These dangling valencies are saturated by use of hydrogen link atoms, consistent with the ONIOM[40] methodology. Inline with refs 30, 104, and 105, the forces on link atoms are transformed back to the real system atoms using the appropriate Jacobians. Next, key interactions between monomers are considered, these interactions are represented by dimer fragments. The interaction distance here is quantified by the sequential displacement between the C_α along the backbone of the peptides considered. The parameter η allows us to tune in the extent of nonbonded interactions considered. For example, $\eta = 2$ implies that only covalently connected dimer units are considered. For $\eta = 3$, noncovalent interactions between, for example, monomer units numbered "1" and "3" are also included, when units "1" and "2" are covalently connected, and "2" and "3" are also covalently connected. (See Figure 5.) Similarly $\eta = 4$ captures a longer range nonbonded interaction, and so on. Most calculations in this paper include nonbonded interaction up until $\eta = 4$ and in some cases $\eta = 5$. In this fashion we are able to tailor-in critical nonbonding interactions. Hence $\eta = n$ implies that there are a total of $(n - 1)$ intervening amino acid monomers that are chemically connected en route to form the dimer fragment and thus depicting the extent of *through space nonbonding interaction* captured in the study. An example of this scheme applied to a 6-alanine helix is seen in Figure 5, where Figure 5a shows the partition of the full system into monomers, Figure 5d−g shows one example dimer which is added to the overall set of dimers with the increase in η. Arising from the graph-theoretic description in ref 38, it is possible to generalize this same idea to trimers, tetramers, etc., but this was not found to be necessary for the applications discussed in this paper.

Table 1. Errors in Conformational Stabilization Energy, Equation 6, for All-Dimer ($\eta = N$) Calculations, with Respect to Double-ζ Basis Functions[a]

	$N_{B,L}$				
$N_{B,S}$	6-31+ +G(d,p)	6-31+ +G(df,pd)	6-31+ +G(2df,2pd)	6-31+ +G(3df,3pd)	aug-cc-pVDZ
			Ala_4		
3-21G	0.782	0.784	0.842	0.875	0.874
6-31G	0.38	0.383	0.441	0.473	0.472
6-31+G	**−0.05**	**−0.048**	**0.011**	**0.044**	**0.042**
6-31G(d)	0.559	0.562	0.62	0.653	0.652
6-31+G(d)	**0.012**	**0.014**	**0.073**	**0.105**	**0.103**
			Ala_6		
3-21G	2.979	3.005	2.978	3.145	2.921
6-31G	1.586	1.611	1.584	1.751	1.528
6-31+G	**−0.213**	**−0.187**	**−0.215**	**−0.048**	**−0.271**
6-31G(d)	2.098	2.124	2.096	2.264	2.041
6-31+G(d)	**0.009**	**0.035**	**0.007**	**0.174**	**−0.048**
			Ala_8		
3-21G	4.946	4.976	4.835	5.138	4.830
6-31G	2.344	2.375	2.234	2.536	2.229
6-31+G	**−0.480**	**−0.449**	**−0.590**	**−0.288**	**−0.596**
6-31G(d)	3.344	3.374	3.233	3.536	3.228
6-31+G(d)	**0.036**	**0.065**	**−0.076**	**−0.180**	**0.497**
			Ala_{10}		
3-21G	5.939	5.961	5.756	6.076	5.482
6-31G	3.233	3.255	3.050	3.370	2.776
6-31+G	**−0.823**	**−0.802**	**−1.006**	**−0.686**	**−1.281**
6-31G(d)	4.507	4.529	4.324	4.644	4.050
6-31+G(d)	**0.142**	**0.164**	**−0.041**	**0.279**	**−0.316**

[a] All errors are in kcal/mol. The columns represent the target basis set, $N_{B,L}$, whereas the rows represent the lower level basis, $N_{B,S}$. Rows in bold show significantly lower errors, and the corresponding $N_{B,S}$ are used later for AIMD simulations.

IV.B. Isomer Stabilization Energies from Equation 5. We begin with a test set of five smaller Pople style basis sets, namely, 3-21G, 6-31G, 6-31+G, 6-31G(d), and 6-31+G(d), and use the results from these basis sets to extrapolate, using eq 5, to much larger basis functions such as 6-311++G(2df,2pd) and 6-311++G(3df,3pd). As noted in Figure 3c,d, the accuracy from these larger basis set calculations is generally within 1 kcal/mol with respect to the corresponding CBS limit. In Tables 1 and 2 we apply eq 5 to obtain the error in the conformational stabilization for transitions between the 3_{10} helix and the β-strand conformations. The error is computed as

$$\Delta E_{error} = (E_{3_{10}\text{-}helix}^{graph\text{-}theoretic} - E_{\beta\text{-}strand}^{graph\text{-}theoretic}) - (E_{3_{10}\text{-}helix}^{N_{B,L}} - E_{\beta\text{-}strand}^{N_{B,L}}) \tag{6}$$

where ΔE_{error} is reported in Tables 1 and 2 and in Figures 3c,d. The quantity, $E^{graph\text{-}theoretic}$, is the system energy obtained from eq 5, and $E^{N_{B,L}}$ is the energy computed using the large basis, $N_{B,L}$. These benchmarks presented here are performed with the B3LYP density functional, using all possible dimer fragments ($\eta = N$, see discussion in Section III). The errors in basis set extrapolations from the lower basis sets ($N_{B,S}$) to larger basis sets ($N_{B,L}$) are shown in Table 1 for larger double-ζ basis sets and in Table 2 for larger triple-ζ basis sets. Based on these tables, it may be concluded that the treatment presented here

Table 2. Errors in Conformational Stabilization Energy, Equation 6, for All-Dimer ($\eta = N$) Calculations, with Respect to Triple-ζ Basis Functions[a]

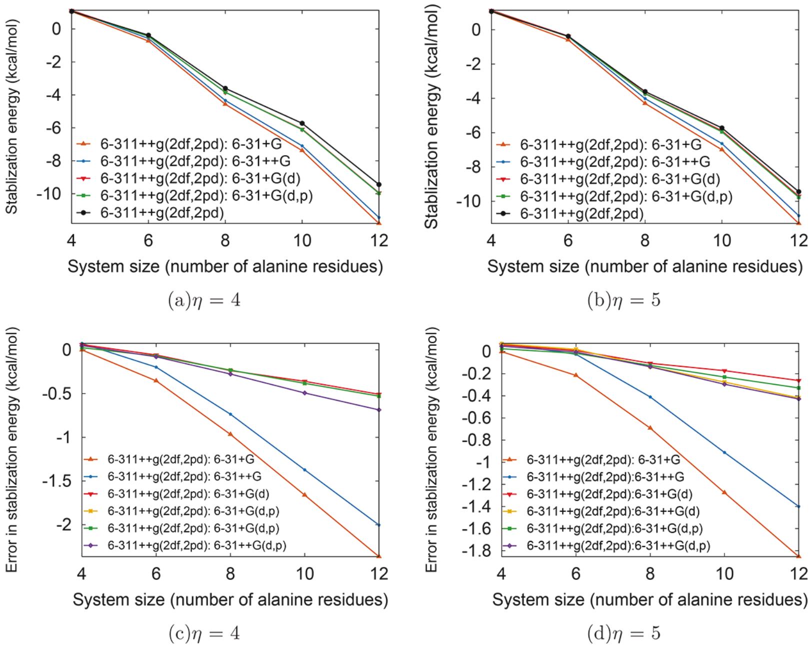
	$N_{B,L}$				
$N_{B,S}$	6-311+ +G(d,p)	6-311+ +G(df,pd)	6-311+ +G(2df,2pd)	6-311+ +G(3df,3pd)	cc-pVTZ
			Ala_4		
3-21G	0.782	0.772	0.830	0.818	0.349
6-31G	0.381	0.37	0.429	0.416	−0.052
6-31+G	**−0.048**	**−0.060**	**−0.001**	**−0.013**	−0.482
6-31G(d)	0.56	0.549	0.608	0.596	0.127
6-31+G(d)	**0.013**	**0.001**	**0.061**	**0.048**	−0.421
			Ala_6		
3-21G	2.857	2.793	2.981	2.832	1.134
6-31G	1.464	1.400	1.588	1.439	−0.260
6-31+G	**−0.336**	**−0.398**	**−0.211**	**−0.360**	−2.059
6-31G(d)	1.976	1.912	2.100	1.951	0.253
6-31+G(d)	**−0.114**	**−0.176**	**0.012**	**−0.137**	−1.837
			Ala_8		
3-21G	4.66	4.558	4.848	4.64	1.769
6-31G	2.058	1.957	2.246	2.039	−0.833
6-31+G	**−0.765**	**−0.867**	**−0.577**	**−0.785**	−3.657
6-31G(d)	3.057	2.956	3.246	3.038	0.167
6-31+G(d)	**−0.250**	**−0.353**	**−0.062**	**−0.270**	−3.141
			Ala_{10}		
3-21G	5.464	5.317	5.763	5.498	1.852
6-31G	2.759	2.612	3.057	2.792	−0.853
6-31+G	−1.298	−1.445	−1.000	−1.264	−4.910
6-31G(d)	4.032	3.885	4.331	4.066	0.420
6-31+G(d)	**−0.333**	**−0.480**	**−0.034**	**−0.299**	−3.945

[a] All errors are in kcal/mol. The columns represent the target basis set, $N_{B,L}$, whereas the rows represent lower level basis, $N_{B,S}$. Rows in bold show significantly lower errors, and the corresponding $N_{B,S}$ are used later for AIMD simulations.

shows significant accuracy when the smaller basis contains additional diffuse functions on heavy atoms ($N_{B,S}$ = 6-31+G), which then is further improved when polarization functions are included ($N_{B,S}$ = 6-31+G(d)). Polarization functions without diffuse functions struggle to capture the stabilization energy. As one would expect, the choice of the smaller basis size primarily affects the extrapolations for the 3_{10} helical structure rather than β-strand. This is because the diffuse and polarized functions would aid in capture of nonbonded, hydrogen-bonding interactions that stabilize the helix. The β-strand structures would primarily have electron density leakage to adjacent atoms in the amino acid chain, which would suffer less perturbation during conformational change and dynamics.

The above analysis shows that within the chosen test set, 6-31+G(d) offers adequate chemical accuracy toward extrapolations to the larger basis sets. The addition of polarization functions at the lower level generally refines the extrapolation, but without the diffuse functions at the low level, the extrapolation remains inaccurate in almost all cases. The combination of polarization and diffuse functions, on the heavy atoms, allows extrapolations to subkcal/mol accuracy to the basis calculations up to basis sets such as 6-311++G(3df,3pd).

We next restrict the extent of long-range interactions by reducing the size of η. We consider η = 4, 5 (see Figure 5), to tailor the nonbonded interactions. A new test set of lower basis sets was also chosen, namely, 6-31+G, 6-31++G, 6-31+G(d), 6-31++G(d), 6-31+G(d,p), and 6-31++G(d,p), based upon our discussion above. Due to the quality of conformational stabilization energies seen in Section II and the previous extrapolation accuracy, 6-311++G(2df,2pd) was selected as the target larger basis set. Figure 6 illustrates the calculated stabilization energy for extrapolations to 6-311++G(2df,2pd) for η = 4 and 5. Consistent with the above studies, the inclusion of polarization functions on heavy atoms offers additional



(a)$\eta = 4$ (b)$\eta = 5$

(c)$\eta = 4$ (d)$\eta = 5$

Figure 6. Conformational stabilization energy approximated to the 6-311++G(2df,2pd) basis (black line) using eq 5. Parts (a) and (b) show the stabilization energy (difference in energy between the 3_{10}-helix and β-strand conformers). Parts (c) and (d) show the corresponding error from using eq 5, in comparison with the larger basis calculation, as defined by eq 6. The errors extrapolating to 6-311++G(2df,2pd), 6-31++G(2df,2pd), and 6-311++(3df,3pd) using dispersion corrected DFT are presented in Figures A-3, A-4, and A-5. Based on these calculations it appears that the pair 6-31+G(d) → 6-311++G(2df,2pd) provides the best choice for extrapolation and hence will be pursued as a part of the AIMD studies.

DOI: 10.1021/acs.jctc.8b00803
J. Chem. Theory Comput. 2018, 14, 5535−5552

Table 3. Energy Conservation Properties for Dynamical Simulations (micro-canonical)

initial config.	$N_{B,S}$[a]	sim. time[b]	T_{Ave} (K)[c]	$\Delta\mathcal{H}$[d]	$\mathcal{H}_{\mathrm{Drift}}$[e]
		Ala_3			
BOMD 3_{10} helix		7.79 ps	336.73 ± 35.62	0.02	−0.01
Frag-BOMD 3_{10} helix	6-31+G	7.32 ps	339.06 ± 37.72	0.06	0.18
Frag-BOMD 3_{10} helix	6-31+G(d)	3.67 ps	338.20 ± 36.99	0.05	0.09
Frag-ADMP 3_{10} helix	6-31+G(d)	4.32 ps	316.76 ± 33.03	0.03	−0.05
BOMD β-sheet		8.44 ps	328.11 ± 39.02	0.02	0.03
Frag-BOMD β-sheet	6-31+G	6.37 ps	334.29 ± 41.12	0.10	0.29
Frag-BOMD β-sheet	6-31+G(d)	2.03 ps	327.92 ± 45.30	0.03	0.05
Frag-ADMP β-sheet	6-31+G(d)	3.82 ps	295.87 ± 36.43	0.07	−0.04
		Ala_4			
BOMD 3_{10} helix		4.80 ps	337.27 ± 31.75	0.03	0.02
Frag-BOMD 3_{10} helix	6-31+G	6.59 ps	337.00 ± 31.40	0.04	−0.01
Frag-BOMD 3_{10} helix	6-31+G(d)	2.87 ps	337.31 ± 31.31	0.02	0.03
Frag-ADMP 3_{10} helix	6-31+G(d)	4.73 ps	310.26 ± 28.81	0.07	0.04
BOMD β-sheet		6.13 ps	310.16 ± 31.32	0.02	0.01
Frag-BOMD β-sheet	6-31+G	6.15 ps	311.80 ± 31.74	0.08	0.23
Frag-BOMD β-sheet	6-31+G(d)	2.90 ps	311.97 ± 32.99	0.03	−0.01
Frag-ADMP β-sheet	6-31+G(d)	4.91 ps	305.21 ± 34.82	0.07	−0.15
		Ala_{12}			
Frag-ADMP 3_{10} helix	6-31+G(d)	1.25 ps	308.49 ± 24.86	0.06	0.00
Frag-ADMP β-sheet	6-31+G	2.99 ps	300.75 ± 20.53	0.07	−0.10

[a] Smaller basis set used in extrapolation to 6-311++G(2df,2pd). [b] Total simulation time in picoseconds. [c] By use of the equipartition theorem, $\frac{3}{2}(N-1)\mathrm{kT}$, we convert the kinetic energy into average and RMS temperatures. Here temperature is a measure of the available energy to sample the conformational space. The initial kinetic energies were randomly distributed along the nuclear degrees of freedom. These random velocities were chosen such that the initial temperatures were 658 K (62.75 kcal/mol) for Ala_3, 627 K (78.44 kcal/mol) for Ala_4, and 656 K (238.45) for Ala_{12}. [d] RMS deviation of the total energy in kcal/mol. [e] The drift in the Hamiltonian (total energy) is computed as the difference between the average for the first 100 fs and last 100 fs. In kcal/mol.

refinement for the fragment-based extrapolations. The addition of diffuse and polarization functions to hydrogens did not offer significant improvement, except for addition of the hydrogen diffusion functions to 6-31+G basis. Increasing the number of fragments from $\eta = 4$ to $\eta = 5$ offers additional refinement to the extrapolation as it includes the $(i \to i + 4)$ interactions necessary for stability of the 3_{10} helical conformer. Overall the large basis set 6-311++G(2df,2pd) can be effectively captured by fragmentation-based extrapolations by basis sets that contain less than half the number of basis functions.

IV.C. Computational Gain.

One significant limitation in electronic structure theory arises from the intrinsic steep algebraic scaling of its methods with number of basis functions. Density functional theory is the most commonly used class of electronic structure methods and formally scales as $O(N^4)$, which is effectively reduced in larger systems to $O(N^{3.5})$ due to reuse of two-electron integrals.[39] Higher level methods in the post-Hartree−Fock regime scale much more rapidly. Due to this scaling the choice of basis set becomes critical to the computational cost of the calculations, thus limiting the utilization of higher quality electronic structure methods. Since ab initio molecular dynamics requires energy and forces at each time step, increased costs for energy and forces would accumulate to result in cost prohibitive scaling with system size. Other methods have been developed to reduce this time step costs in AIMD, such as dual-basis dynamics[85] and multiple time-step basis set partitioning.[86] Here we aim to approximate larger basis calculations using the fragmentation-based extrapolation scheme discussed above to achieve scaling cost associated with smaller basis sets.

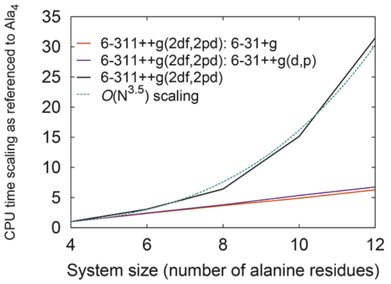
To derive a formal scaling for the extrapolation techniques depicted by eqs 5 and 3, we assume that the size of the larger basis is $N_{B,L}$ per monomer and similarly the size of the smaller basis is $N_{B,S}$ per monomer. Thus, for a system with N monomers (i.e., in this case, the length of Ala_N) the expected computational effort for the larger basis calculations is approximately $O[(N \times N_{B,L})^4]$, where we have accounted for the formal fourth order scaling of DFT with basis-set size and have ignored negligible modifications to CPU times due to presence of link atoms. Hence, the computational effort for the scheme outlined through eq 5 is

$$O[N \times (N-1)/2 \times \{(2 \times N_{B,S})^4 + (2 \times N_{B,L})^4\} + (N \times N_{B,S})^4] \tag{7}$$

when all possible dimer interactions are considered in eq 5 and

$$O[N \times (\eta - 1) \times \{(2 \times N_{B,S})^4 + (2 \times N_{B,L})^4\} + (N \times N_{B,S})^4] \tag{8}$$

when only η edges (dimers) are considered for each node (monomer). Both equations above approach $O[(N \times N_{B,S})^4]$ scaling in the large N limit; that is, the method scales as the small basis calculation in the large basis limit. In Figure 7, we showcase the computational scaling advantage from this fragmentation scheme and compare it to the cost incurred in computing the full system energy and forces with a larger basis set calculation. The costs in Figure 7 are reported as the ratio of the CPU times required to perform calculations for Ala_N with respect to those for Ala_4:

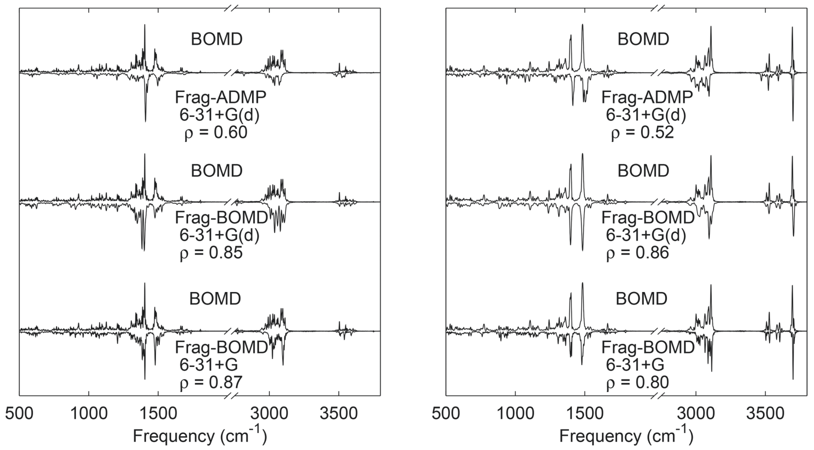


(a)

Figure 7. Computational cost for the fragment based basis-set extrapolation (for $\eta = 5$). The scaling costs are presented here with respect to that for Ala_4. (See eq 9.) The CPU times used are average of those obtained for the β-strand and 3_{10} helical forms.

$$T(Ala_N) = \frac{cputime(Ala_N)}{cputime(Ala_4)} \tag{9}$$

where the quantity, $cputime(Ala_N)$, is the total CPU time required for Ala_N calculations and, similarly, $cputime(Ala_4)$ is the corresponding time for Ala_4. The ratio $T(Ala_N)$ is reported on the left vertical axis of Figure 7 and allows us to study the relative change in system size and corresponding change in CPU time. The proposed scheme offers significant cost savings as the system size grows as seen from Figure 7 and also from



(a)Ala_3 3_{10} helix

(b)Ala_4 3_{10} helix

Figure 8. Vibrational density of states calculated for B3LYP/6-311++G(2df,2pd) dynamics for 3_{10} helical initial conformation for Ala_3 (a) and Ala_4 (b). The full system at 6-311++G(2df,2pd) is presented with a positive amplitude and the fragment dynamics with a negative amplitude for ease of comparison. In the case of the latter, the smaller basis used for extrapolation is noted. The correlation coefficient (ρ from eq 13) is presented to quantify the comparison of the spectra.

eq 8. Furthermore, since the fragment calculations are independent, further computational gain is afforded through the use of MPI-parallelization. [Parallelism is not taken into account in constructing Figure 7 but is used in the AIMD simulations conducted later in this paper.]

The discussion above uses dimer fragments, which is the implementation in eq 5. When higher order many-body interactions are necessary, then eq 3 must be used. In this case the scaling could be more steep; for example, if many-body interactions of the order M are necessary for accuracy, the corresponding scaling then becomes

$$O\left[\left\{\sum_{l=2}^{M}\binom{N}{l} \times \{(l \times N_{B,S})^4 + (l^* N_{B,L})^4\}\right\} + (N \times N_{B,S})^4\right] \tag{10}$$

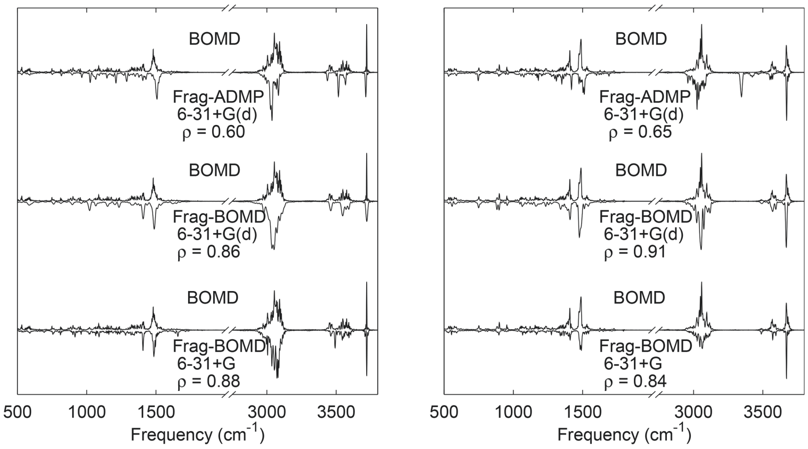
But, as seen in the previous section, dimer fragments with reasonable choice of η are sufficient to maintain good accuracy, and as seen in Figure 7 the scaling is much reduced from use of eq 5.

V. EXTENDED LAGRANGIAN BASED AB INITIO DYNAMICS CONSTRUCTED USING EQUATION 5

As the system size grows, the number of basis functions used for the full system ($N_{B,S}$) increases and determines the overall computational effort. This aspect is already clear from eq 8, where it is shown that the scaling of the extrapolation is constrained only by the full system lower basis calculations. Thus, this full system calculation could potentially become a bottleneck for a larger-sized systems. It is possible to reduce this complexity by introducing an additional layer of basis functions, or by introducing linear scaling methods[106−111] with SCF parallelism[106,112] for the full system smaller basis calculation. But in this paper we also introduce an extended Lagrangian[113,114] implementation where the electronic parameters that depict the energy for the full system smaller basis calculation, $E^{N_{B,S}}$ in eq 5, are treated as dynamical variables. Specifically here, the electronic parameters that determine $E^{N_{B,S}}$ are propagated along with the nuclear degrees of freedom through an adjustment of the relative time scales between the full system, small basis calculation and nuclear degrees of freedom. This is essentially a Car−Parrinello-style method,[21] but is implemented using the atom-centered Gaussian basis functions and single particle density matrices that determine $E^{N_{B,S}}$ and hence follow the atom-centered density matrix propagation (ADMP)[22,115−117] protocol. This methodology, thus, is in similar spirit to the recently developed Atom-centered Density Matrix Propagation with post-Hartree−Fock accuracy (ADMP-pHF).[36,37] Other complementary methods include the dual basis methods[85] and multistep basis set partitioning.[86] The associated multibasis extended Lagrangian is

$$\begin{aligned}\mathcal{L} = &\frac{1}{2}\mathrm{Tr}[V^T \mathrm{MV}] + \frac{1}{2}\mathrm{Tr}[(\boldsymbol{\mu}_{N_{B,S}}^{1/4}\mathbf{W}_{N_{\mathbf{B,S}}}\boldsymbol{\mu}_{N_{B,S}}^{1/4})^2] \\ &- E_{R=2}^{graph\text{-}theoretic}(\mathbf{R}, \mathbf{P}_{\mathbf{N_{B,S}}}) - \mathbf{Tr}[\boldsymbol{\Lambda}_{\mathbf{N_{B,S}}}(\mathbf{P}_{\mathbf{N_{B},S}}{}^2 - \mathbf{P}_{\mathbf{N_{B,S}}})]\end{aligned} \tag{11}$$

Here the parameters $\mathbf{R}$ and $\mathbf{V}$ represent the classical nuclear positions and velocities, with masses, $\mathbf{M}$. The single particle density matrix $\mathbf{P}_{\mathbf{N_{B,S}}}$ represents the full system, at the lower level of basis, and is propagated to determine $E^{N_{B,S}}$, which is part of $E_{R=2}^{graph\text{-}theoretic}$ in eq 11 (see eq 5). This density matrix dynamics is tempered by a fictitious velocity, $\mathbf{W}_{\mathbf{N_{B,S}}}$ (in the spirit of Car−Parrinello,[21] ADMP[115] and ADMP-pHF[36,37]), with fictitious inertia tensor $\boldsymbol{\mu}_{\mathbf{N_{B,S}}}$. Velocity Verlet[118] integration is used to evolve the dynamic parameters of the full system $\{\mathbf{R}, \mathbf{V}; \mathbf{P}_{\mathbf{N_{B,S}}}, \mathbf{W}_{\mathbf{N_{B,S}}}\}$. The choice of the fictitious inertia tensor, $\boldsymbol{\mu}_{\mathbf{N_{B,S}}}$, determines deviations from the Born−Oppenheimer surface. These precise deviations from the Born−Oppenheimer surface have been discussed in Appendix A of ref 36. (Also see refs 115, 117.) There are additional nuclear forces that arise as a result of this propagation, and these forces are proportional to the commutator of the single particle description of the full system, i.e., the associated Fock matrix using the smaller basis and the density matrix $\mathbf{P}_{\mathbf{N_{B,S}}}$. Based on these criteria the values for the fictitious inertia tensor, $\mu_{N_{B,S}}$, are chosen as discussed in refs 36, 37, and 115. As a result, the time-scales for the orbitals within $\mathbf{P}_{\mathbf{N_{B,S}}}$ are adjusted based on the respective diagonal Fock matrix values, so as to provide greater inertia to the core orbitals over the valence orbitals, as outlined in ref 37. This then adjusts the time scales such that there is simultaneous propagation for

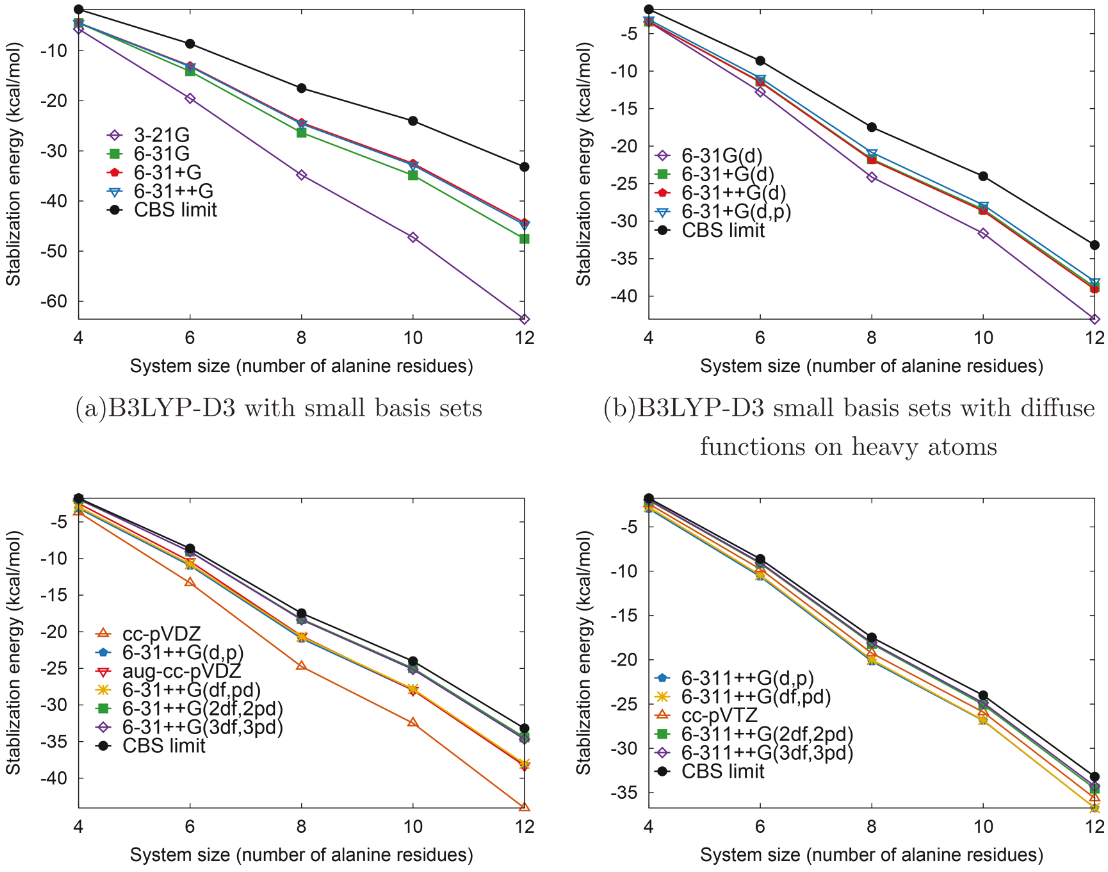


(a)β strand Ala_3

(b)β strand Ala_4

Figure 9. Vibrational density of states calculated from B3LYP/6-311++G(2df,2pd) dynamics for β strand initial conformations of Ala_3 (a) and Ala_4 (b). As in Figure 8, we present the full system with a positive amplitude, and each of the fragment-based dynamics results are presented with a negative amplitude for ease of comparison. The smaller basis used for extrapolation is noted along with the correlation coefficient from eq 13.

both the nuclear, in **R**, and electronic degrees of freedom represented within $\mathbf{P}_{N_{B,S}}$, through eq 11. However, the use of the parameter $\boldsymbol{\mu}_{N_{B,S}}$ in extended Lagrangian schemes[37,116,119−121] couples the electron density to the Born−Oppenheimer surface leading to oscillations which perturb the nuclear motion quadratically.[37] A scaling factor is thus introduced here, as per the prescriptions of ref 36 to obtain the observable vibrational frequencies. As shown in ref 37, this scaling factor is system independent, and we use the same scaling factor in previous studies.[36,38] (These are listed in Section VI).



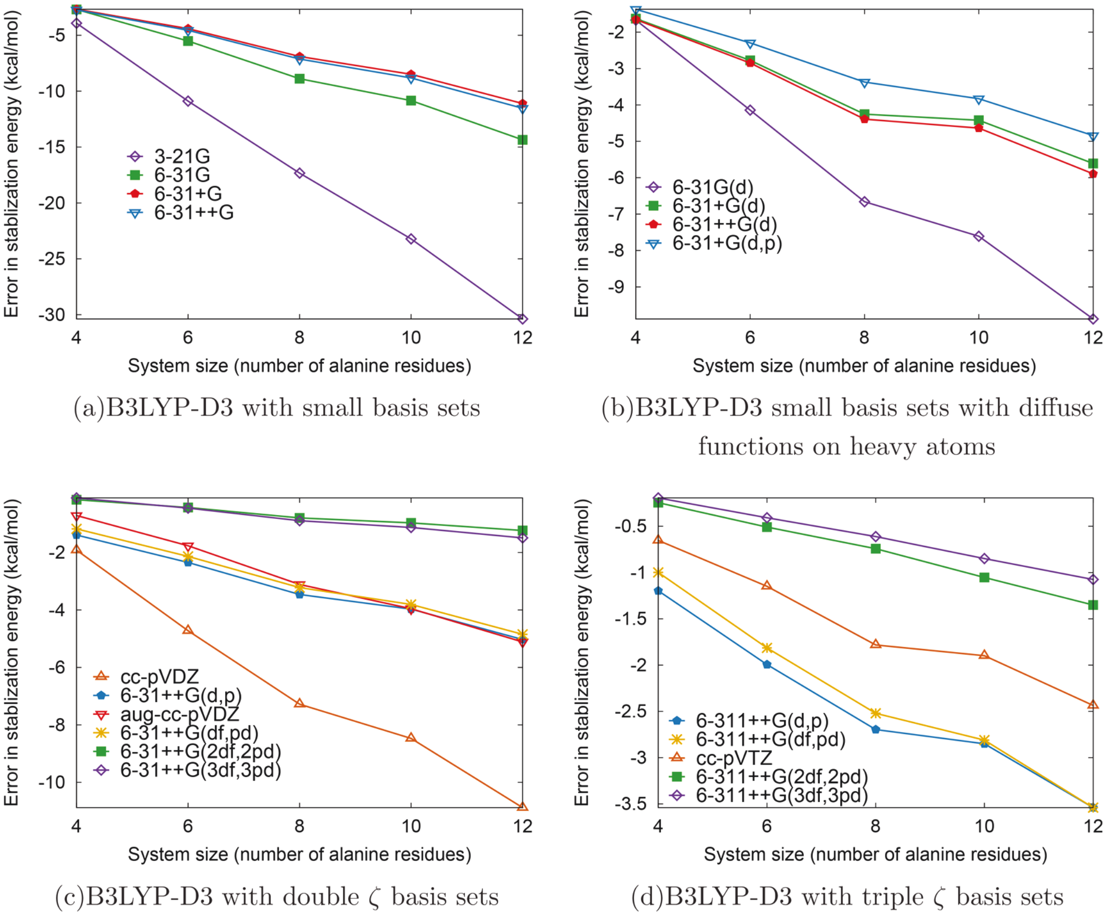
(a)B3LYP-D3 with small basis sets

(b)B3LYP-D3 small basis sets with diffuse functions on heavy atoms

(c)B3LYP-D3 with double ζ basis sets

(d)B3LYP-D3 with triple ζ basis sets

Figure A-1. Conformational stabilization energy dependence on the choice of basis, with dispersion corrected B3LYP,[45] to complement Figures 3 and 4.



(a)B3LYP-D3 with small basis sets

(b)B3LYP-D3 small basis sets with diffuse functions on heavy atoms

(c)B3LYP-D3 with double ζ basis sets

(d)B3LYP-D3 with triple ζ basis sets

Figure A-2. Stabilization energy difference with respect to the CBS limit for the studies shown in Figure A-1, with dispersion corrected B3LYP.[45] The results here are a complement to Figures 3, 4, and A-1. Part (d) here is identical to Figure 4d for the same reason as that highlighted in the caption of Figure 4.

The Lagrange multiplier matrix, $\mathbf{\Lambda}_{\mathbf{N}_{B,S}}$, conserves the N-representability[22,115,122] of $\mathbf{P}_{\mathbf{N}_{B,S}}$. This is done by (a) conserving the idempotency of $\mathbf{P}_{\mathbf{N}_{B,S}}$ through an iterative procedure[36,37,115] and (b) by conserving the particle number. Thus, to assemble the full energy in eqs 11 and 5, we utilize the propagated density matrix, $\mathbf{P}_{\mathbf{N}_{B,S}}$, to obtain $E^{N_{B,S}}$, and the remaining parts of $E_{R=2}^{graph\text{-}theoretic}$ in eq 5 are obtained through full SCF on the fragments with small and large basis sets yielding $\{\mathcal{S}(\alpha, 1)\}$ and $\{\mathcal{S}(\alpha, 2)\}$. The gradients, nuclear and density matrix, were found as described in ref 36. In Section VI, we present both extended Lagrangian and Born–Oppenheimer versions of AIMD.

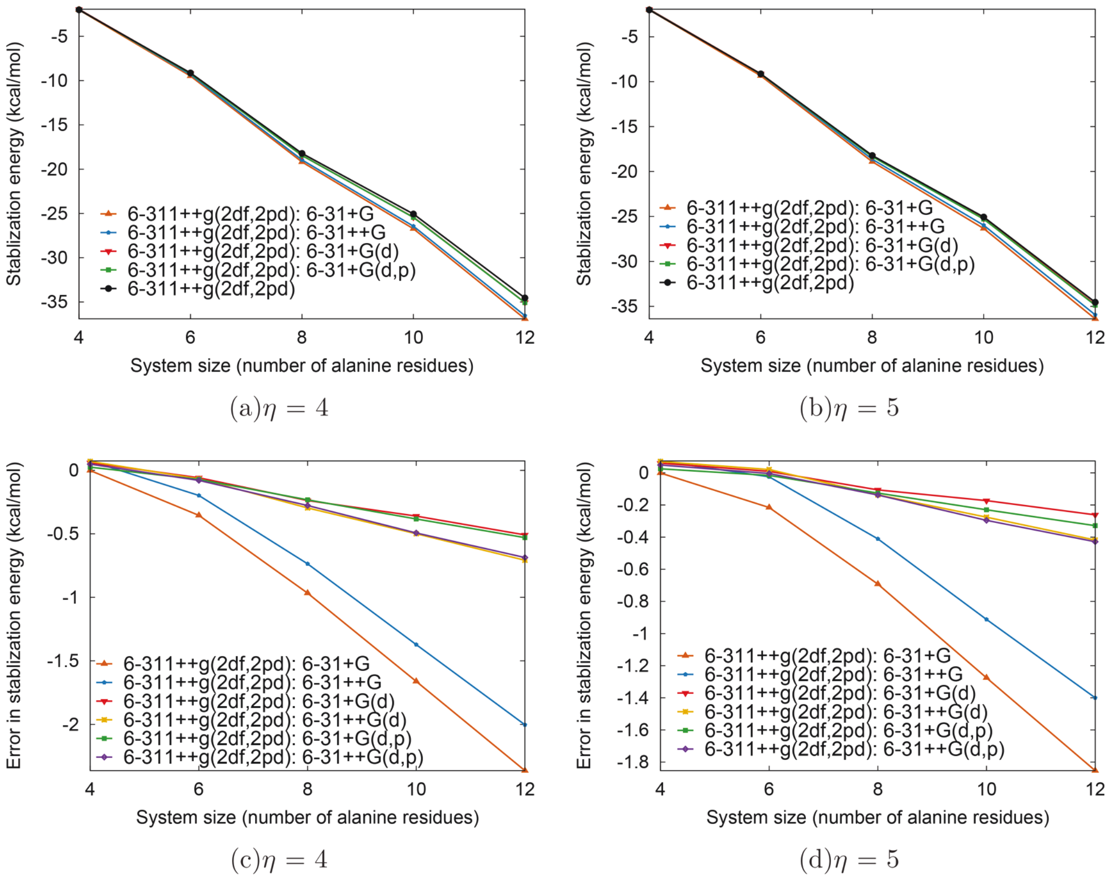
VI. AB INITIO MOLECULAR DYNAMICS TRAJECTORIES INCLUDING BASIS SET EXTRAPOLATION: BORN–OPPENHEIMER AND EXTENDED LAGRANGIAN IMPLEMENTATIONS

In this section, we use our basis set extrapolation scheme to efficiently compute classical trajectories in poly peptide systems. We employ Born–Oppenheimer molecular dynamics where the gradients associated with the system energy in eq 5 are used at every step to propagate the nuclei using the velocity Verlet scheme.[118] The full-system gradients are assembled from those obtained from the nuclear gradients from the full system low level as well as fragment calculations, with appropriate coefficients from eqs 1, 2, 3, and 5. Link atom gradients, are transformed back to the atoms in the full system using the standard Jacobians outlined in refs 30 and 35–37.

When the system size is increased, the basis set for the full system dominates the scaling of the calculations, becoming the bottleneck for dynamics. As was discussed in Section V, this obstacle is alleviated here by the propagation of the electronic density matrix for the full system basis using the extended Lagrangian treatment introduced in eq 11. The fictitious inertia tensor choice provides a bound for the maximum time step for the extended Lagrangian formalism with larger values allowing larger time steps. For the production simulations presented here, we chose the fictitious inertia tensor based on past studies on hydrogen bonded systems,[31,36,116,123–127] where that the valence orbitals have an inertia of 180 au (0.1 amu·bohr^2) and the core orbitals are weighted as per their respective diagonal Fock matrix value as discussed in refs 36, 37, and 115.

Here we consider four benchmark AIMD studies with initial conditions including β-strand and 3_{10}-helical forms of Ala_3 and Ala_4. In these cases we were also able to perform BOMD calculations for the target level of theory with large basis (B3LYP/6-311++G(2df,2pd)), thus allowing detailed comparisons. In addition, as a demonstration of the power of our approach, we also present a trajectory for Ala_{12} where the starting geometry is chosen as a β-strand conformation. The target basis BOMD calculations are cost prohibitive for this case.

The initial structures in all cases are optimized in the gas phase at B3LYP/6-31++G(d,p) level of theory as discussed in



(a)$\eta = 4$

(b)$\eta = 5$

(c)$\eta = 4$

(d)$\eta = 5$

Figure A-3. This figure complements Figure 6 and presents the error in stabilization energy for the target basis set of 6-311++G(2df,2pd) with dispersion corrected B3LYP functional.[45] As in Figure 4b,d, Figure 6c is identical to part (c) here. Similarly Figure 6d and part (d) here are also identical. Furthermore, η = 5 marginally improves over η = 4.

Section IIIA. Both 6-31+G and 6-31+G(d) are used as lower level basis sets for extrapolation toward 6-311++G(2df,2pd) with the B3LYP density functional. The initial nuclear velocities were randomly assigned such that the total kinetic energy was 658 K (62.75 kcal/mol) for Ala_3, 627 K (78.44 kcal/mol) for Ala_4 and 656 K (238.45) for Ala_{12}. Simulation details are shown in Table 2. All trajectories were computed with a velocity Verlet integration scheme[118] with a time step of 0.25 fs. The trajectories were microcanonical in the gas phase, and energy conservation was used as a critical gauge of the integration scheme and smoothness of the energy functional and is quantified by the total energy drift ($\mathcal{H}_{\text{Drift}}$) and standard deviation of total energy ($\Delta\mathcal{H}$). For all trajectories considered here, η was set to 4, with the monomers defined as described in Section VI.A. For Ala_3 and Ala_4 this choice of η provides all-dimer interactions, while for Ala_{12} this choice captures the essential spatial interactions that signify the 3_{10}-helical interaction and also gives a modest number of fragments. All dynamics calculations are done using MPI-parallelism. In Table 3 the dynamics details are provided for fragment-based dynamics and full system dynamics, respectively. All calculations conserve the total energy to within 0.10 kcal/mol and have drifts of the order of 0.10 kcal/mol or less.

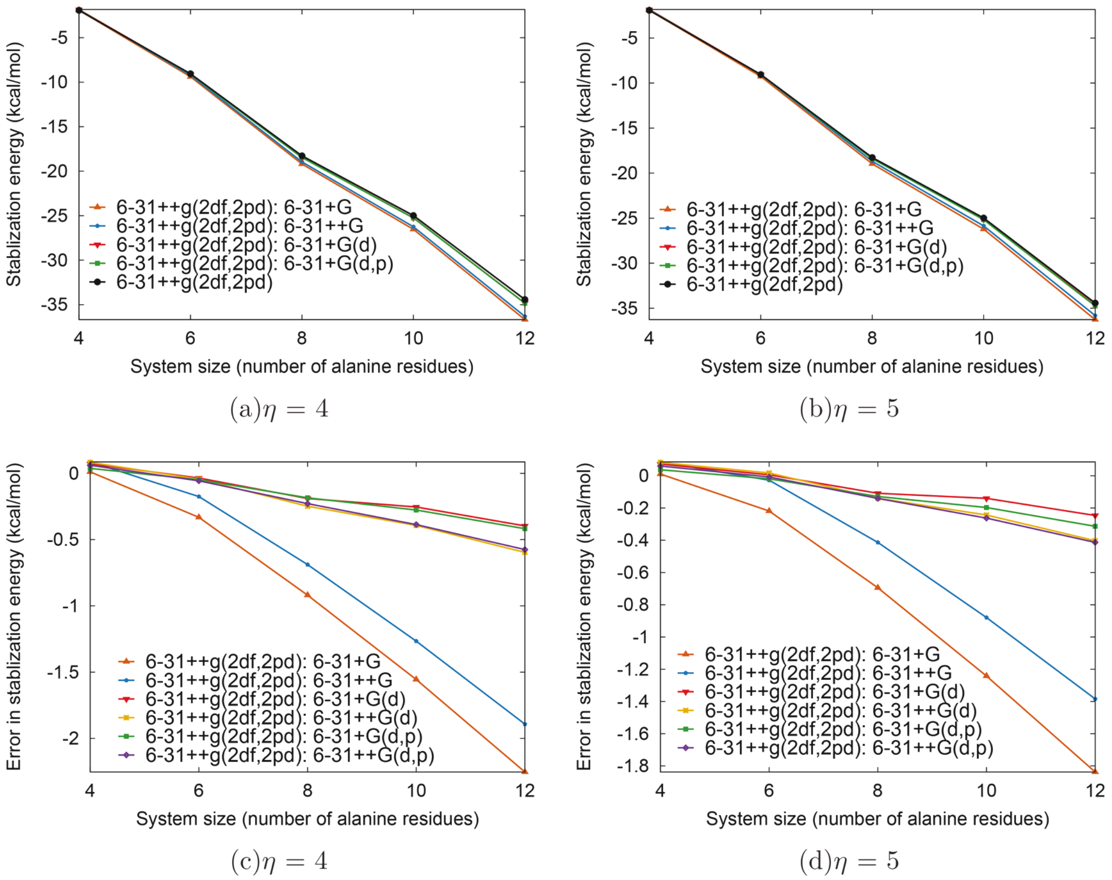
VI.A. Comparison of Vibrational Density of States from AIMD Trajectories. In order to gauge the veracity of the extrapolated dynamics, the vibrational densities of states were computed for both the fragment-based trajectories and the full basis benchmark trajectories. We compute the density of states by use of the Fourier transform of the velocity autocorrelation[123,124,128−130] function. The velocity autocorrelation function, which is simplified by use of the convolution theorem[131] and determines the vibrational density of states ($I_V(\omega)$) from the nuclear velocities, is as follows:

$$\mathbf{I_V}(\omega) = \lim_{T\to\infty}\int_{t=0}^{t=T} dt\ \exp(-\imath\omega t)\langle\mathbf{V}(0)\cdot\mathbf{V}(t)\rangle$$
$$= \tilde{\mathbf{V}}(\omega)\cdot\tilde{\mathbf{V}}(\omega) = |\tilde{\mathbf{V}}(\omega)|^2 \qquad (12)$$

Equation 12 also provides a spectral representation of the trajectory, that is the partitioning of velocities, and hence distribution kinetic energy, as a function of frequency. Here $\mathbf{V}(t)$ is a vector of nuclear velocities at time t, whereas, $\tilde{\mathbf{V}}(\omega)$ are the associated Fourier transforms. In essence, $\mathbf{I}_\mathbf{V}(\omega)$ is roughly related to the amount of energy present in the specific AIMD trajectory at a given frequency and we compare here the spectral densities between the full basis trajectories and the extrapolated trajectories. Figure 8 shows the spectral results for the fragment-based dynamics (Frag-BOMD) and the full system BOMD. In order to quantitatively probe the agreement between each pair of trajectories we compute the Cosine similarity index[132] between the density of states for the fragment based trajectories ($\mathbf{I}_{\mathbf{V,Frag}}$) and those obtained from BOMD simulations that employ the larger (6-311++(2df,2pd)) basis ($\mathbf{I}_\mathbf{V}$):

$$\rho(\mathbf{I}_\mathbf{V}, \mathbf{I}_{\mathbf{V,frag}}) = \frac{\mathbf{I}_\mathbf{V}\cdot\mathbf{I}_{\mathbf{V,Frag}}}{\left|\left|\mathbf{I}_\mathbf{V}\right|\right|\left|\left|\mathbf{I}_{\mathbf{V,Frag}}\right|\right|} \qquad (13)$$

DOI: 10.1021/acs.jctc.8b00803
J. Chem. Theory Comput. 2018, 14, 5535−5552



(a)$\eta = 4$

(b)$\eta = 5$

(c)$\eta = 4$

(d)$\eta = 5$

Figure A-4. These figures complement Figure 6 and present the error in stabilization energy for the target basis set of 6-31++G(2df,2pd) with dispersion corrected B3LYP functional.[45]

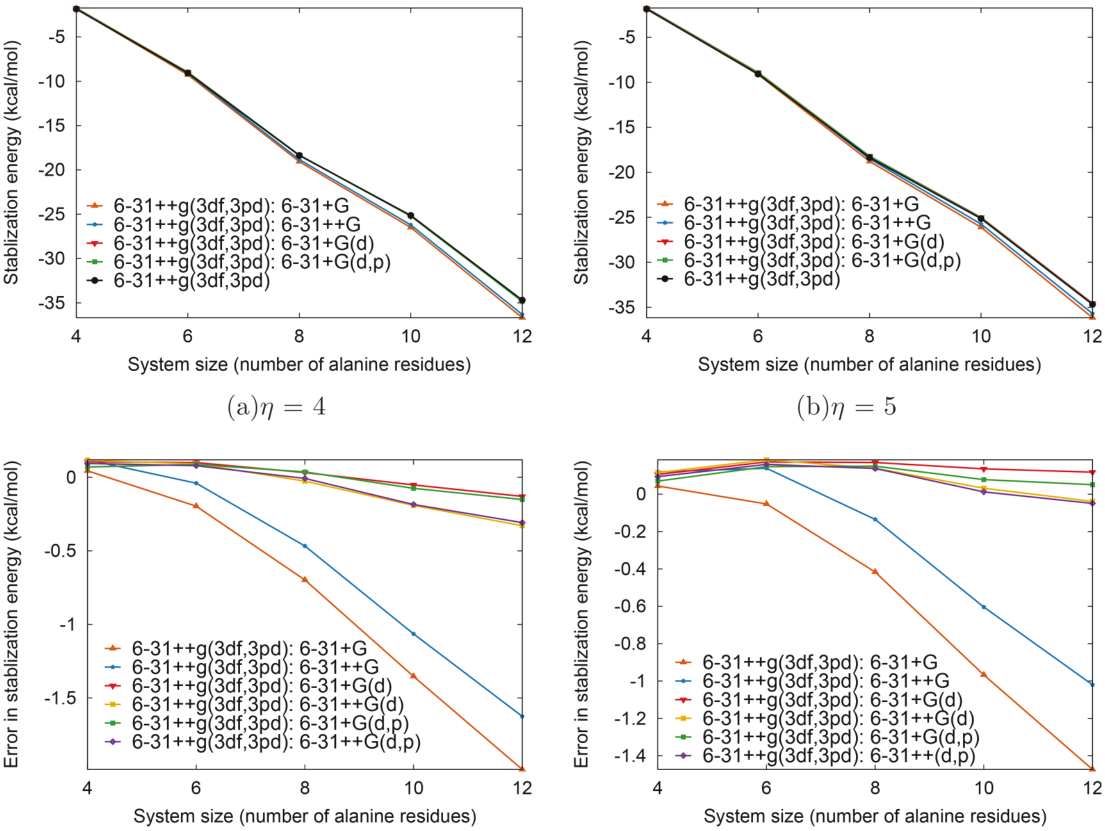
The Cosine similarity index treats the spectral densities as vectors (in frequency) and computes the cosine of the angle between these vectors as indicated in eq 12. This quantity is generally chosen to gauge similarity between positive definite spaces, and since eq 13 is positive definite, we have used the quantity in eq 13 to gauge the similarity here. Furthermore, the quantity in eq 13 is also complementary to Pearson correlation, which has a similar form, but the similarity in that case would be computed by removing the average values of I_V and $I_{V,Frag}$. In all cases, a perfect replication of the spectra would have a coefficient (ρ) of identity.

The spectra in Figures 8 and 9 are the trajectories that begin at the helical and β-strand conformations, respectively. The Amide I and II peaks on the lower frequency end of the spectra are apparent, which are often used in the characterization of polypeptide conformations.[133,134] The methyl and α-carbon–hydrogen stretch regime (about 3100 cm^{-1} to 3200 cm^{-1}) are prominent within these spectra as well. Based on the similarity index in eq 13, the fragment-based trajectories for Ala_3 and Ala_4 quantitatively show close agreement with the benchmark trajectories.

VII. CONCLUSION

In this paper, we have discussed a new approach to perform large basis ab initio molecular dynamics calculations at much reduced computational overhead. The method is based on molecular fragmentation through the adaptation of ONIOM using the set-theoretic inclusion–exclusion principle. Here, a large system is fragmented into monomers or nodes in a graph, and these individual units are allowed to interact up to arbitrary orders based on a truncated many-body expansion. At each level of many-body truncation, the energy and gradients are computed at both the target high-level basis and a specific lower level basis. In addition, the full system energy and gradients are computed using the lower level basis, and these together provide a very good estimate for the high level basis calculations when the individual components are assembled in a fashion consistent with ONIOM. In this sense, the method is related to several previous methods. On the one hand the approach is trivially related to many fragmentation methods. On the other hand, the approach is also related to the well-known double many body expansion popularized by Varandas and co-workers[82] but is now constructed through ONIOM on-the-fly. In all cases, the computational implementation greatly benefits from a geometric network interpretation which reduces computational cost.

We show that basis set choice may gravely affect the determination of conformational stability in peptide secondary structures. These effects grow drastically as the system size grows. To rectify this problem without adversely affecting the computational scaling, we utilize the approach discussed in the previous paragraph for the extrapolation of energy and gradients at a complete basis set using modest levels of computation. Thus, we represent the full system with a smaller basis but include localized many-body interactions with a larger basis representation through an ONIOM-like method with fragment contributions determined by the inclusion–exclusion principle.[135] The method is shown to approach the complete basis accuracy at significantly lower computational cost. In these calculations we note that the inclusion



(a)$\eta = 4$ (b)$\eta = 5$

(c)$\eta = 4$ (d)$\eta = 5$

Figure A-5. These figures complement Figure 6 and present the error in stabilization energy for the target basis set of 6-31++G(3df,3pd) with dispersion corrected B3LYP functional.[45]

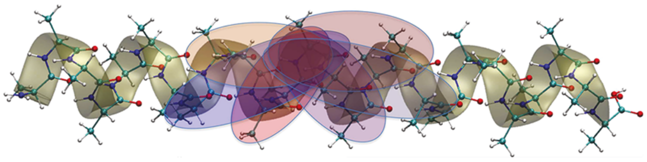
of diffuse functions, in the chosen smaller basis set, are necessary to capture the quintessential character of the electronic density needed to determine stabilization energy in the fragmentation scheme. These extrapolations from a smaller basis set can achieve subkcal/mol accuracy. By varying the lower basis sets used in the extrapolations we note that polarization functions on the heavy atoms are needed to correctly capture the target energy. But the addition of polarization functions on the hydrogen atoms does not lead to a significant gain. This fragmentation method offers a significant cost scaling advantage over the full system calculation. But as system size grows, the expenses are dominated by the lower level full system. To address this, we also introduce an extended Lagrangian scheme, where the full system, lower basis density matrix is propagated with the nuclear degrees of freedom using an extended-Lagrangian formalism. The approach has the potential for accurate AIMD in large systems.

■ APPENDIX A: MORE DETAILS ON BASIS SET DEPENDENCE AND EXTRAPOLATION FOR DISPERSION CORRECTED DENSITY FUNCTIONALS

The conformational stabilization B3LYP energies including the Grimme[45] dispersion correction in Figure A-1 are analogous to Figure 3 but includes dispersion corrections. The extrapolations presented in Figure 6 in the main text are supplemented here through Figures A-3, A-4, and A-5 with dispersion corrections for all calculations. Clearly the basis set effects are similar for the dispersion corrected and uncorrected calculations.

■ APPENDIX B: AN ILLUSTRATION OF THE ISOMORPHISM BETWEEN THE *SIMPLEX DECOMPOSITION* AND *SET-THEORETIC INCLUSION–EXCLUSION PRINCIPLE*

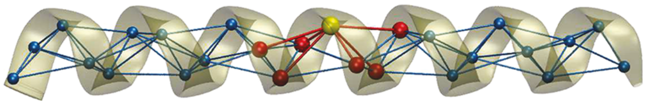
The *simplex decomposition* in ref 38 is isomorphic to the *set-theoretic decomposition.*[35−37] Consider a set-theoretic decomposition of polyalanine, a portion of which is shown below.



Using the inclusion exclusion principle generalization of ONIOM,[35,36,38] the energy expression is

$$E^{\text{PIE}-\text{ONIOM}} \equiv E_{\text{low}}(0) + \sum_{i=1}^{n} \mathcal{S}(i) - \sum_{1\le i<j\le n} \mathcal{S}(i \cap j) + \sum_{1\le i<j<k\le n} \mathcal{S}(i \cap j \cap k) - \cdots + (-1)^{n-1} \sum \mathcal{S}(1 \cap \cdots \cap n)$$

The following geometric network, or simplex decomposition, captures the same interactions in a more efficient way:[38]



Considering dimer interactions (edges) leads to

$$\begin{aligned} E_{\text{network}}^{\text{edge/dimers}} &= E^{\text{level},0} + \sum_{\alpha}^{\text{Edges}} \Delta E_{\alpha}^{\text{cor.},1} \\ &\quad - \sum_{I}^{\text{Nodes}} \Delta E_{I}^{\text{cor.},1}[p_{I}^{1,2} - 1]\Delta E_{\alpha}^{\text{cor.},1} \\ &= E_{\alpha}^{\text{level},\mathbf{1}} - E_{\alpha}^{\text{level},0} \end{aligned}$$

The expression above is similar to eq 5 truncated at the level of edges. A more general expression that includes *all embedded simplexes* yields a generalized (geometric) description of **many-body interactions**

$$\begin{aligned} E_{\text{graph-theoretic}}^{R\text{-rank}} &= E^{\text{level},0} + \sum_{r=1}^{R} (-1)^{r} \\ &\quad \times \left\{ \sum_{\alpha} \Delta E_{\alpha,r}^{\text{cor.},\mathbf{1}} \left[\sum_{m=r}^{R} (-1)^{m} p_{\alpha}^{r,m} \right] \right\} \end{aligned}$$

and this expression is similar to eq 3. The square bracketed term contains the overcounting correction from ref 38, where $p_{\alpha}^{r,m}$ is the number of times the αth rank-r simplex appears in simplexes of higher rank. This last expression has close connections to the well-known many body expansions.[51−57]

AUTHOR INFORMATION

Corresponding Author

*(S.S.I.) E-mail: iyengar@indiana.edu.

ORCID

Srinivasan S. Iyengar: 0000-0001-6526-2907

Funding

This research is supported by the National Science Foundation Grant NSF CHE-1665336 to S.S.I.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Science Publications: New York, 1987.

(2) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **2000**, *33*, 889.

(3) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589.

(4) Petersen, M. K.; Iyengar, S. S.; Day, T. J. F.; Voth, G. A. The Hydrated Proton at Water Liquid/Vapour Interfaces. *J. Phys. Chem. B* **2004**, *108*, 14804.

(5) Jungwirth, P.; Tobias, D. J. Molecular structure of salt solutions: a new view of the interface with implications for heterogeneous atmospheric chemistry. *J. Phys. Chem. B* **2001**, *105*, 10468.

(6) Tse, Y.-L. S.; Herring, A. M.; Voth, G. A.; Kim, K. Molecular Dynamics Simulations of Proton Transport in 3M and Nafion Perfluorosulfonic Acid Membranes. *J. Phys. Chem. C* **2013**, *117*, 8079.

(7) Sørensen, M. R.; Mishin, Y.; Voter, A. F. Diffusion mechanisms in Cu grain boundaries. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *62*, 3658.

(8) Han, Y.; Elliott, J. Molecular dynamics simulations of the elastic properties of polymer/carbon nanotube composites. *Comput. Mater. Sci.* **2007**, *39*, 315.

(9) Berthier, L.; Biroli, G. Theoretical perspective on the glass transition and amorphous materials. *Rev. Mod. Phys.* **2011**, *83*, 587.

(10) Warshel, A.; Weiss, R. M. An Empirical Valence Bond Approach for Comparing Reactions in Solutions and in Enzymes. *J. Am. Chem. Soc.* **1980**, *102*, 6218.

(11) Åqvist, J.; Warshel, A. Simulation of Enzyme Reactions Using Valence Bond Force Fields and Other Hybrid Quantum/classical Approaches. *Chem. Rev.* **1993**, *93*, 2523.

(12) Warshel, A. Computer Simulations of Enzyme Catalysis: Methods, Progress, and Insights. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425.

(13) Schmitt, U. W.; Voth, G. A. The Computer Simulation of Proton Transport in Water. *J. Chem. Phys.* **1999**, *111*, 9361.

(14) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396.

(15) Marx, D.; Hutter, J. Ab initio molecular dynamics: theory and implementation. In *Modern methods and algorithms of quantum chemistry*; NIC Series; 2000.

(16) Remler, D. K.; Madden, P. A. Molecular Dynamics Without Effective Potentials Via the Car-Parrinello Approach. *Mol. Phys.* **1990**, *70*, 921.

(17) Wang, I. S. Y.; Karplus, M. Dynamics of Organic Reactions. *J. Am. Chem. Soc.* **1973**, *95*, 8160.

(18) Leforestier, C. Classical Trajectories Using the Full Ab Initio Potential Energy Surface $H^{-}+CH_4 \rightarrow CH_4+H^{-}$. *J. Chem. Phys.* **1978**, *68*, 4406.

(19) Helgaker, T.; Uggerud, E.; Jensen, H. J. A. Integration of the Classical Equations of Motion on Ab Initio Molecular Potential Energy Surfaces Using Gradients and Hessians: Application to Translational Energy Release upon Fragmentation. *Chem. Phys. Lett.* **1990**, *173*, 145.

(20) Bolton, K.; Hase, W. L.; Peslherbe, G. H. In *Modern Methods for Multidimensional Dynamics Computation in Chemistry*; Thompson, D. L., Ed.; World Scientific: Singapore, 1998; Chapter Direct Dynamics of Reactive Systems, p 143.

(21) Car, R.; Parrinello, M. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Phys. Rev. Lett.* **1985**, *55*, 2471.

(22) Schlegel, H. B.; Millam, J. M.; Iyengar, S. S.; Voth, G. A.; Daniels, A. D.; Scuseria, G. E.; Frisch, M. J. Ab Initio Molecular Dynamics: Propagating the Density Matrix with Gaussian Orbitals. *J. Chem. Phys.* **2001**, *114*, 9758.

(23) Peverati, R.; Truhlar, D. Quest for a Universal Density Functional: The Accuracy of Density Functionals Across a Broad Spectrum of Databases in Chemistry and Physics. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120476.

(24) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315.

(25) Tse, J. S. Ab initio molecular dynamics with density functional theory. *Annu. Rev. Phys. Chem.* **2002**, *53*, 249.

(26) Wong, K.-Y.; Gao, J. Insight into Phosphodiesterase Mechanism from Combined QM/MM Molecular Dynamics Simulations. *FEBS J.* **2011**, *278*, 2579.

(27) Harris, D. L. Oxidation and Electronic State Dependence of Proton Transfer in the Enzymatic Cycle of Cytochrome P450eryF. *J. Inorg. Biochem.* **2002**, *91*, 568.

(28) van der Kamp, M. W.; Mulholland, A. J. Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **2013**, *52*, 2708.

(29) Gao, J.; Truhlar, D. G. Quantum mechanical methods for enzyme kinetics. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467.

(30) Rega, N.; Iyengar, S. S.; Voth, G. A.; Schlegel, H. B.; Vreven, T.; Frisch, M. J. Hybrid Ab-Initio/Empirical Molecular Dynamics:

Combining the ONIOM Scheme with the Atom-Centered Density Matrix Propagation (ADMP) Approach. *J. Phys. Chem. B* **2004**, *108*, 4210.

(31) Phatak, P.; Sumner, I.; Iyengar, S. S. Gauging the Flexibility of the Active Site in Soybean Lipoxygenase-1 (SLO-1) Through an Atom-Centered Density Matrix Propagation (ADMP) Treatment That Facilitates the Sampling of Rare Events. *J. Phys. Chem. B* **2012**, *116*, 10145.

(32) Klimes, J.; Michaelides, A. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *J. Chem. Phys.* **2012**, *137*, 120901.

(33) Mori-Sanchez, P.; Cohen, A. J.; Yang, W. T. Many-electron self-interaction error in approximate density functionals. *J. Chem. Phys.* **2006**, *125*, 201102.

(34) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. T. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289.

(35) Li, J.; Iyengar, S. S. Ab initio Molecular Dynamics using Recursive, Spatially Separated, Overlapping Model Subsystems Mixed Within an ONIOM Based Fragmentation Energy Extrapolation Technique. *J. Chem. Theory Comput.* **2015**, *11*, 3978.

(36) Li, J.; Haycraft, C.; Iyengar, S. S. Hybrid extended Lagrangian, post-Hartree-Fock Born-Oppenheimer ab initio molecular dynamics using fragment-based electronic structure. *J. Chem. Theory Comput.* **2016**, *12*, 2493.

(37) Haycraft, C.; Li, J.; Iyengar, S. S. On-the-fly" Ab initio molecular dynamics with coupled cluster accuracy. *J. Chem. Theory Comput.* **2017**, *13*, 1887.

(38) Ricard, T. C.; Haycraft, C.; Iyengar, S. S. Adaptive, geometric networks for efficient coarse-grained *ab initio* molecular dynamics with post-Hartree-Fock accuracy. *J. Chem. Theory Comput.* **2018**, *14*, 2852.

(39) Schlegel, H. B.; Frisch, M. J. Computational Bottlenecks In Molecular-Orbital Calculations. In *Theoretical and Computational Models for Organic Chemistry*; Formosinho, S. J., Csizmadia, I. G., Arnaut, L. G., Eds.; Nato Advanced Science Institutes Series, Series C, Mathematical And Physical Sciences; 1991; Vol. *339*, pp 5−33.

(40) Maseras, F.; Morokuma, K. A New "Ab Initio + Molecular Mechanics" Geometry Optimization Scheme of Equilibrium Structures and Transition States. *J. Comput. Chem.* **1995**, *16*, 1170.

(41) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J. Basis-set convergence of the energy in molecular Hartree-Fock calculations. *Chem. Phys. Lett.* **1999**, *302*, 437.

(42) Dunning, T. H., Jr Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. the Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.

(43) Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martinez, T. J. Ab Initio Quantum Chemistry for Protein Structures. *J. Phys. Chem. B* **2012**, *116*, 12501.

(44) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martinez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381.

(45) 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*, 154104.

(46) Binkley, J. S.; Pople, J. A.; Hehre, W. J. Self-Consistent Molecular Orbital Methods. 21. Small Split-Valence Basis Sets for First-Row Elements. *J. Am. Chem. Soc.* **1980**, *102*, 939.

(47) van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. State of Art in Counterpoise Theory. *Chem. Rev.* **1994**, *94*, 1873.

(48) Clementi, E. Study of the Electronic Structure of Molecules. II. Wavefunctions for the NH_3+HCl → NH_4Cl Reaction. *J. Chem. Phys.* **1967**, *46*, 3851.

(49) Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19*, 553.

(50) Iyengar, S. S.; Frisch, M. J. Effect of Time-Dependent Basis Functions and Their Superposition Error on Atom-Centered Density Matrix Propagation (ADMP): Connections to Wavelet Theory of Multi-Resolution Analysis. *J. Chem. Phys.* **2004**, *121*, 5061.

(51) Murrell, J.; Carter, S.; Farantos, S.; Huxley, P.; Varandas, A. *Molecular Potential Energy Functions*; Wiley: New York, 1984.

(52) Varandas, A.; Pais, A. A realistic double many-body expansion (DMBE) potential energy surface for ground-state O_3 from a multiproperty fit to ab initio calculations, and to experimental spectroscopic, inelastic scattering, and kinetic isotope thermal rate data. *Mol. Phys.* **1988**, *65*, 843.

(53) Burnham, C. J.; Xantheas, S. S. Development of Transferable Interaction Models for Water. IV. a Flexible, All-Atom Polarizable Potential (TTM2-F) Based on Geometry Dependent Charges Derived from an Ab Initio Monomer Dipole Moment Surface. *J. Chem. Phys.* **2002**, *116*, 5115.

(54) Hirata, S. Fast Electron-Correlation Methods for Molecular Crystals: an Application to the α, $\beta(1)$, and $\beta(2)$ Modifications of Solid Formic Acid. *J. Chem. Phys.* **2008**, *129*, 204104.

(55) Kamiya, M.; Hirata, S.; Valiev, M. Fast Electron-Correlation Methods for Molecular Crystals Without Basis Set Superposition Errors. *J. Chem. Phys.* **2008**, *128*, 074103.

(56) Jacobson, L. D.; Herbert, J. M. An Efficient, Fragment-Based Electronic Structure Method for Molecular Systems: Self-Consistent Polarization with Perturbative Two-Body Exchange and Dispersion. *J. Chem. Phys.* **2011**, *134*, 094118.

(57) Góra, U.; Podeszwa, R.; Cencek, W.; Szalewicz, K. Interaction energies of large clusters from many-body expansion. *J. Chem. Phys.* **2011**, *135*, 224102.

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

(59) 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*, 1563.

(60) Hopkins, B. W.; Tschumper, G. S. Multicentred QM/QM Methods for Overlapping Model Systems. *Mol. Phys.* **2005**, *103*, 309.

(61) 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*, 104109.

(62) Sahu, N.; Yeole, S. D.; Gadre, S. R. Appraisal of molecular tailoring approach for large clusters. *J. Chem. Phys.* **2013**, *138*, 104101.

(63) Guo, W.; Wu, A.; Xu, X. XO: An Extended ONIOM Method for Accurate and Efficient Geometry Optimization of Large Molecules. *Chem. Phys. Lett.* **2010**, *498*, 203.

(64) Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115*, 5643.

(65) 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*, 1336.

(66) 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*, 2012.

(67) Mayhall, N. J.; Raghavachari, K. Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Nondisjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2669.

(68) 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*, 701.

(69) Deev, V.; Collins, M. A. Approximate ab initio energies by systematic molecular fragmentation. *J. Chem. Phys.* **2005**, *122*, 154102.

(70) Gordon, M.; Mullin, J.; Pruitt, S.; Roskop, L.; Slipchenko, L.; Boatz, J. Accurate Methods for Large Molecular Systems. *J. Phys. Chem. B* **2009**, *113*, 9646.

(71) 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*, 7578.

(72) Li, S.; Li, W.; Ma, J. Generalized Energy-Based Fragmentation Approach and Its Applications to Macromolecules and Molecular Aggregates. *Acc. Chem. Res.* **2014**, *47*, 2712.

(73) Collins, M. A.; Bettens, R. P. A. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015**, *115*, 5607.

(74) Thapa, B.; Beckett, D.; Jovan Jose, K.; Raghavachari, K. Assessment of Fragmentation Strategies for Large Proteins Using the Multilayer Molecules-in-Molecules Approach. *J. Chem. Theory Comput.* **2018**, *14*, 1383.

(75) Brorsen, K. R.; Minezawa, N.; Xu, F.; Windus, T. L.; Gordon, M. S. Fragment Molecular Orbital Molecular Dynamics with the Fully Analytic Energy Gradient. *J. Chem. Theory Comput.* **2012**, *8*, 5008.

(76) Li, J.; Sode, O.; Hirata, S. Second-Order Many-Body Perturbation Study on Thermal Expansion of Solid Carbon Dioxide. *J. Chem. Theory Comput.* **2015**, *11*, 224.

(77) Willow, S. Y.; Salim, M. A.; Kim, K. S.; Hirata, S. Ab initio molecular dynamics of liquid water using embedded fragment second-order many-body perturbation theory towards its accurate property prediction. *Sci. Rep.* **2015**, *5*, 14358.

(78) Liu, J.; Zhu, T.; Wang, X.; He, X.; Zhang, J. Z. H. Quantum Fragment Based ab Initio Molecular Dynamics for Proteins. *J. Chem. Theory Comput.* **2015**, *11*, 5897.

(79) Pruitt, S. R.; Nakata, H.; Nagata, T.; Mayes, M.; Alexeev, Y.; Fletcher, G.; Fedorov, D. G.; Kitaura, K.; Gordon, M. S. Importance of Three-Body interactions in molecular dynamics simulations of water demonstrated with the fragment molecular orbital method. *J. Chem. Theory Comput.* **2016**, *12*, 1423.

(80) Collins, M. A. Can Systematic Molecular Fragmentation Be Applied to Direct Ab Initio Molecular Dynamics? *J. Phys. Chem. A* **2016**, *120*, 9281.

(81) Liu, J.; He, X.; Zhang, J. Z.; Qi, L.-W. Hydrogen-bond structure dynamics in bulk water: insights from ab initio simulations with coupled cluster theory. *Chem. Sci.* **2018**, *9*, 2065.

(82) Varandas, A. J.; Murrell, J. N. A many-body expansion of polyatomic potential energy surfaces: application to H n systems. *Faraday Discuss. Chem. Soc.* **1977**, *62*, 92.

(83) 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*, 46.

(84) 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*, 064113.

(85) Steele, R. P.; Head-Gordon, M.; Tully, J. C. Ab Initio Molecular Dynamics with Dual Basis Set Methods. *J. Phys. Chem. A* **2010**, *114*, 11853.

(86) Steele, R. P. Multiple-timestep ab initio molecular dynamics using an atomic basis set partitioning. *J. Phys. Chem. A* **2015**, *119*, 12119.

(87) Bowyer, A. Computing Dirichlet tessellations. *Comput. J.* **1981**, *24*, 162.

(88) Watson, D. Computing the n-dimensional Delaunay tessellation with applications to Voronoi polytopes. *Comput. J.* **1981**, *24*, 167.

(89) Aurenhammer, F. Voronoi Diagrams — A survey of a fundamental geometric data structure. *ACM Comput. Survey* **1991**, 23, 345.

(90) Okabe, A.; Boots, B.; Sugihara, K.; Chiu, S. N. *Spatial Tessellations — Concepts and applications of Voronoi diagrams*; John Wiley and Sons: 2000.

(91) Hert, S.; Seel, M. dD Convex Hulls and Delaunay Triangulations. In *CGAL User and Reference Manual*, 4.10 ed.; CGAL Editorial Board: 2017.

(92) Farin, G. Surfaces over Dirichlet Tessellations. *Computer Aided Geometric Design* **1990**, *7*, 281.

(93) DeGregorio, N.; Iyengar, S. S. Efficient and adaptive methods for computing accurate potential surfaces for quantum nuclear effects: Applications to hydrogen transfer reactions. *J. Chem. Theory Comput.* **2018**, *14*, 30.

(94) Coffey, T. M.; Wyatt, R. E.; Schieve, W. C. Reconstruction of the Time-Dependent Wave Function Exclusively from Position Data. *Phys. Rev. Lett.* **2011**, *107*, 230403.

(95) Sun, L.; Yeh, G.-T.; Ma, X.; Lin, F.; Zhao, G. Engineering applications of 2D and 3D finite element mesh generation in hydrogeology and water resources. *Comput. Geosci* **2017**, *21*, 733.

(96) Dey, T. K.; Shah, N. R. On the number of simplicial complexes in R^d. *Computational Geometry* **1997**, *8*, 267.

(97) Nelson, D. L.; Cox, M. M. In *Lehninger Principles of Biochemistry*, 4th ed.; Freeman: 2004.

(98) Creighton, T. Protein structure. Stability of alpha-helices. *Nature* **1987**, *326*, 547−548.

(99) Jagielska, A.; Skolnick, J. Origin of intrinsic 3_{10}-helix versus strand stability in homopolypeptides and its implications for the accuracy of the Amber force field. *J. Comput. Chem.* **2007**, *28*, 1648.

(100) Wieczorek, R.; Dannenberg, J. J. Comparison of Fully Optimized α- and 3_{10}−Helices with Extended ß-Strands. An ONIOM Density Functional Theory Study. *J. Am. Chem. Soc.* **2004**, *126*, 14198.

(101) Li, J.; Wang, Y.; Chen, J.; Liu, Z.; Bax, A.; Yao, L. Observation of α−Helical Hydrogen-Bond Cooperativity in an Intact Protein. *J. Am. Chem. Soc.* **2016**, *138*, 1824.

(102) Simon, S.; Duran, M.; Dannenberg, J. J. Effect of Basis Set Superposition Error on the Water Dimer Surface Calculated at Hartree Fock, Møller-Plesset, and Density Functional Theory Levels. *J. Phys. Chem. A* **1999**, *103*, 1640.

(103) Chalasinski, G.; Gutowski, M. Weak Interactions between Small Systems. Models for Studying the Nature of Intermolecular Forces and Challenging Problems for ab Initio Calculations. *Chem. Rev.* **1988**, *88*, 943.

(104) Bakowies, D.; Thiel, W. Hybrid Models for Combined Quantum Mechanical and Molecular Mechanical Approaches. *J. Phys. Chem.* **1996**, *100*, 10580.

(105) Dapprich, S.; Komáromi, I.; Byun, K.; Morokuma, K.; Frisch, M. J. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct.: THEOCHEM* **1999**, *461*, 1.

(106) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree-Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree-Fock exchange. *Chem. Phys.* **2009**, *356*, 98.

(107) Kussmann, J.; Beer, M.; Ochsenfeld, C. Linear-scaling self-consistent field methods for large molecules. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2013**, *3*, 614.

(108) Scuseria, G. E. Linear Scaling Density Functional Calculations with Gaussian Orbitals. *J. Phys. Chem. A* **1999**, *103*, 4782.

(109) Goedecker, S. Linear Scaling Electronic Structure Methods. *Rev. Mod. Phys.* **1999**, *71*, 1085.

(110) White, C. A.; Head-Gordon, M. Derivation and Efficient Implementation of the Fast Multipole Method. *J. Chem. Phys.* **1994**, *101*, 6593.

(111) Strain, M. C.; Scuseria, G. E.; Frisch, M. J. *Science* **1996**, *271*, 51.

(112) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477.

(113) Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure And/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384.

(114) Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196.

(115) Iyengar, S. S.; Schlegel, H. B.; Millam, J. M.; Voth, G. A.; Scuseria, G. E.; Frisch, M. J. Ab Initio Molecular Dynamics: Propagating the Density Matrix with Gaussian Orbitals. II. Generalizations Based on Mass-Weighting, Idempotency, Energy Conserva-

tion and Choice of Initial Conditions. *J. Chem. Phys.* **2001**, *115*, 10291.

(116) Schlegel, H. B.; Iyengar, S. S.; Li, X.; Millam, J. M.; Voth, G. A.; Scuseria, G. E.; Frisch, M. J. Ab Initio Molecular Dynamics: Propagating the Density Matrix with Gaussian Orbitals. III. Comparison with Born-Oppenheimer Dynamics. *J. Chem. Phys.* **2002**, *117*, 8694.

(117) Iyengar, S. S.; Schlegel, H. B.; Voth, G. A.; Millam, J. M.; Scuseria, G. E.; Frisch, M. J. Ab Initio Molecular Dynamics: Propagating the Density Matrix with Gaussian Orbitals. IV. Formal Analysis of the Deviations from Born-Oppenheimer Dynamics. *Isr. J. Chem.* **2002**, *42*, 191.

(118) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer-Simulation Method for the Calculation of Equilibrium-Constants for the Formation of Physical Clusters of Molecules - Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76*, 637.

(119) Hutter, J. Car-Parrinello Molecular Dynamics. *WIREs-Comp. Mol. Sci.* **2012**, *2*, 604.

(120) Tangney, P.; Scandolo, S. How Well do Car-Parrinello Simulations Reproduce the Born-Oppenheimer Surface? Theory and Examples. *J. Chem. Phys.* **2002**, *116*, 14.

(121) Herbert, J. M.; Head-Gordon, M. Curvy-Steps Approach to Constraint-Free Extended-Lagrangian Ab Initio Molecular Dynamics, Using Atom-Centered Basis Functions: Convergence Toward Born Oppenheimer Trajectories. *J. Chem. Phys.* **2004**, *121*, 11542.

(122) McWeeny, R. Some Recent Advances in Density Matrix Theory. *Rev. Mod. Phys.* **1960**, *32*, 335.

(123) Iyengar, S. S.; Petersen, M. K.; Day, T. J. F.; Burnham, C. J.; Teige, V. E.; Voth, G. A. The Properties of Ion-Water Clusters. I. the Protonated 21-Water Cluster. *J. Chem. Phys.* **2005**, *123*, 084309.

(124) Iyengar, S. S. Further Analysis of the Dynamically Averaged Vibrational Spectrum for the "Magic" Protonated 21-Water Cluster. *J. Chem. Phys.* **2007**, *126*, 216101.

(125) Iyengar, S. S.; Day, T. J. F.; Voth, G. A. On the Amphiphilic Behavior of the Hydrated Proton: An *Ab Initio* Molecular Dynamics Study. *Int. J. Mass Spectrom.* **2005**, *241*, 197.

(126) Iyengar, S. S. Dynamical Effects on Vibrational and Electronic Spectra of Hydroperoxyl Radical Water Clusters. *J. Chem. Phys.* **2005**, *123*, 084310.

(127) Vimal, D.; Pacheco, A. B.; Iyengar, S. S.; Stevens, P. S. Experimental and Ab Initio Dynamical Investigations of the Kinetics and Intramolecular Energy Transfer Mechanisms for the OH + 1,3-Butadiene Reaction Between 263 and 423 K at Low Pressure. *J. Phys. Chem. A* **2008**, *112*, 7227.

(128) Li, X.; Moore, D. T.; Iyengar, S. S. Insights from First Principles Molecular Dynamics Studies Towards Infra-Red Multiple-Photon and Single-Photon Action Spectroscopy: Case Study of the Proton-Bound Di-Methyl Ether Dimer. *J. Chem. Phys.* **2008**, *128*, 184308.

(129) Li, X.; Oomens, J.; Eyler, J. R.; Moore, D. T.; Iyengar, S. S. Isotope Dependent, Temperature Regulated, Energy Repartitioning in a Low-Barrier, Short-Strong Hydrogen Bonded Cluster. *J. Chem. Phys.* **2010**, *132*, 244301.

(130) Dietrick, S. M.; Iyengar, S. S. Constructing Periodic Phase Space Orbits from Ab Initio Molecular Dynamics Trajectories to Analyze Vibrational Spectra: Case Study of the Zundel ($H_5O_2^+$) Cation. *J. Chem. Theory Comput.* **2012**, *8*, 4876.

(131) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: New York, 1992.

(132) Manning, C. D.; Raghavan, P.; Schutze, H. *Introduction to information retrieval*; Cambridge University Press: New York, 2008; pp 279−288.

(133) Krishnan, M.; Gupta, V. Vibration Spectra of Alpha−helix of Poly−alanine. *Chem. Phys. Lett.* **1970**, *6*, 231.

(134) Dousseau, F.; Pézolet, M. Determination of the Secondary Structure Content of Proteins in Aqueous Solutions from Their Amide I and Amide II Infrared Bands. Comparison between Classical and Partial Least-Squares Methods. *Biochemistry* **1990**, *29*, 8771.

(135) Björklund, A.; Husfeldt, T.; Koivisto, M. Set Partitioning via Inclusion Exclusion. *SIAM J. Comput.* **2009**, *39*, 546.