Received: 20 December 2019 | Revised: 19 March 2020 | Accepted: 20 March 2020

DOI: 10.1002/qua.26244

FULL PAPER

International Journal of QUANTUM CHEMISTRY WILEY

Embedded, graph-theoretically defined many-body approximations for wavefunction-in-DFT and DFT-in-DFT: Applications to gas- and condensed-phase ab initio molecular dynamics, and potential surfaces for quantum nuclear effects

Timothy C. Ricard | Anup Kumar | Srinivasan S. Iyengar

Department of Chemistry and Department of Physics, Indiana University, Bloomington, Indiana

Correspondence
Srinivasan S. Iyengar, Department of Chemistry and Department of Physics, Indiana University, 800 E. Kirkwood Ave, Bloomington, IN 47405.
Email: iyengar@indiana.edu

Funding information
XSEDE, Grant/Award Number: CHE170081; National Science Foundation, Grant/Award Number: CHE-1665336

Abstract

We present a graph-theoretic approach to adaptively compute many-body approximations in an efficient manner to perform (a) accurate post-Hartree–Fock (HF) ab initio molecular dynamics (AIMD) at density functional theory (DFT) cost for medium- to large-sized molecular clusters, (b) hybrid DFT electronic structure calculations for condensed-phase simulations at the cost of pure density functionals, (c) reduced-cost on-the-fly basis extrapolation for gas-phase AIMD and condensed phase studies, and (d) accurate post-HF-level potential energy surfaces at DFT cost for quantum nuclear effects. The salient features of our approach are ONIOM-like in that (a) the full system (cluster or condensed phase) calculation is performed at a lower level of theory (pure DFT for condensed phase or hybrid DFT for molecular systems), and (b) this approximation is improved through a correction term that captures all many-body interactions up to any given order within a higher level of theory (hybrid DFT for condensed phase; CCSD or MP2 for cluster), combined through graph-theoretic methods. Specifically, a region of chemical interest is coarse-grained into a set of nodes and these nodes are then connected to form edges based on a given definition of local envelope (or threshold) of interactions. The nodes and edges together define a graph, which forms the basis for developing the many-body expansion. The methods are demonstrated through (a) ab initio dynamics studies on protonated water clusters and polypeptide fragments, (b) potential energy surface calculations on one-dimensional water chains such as those found in ion channels, and (c) conformational stabilization and lattice energy studies on homogeneous and heterogeneous surfaces of water with organic adsorbates using two-dimensional periodic boundary conditions.

KEYWORDS

AIMD, graph-theoretic treatment, many-body theory, molecular fragmentation, potential energy surfaces

1 | INTRODUCTION

Accurate treatment of molecular properties often requires the correlated study of electronic structure with extensive basis sets. But the size of systems that can be considered by standard approaches in electronic structure is greatly limited due to the intrinsic computational scaling of electron correlation methods and the number and quality of the basis functions used. Although there has been significant progress through algorithmic improvements for the treatment of electron correlation,[1–6] the computational expense is still often too high for most systems of chemical interest. This is especially true when dynamics calculations are to be performed with on-the-fly electronic structure or when quantum nuclear effects are to be studied. Alternatively, application of density functional theory (DFT) has allowed the treatment of moderately sized systems at reasonable accuracy and computational scaling costs, although there still remain significant challenges.[7–9] Another approach that has gained popularity in the past two decades is the treatment of local interactions with high-quality ab initio methods, which are then combined to form composite energy expressions for the full system. Methods that employ this idea are often presented as embedding or fragmentation approaches.[10–27] These approaches are complementary to the intensive work on many-body methods[28–36] that have also greatly contributed towards the development of reduced scaling methods. These methods have allowed localized treatments of correlation with larger basis sets with substantial computational cost reductions and often in a manner that is trivially parallelizable.[37,38]

Inspired by these developments, in a series of publications[39–45] we have shown how graph-theoretic methods may be used to adaptively construct many-body approximations to potential surfaces and ab initio molecular dynamics (AIMD) calculations. In the process, both extended Lagrangian and Born–Oppenheimer-based AIMD simulations can be performed at accuracy comparable to coupled cluster singles and doubles (CCSD) and second-order Møller-Plesset (MP2) levels of theory with DFT computational cost.[39–42] Hence, for the first time, in refs. [40,41] we presented Car–Parrinello-style dynamics, but with CCSD accuracy. Similarly, we have shown how multiple graphical representations of molecular systems may be simultaneously utilized to construct accurate potential surfaces in agreement with MP2 and CCSD levels of theory, again at DFT cost.[44] In ref. [43], we have also shown that weak interactions (specifically hydrogen bonds) can be accurately captured and efficient approximations to large-basis AIMD trajectories, such as 6-311++G(2df,2pd), can be constructed through computational effort commensurate with much smaller basis set sizes, sets such as 6-31+G(d). Furthermore, we have also shown in ref. [45] how condensed-phase simulations on interfaces and liquids may be constructed with hybrid DFT accuracy at gradient-corrected DFT accuracy. The approach utilizes graph-theoretic methods to perform an ONIOM-type decomposition of the system where the individual extrapolatory components in the ONIOM expression are obtained from many-body theory.

The critical aim of these methods is to make it possible to construct accurate post-Hartree–Fock (HF) methods with large basis sets for accurate treatment of molecular clusters as well as condensed-phase surface problems. The conventional methods that approach chemical accuracy[46–49] scale catastrophically as the system size grows. Because of these prohibitive costs, most AIMD calculations are limited to “on-the-fly” DFT calculations with a modest basis, but these studies are restricted in many ways.[7,50,51] As illustrated in Figure 1A, one target of electronic structure theory, which is its gold standard, is to provide good estimates for properties derived from post-HF theories in the large basis limit. The absolute limit of these calculations is the full-CI limit. When nuclear degrees of freedom of a system are also

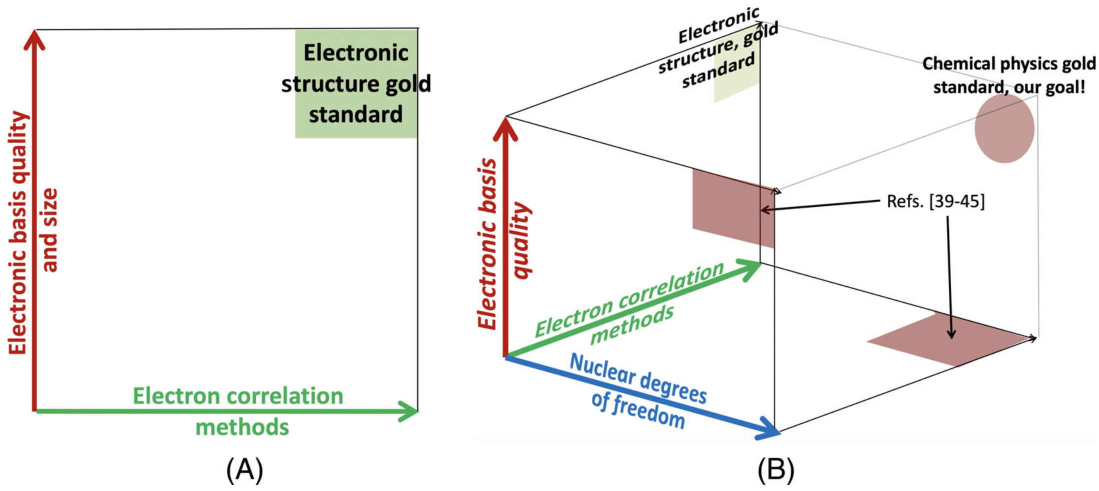
Electronic structure gold standard
Electronic basis quality and size
Electron correlation methods
(A)
Electronic structure, gold standard
Chemical physics gold standard, our goal!
Refs. [39-45]
Electronic basis quality
Electron correlation methods
Nuclear degrees of freedom
(B)

FIGURE 1 A, Gold standard for electronic structure treatment in the limit of large basis sets and accurate post-Hartree-Fock methods. The complexity of this gold standard is expanded through consideration of nuclear degrees of freedom B, that is, classical degrees of freedom in ab initio molecular dynamics and quantum freedom when considering potential energy surfaces

considered (Figure 1B)—whether in AIMD, potential energy surface calculations, or other similar treatments—this problem becomes even more complicated through the shear exponential scaling nature[52–55] of electron-nuclear methods. For the case of potential energy surfaces, the number of such calculations can be significantly reduced by sampling methods,[56,57] and data storage bottlenecks may be addressed through sophisticated techniques such as tensor networks.[57,58] The graph-theoretic approach to adaptive many-body approximations described here are one facet of the grand attempt to achieve the larger goal of computing accurate AIMD and potential surfaces at reduced computational costs.

The rest of the paper is organized as follows: In Section 2 we present a brief survey of our graph-theoretic approach to low-cost electronic structure and many-body theory, AIMD, and potential surface calculations. Computational studies are presented in Section 3. Conclusions are given in Section 4.

2 | EMBEDDED LOCAL MANY-BODY INTERACTIONS FOR AB INITIO MOLECULAR DYNAMICS AND POTENTIAL SURFACES

The main features of the method presented in refs. [39-45] is the partitioned treatment of a large reactive system based on the significance of short-range and long-range correlation and basis set completeness effects. In a spirit compatible with the ONIOM formalism,[59] the long-range effects are treated at a lower level of theory whereas the short-range effects in turn are treated as a many-body expansion that extrapolates between a lower and a higher level of theory to provide an estimate of the total electronic energy of a system at a higher level of theory. Thus

$$E^{\text{Embed}}_{\mathscr{R}} = E^{\text{level,0}} + E^{\text{level,1}}_{\text{MBE},\mathscr{R}} - E^{\text{level,0}}_{\text{MBE},\mathscr{R}} \tag{1}$$

where the quantity $\mathscr{R}$ represents the level of the many-body expansion. The innovative idea in refs. [39-45] is that the terms on the right side of Equation (1) are obtained using graph-theoretic and set-theoretic paradigms. Specifically

$$E^{\text{level,1}}_{\text{MBE},\mathscr{R}} \equiv \sum_{r=0}^{\mathscr{R}} (-1)^r \left\{ \sum_{\alpha_r}^{r-\text{rank}} E^{\text{level,1}}(\alpha_r, r) \left[\sum_{m=r}^{\mathscr{R}} (-1)^m p^{r,m}_{\alpha_r} \right] \right\} \tag{2}$$

where the summation over α_r is over all rank-r or order-r many-body terms, and $E^{\text{level, 1}}(\alpha_r, r)$ is an electronic energy for the α_rth rank-r many-body contribution. In Section 2.1 we develop a graph-theoretic procedure to evaluate Equation (1), where the order-r many-body terms are obtained from rank-r geometric objects known as simplexes.[60] As a result, we use order-r many-body terms, rank-r many-body terms, and eventually, following Section 2.1, rank-r simplexes in an interchangeable manner. A similar expression as Equation (2) can be written for $E^{\text{level,0}}_{\text{MBE},\mathscr{R}}$, which is a function of $E^{\text{level, 0}}(\alpha_r, r)$. The square-bracketed term in Equation (2) contains an over-counting correction,[42,43] where $p^{r,m}_{\alpha_r}$ refers to the number of times the α_rth rank-r term appears in all rank-m terms ($m \geq r$), with phase $(-1)^m$. This over-counting correction is along similar lines as those present in molecular fragmentation[11–13,26,33,61–67] as well as many-body and double many-body expansions.[28,30–35] Together, $E^{\text{level,1}}_{\text{MBE},\mathscr{R}}$ and $E^{\text{level,0}}_{\text{MBE},\mathscr{R}}$ in Equation (1) provide an ONIOM-like perturbation to $E^{\text{level, 0}}$ and, in ONIOM-style, the electronic energy for the α_rth rank-r many-body-term is computed at two levels of theory, referred to as "level,1" and "level,0" above. Combining Equations (1) and (2), the final expression for the graph-theoretic many-body embedding energy is

$$E^{\text{Embed}}_{\mathscr{R}} = E^{\text{level,0}} + \sum_{r=0}^{\mathscr{R}} (-1)^r \left\{ \sum_{\alpha_r}^{r-\text{rank}} \Delta E(\alpha_r, r) \left[\sum_{m=r}^{\mathscr{R}} (-1)^m p^{r,m}_{\alpha_r} \right] \right\} \tag{3}$$

where $\Delta E(\alpha_r, r)$ is an ONIOM-like energy correction or extrapolation term for the α_rth rank-r many-body term:

$$\Delta E(\alpha_r, r) = E^{\text{level,1}}(\alpha_r, r) - E^{\text{level,0}}(\alpha_r, r). \tag{4}$$

In this manner, local defects, reactive effects, and weak interactions are treated at the necessary higher level of electronic structure and larger basis, level, 1, whereas long-range effects are considered through an ONIOM-like extrapolation procedure, all of this couched within the successful many-body-expansion[28,30–33] procedure. The central idea here is the introduction of long-range interactions at a lower level of theory (such as DFT functionals using smaller basis sets), with the perturbative introduction of higher levels of theory (such as hybrid density functionals, MP2, and coupled cluster theories) and larger basis sets using local many-body expansions that are obtained from graph theory. Hence, the method discussed in this work aims to form an approximation to the overall energy of a large molecular system through the addition of perturbative local terms that progressively build a suitable approximation to the full system electronic structure.

2.1 | Graph-theoretic representation of Equation (1)

The many-body contributions in Equation (3) form an ordered set, where α_r identifies one particular rank-r or order-r many-body term from a set of all rank-r many-body terms V_r, that is, $\{\alpha_r | \alpha_r \in V_r; r = 0 \cdots \mathscr{R}\}$, where V_r is the set of all rank-r many-body terms up to a maximum rank $\mathscr{R}$. The critical contribution in refs. [42,44] is to use graph theory to simplify this process. That is, $E^{\text{level},1}_{\text{MBE},\mathscr{R}}$ and $E^{\text{level},0}_{\text{MBE},\mathscr{R}}$ provide an ONIOM-like perturbation to $E^{\text{level},0}$ in Equation (1) obtained in refs. [42-44] from a graph-theoretic decomposition of the molecular structure. We have shown[42,44,45] that this procedure captures the accuracy at higher levels of theory and basis (level, 1 above) at computational costs commensurate with that from level, 0. The low computational cost associated with this procedure may be inferred from Figure 2.

Here, the region of chemical interest is coarse-grained into a set of nodes, or vertices, and these nodes are then connected to form edges based on a given definition of local envelope of interactions (or edge formation threshold). This "local envelope" essentially defines the spatial extent to which two-body interactions are to be considered. In the graph-theoretic scheme, it then defines a graph and, by inference, edges (two-body interactions), triangles (three-body interactions), and higher order simplexes (or arbitrary-rank many-body interactions), which are then employed with Equations (1) and (2). In this work, we discuss results from distance-based[39–41,44,45] and connectivity-based[42,43] envelopes. In ref. [42], we have also evaluated procedures that utilized Delaunay triangulation[68–75] of coarse-grained representations of molecular structure to compute local envelopes. Our conception of these local interaction envelopes is similar to the distance-based cutoffs developed in fragment molecular orbital (FMO)[76–78] and effective FMO (EFMO),[79,80] and are currently employed in methods such as molecules-in-molecules (MIM),[62,81] generalized many-body expansion (GMBE),[17,33] and hybrid many-body interaction (HBMI),[82,83] along with many other methods discussed in refs. [11,13,36,84].

The nodes and edges together define a graph $\mathcal{G}$, which forms the basis for developing the many-body expansions as per Equation (3). Specifically, the nodes and edges define faces, tetrahedrons, and other higher order objects referred to as simplexes in the graph, and these capture local (bonded/nonbonded) as well as nonlocal (nonbonded) many-body interactions. Such a geometrical decomposition is shown in Figure 3. Furthermore, the nodal structure or graph in Figure 3 has an associated topological invariance known as an Euler characteristic,[85] χ, defined in Equation (5), and guides our definition of $E^{\text{level},1}_{\text{MBE},\mathscr{R}}$ and $E^{\text{level},0}_{\text{MBE},\mathscr{R}}$:

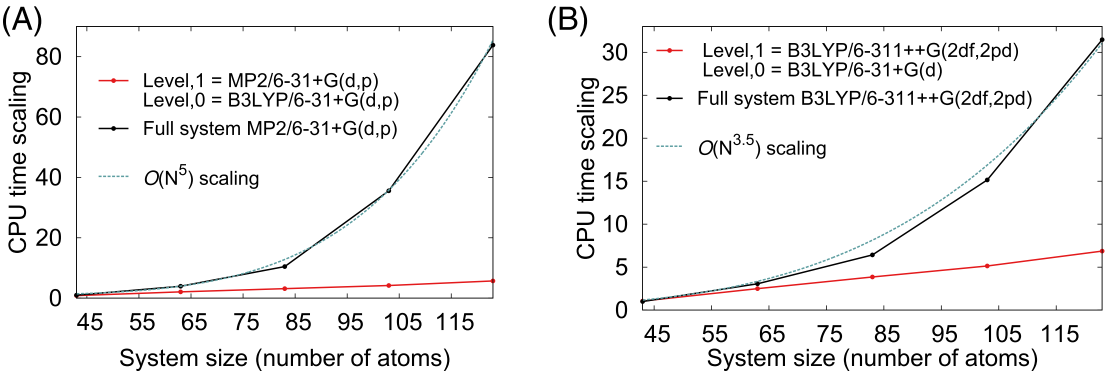

FIGURE 2 Computational scaling for implementation of Equation (3) (in red) contrasted with the case where the full system calculations are performed at level, 1 theory (MP2 for A, and the larger basis set for B). The system considered for this figure is polyalanine, $(\text{Ala})_n$ for number of monomers ranging from 4 to 12. The CPU scaling (vertical axis) in both figures is presented as a ratio with respect to the computational cost for $(\text{Ala})_4$

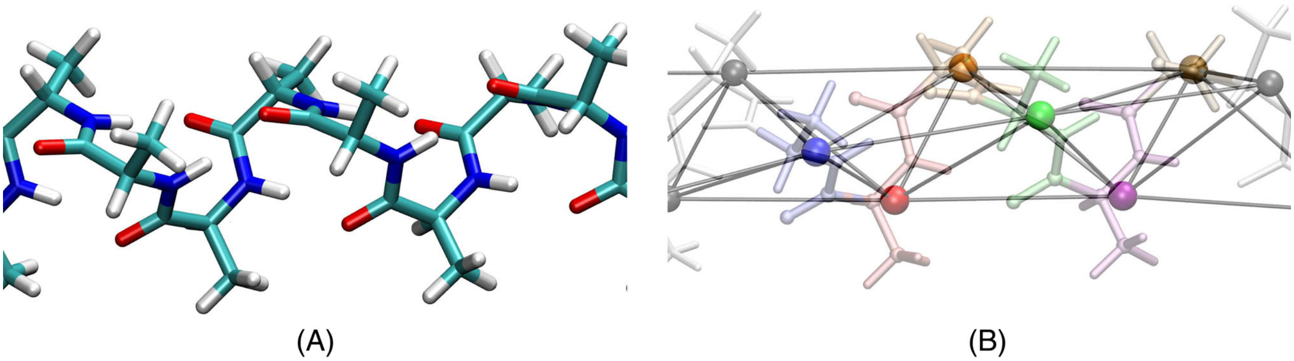

FIGURE 3 Illustration of the graphical representation for a 3_{10} helical form of polyalanine. The Ala_N structure A, is partitioned into nodes, which are then connected by edges B, to define higher order simplexes or higher order many-body expansions as in Equation (3)

$$\chi = \eta_0 - \eta_1 + \eta_2 - \cdots + (-1)^r \eta_r + \cdots + (-1)^{\mathcal{R}} \eta_{\mathcal{R}} = \sum_{r=0}^{\mathcal{R}} (-1)^r \eta_r. \tag{5}$$

Here, η_r is the number of geometric entities (nodes, edges, faces, and higher order simplexes) of rank r. That is, η_0 is the number of nodes (rank-0 simplexes) in the graph, η_1 is the number of edges (rank-1 simplexes) in the graph, η_2 is the number of faces (rank-2 simplexes) in the graph, and so on.[60,68,71–73,75,85,86] The quantity $\mathcal{R}$ is the largest simplex rank included in the truncated expansion on the right side of Equation (5), which in general can go up to arbitrary orders

Following the graphical decomposition of molecular system as per the above prescriptions, in refs. [44,45], we replace the appearance of each rank-r simplex in Equation (5), numbered using the index α_r, by an energy correction corresponding to the molecular interactions captured by the specific simplex:

$$\eta_r \to \left\{ \sum_{\alpha_r}^{r-\text{rank}} \Delta E(\alpha_r, r) \left[\sum_{m=r}^{\mathcal{R}} (-1)^m p_{\alpha_r}^{r,m} \right] \right\} \tag{6}$$

where the summation over α_r is over all rank-r simplexes, that is, rank-r many-body expansions. Furthermore, by comparison with Equation (3), it is clear that Equation (6) represents one part of the summation in the many-body expansion, and hence here the graphical decomposition is essentially used to adaptively determine the many-body expansion that is *embedded* within a calculation that represents the full system, consistent with Equation (3). That is, with the substitution of Equation (6) into Equation (5), we again recover the many-body correction in Equation (3), where $E_{\text{MBE},\mathcal{R}}^{\text{level},1} - E_{\text{MBE},\mathcal{R}}^{\text{level},0}$ is now a perturbative term based on the Euler characteristic of the graph:

$$E_{\text{MBE},\mathcal{R}}^{\text{level},1} - E_{\text{MBE},\mathcal{R}}^{\text{level},0} \equiv \sum_{r=0}^{\mathcal{R}} (-1)^r \left\{ \sum_{\alpha_r}^{r-\text{rank}} \Delta E(\alpha_r, r) \left[\sum_{m=r}^{\mathcal{R}} (-1)^m p_{\alpha_r}^{r,m} \right] \right\} \tag{7}$$

This perturbative term corrects the full system energy $E^{\text{level},0}$ to reproduce Equation (3), where the difference of two many-body expansions is re-interpreted as a graph-based correction. The advantage of this graphical interpretation is the ability to consider the energy correction up to arbitrary many-body terms. For example, a given spatial envelope, that is, edge definition or range of maximum spatial range of two-body interactions, in conjunction with the set of nodes, or monomers, defines a graph. Following this, truncation of Equation (2) through $\mathcal{R}$-rank simplexes in the graph includes many-body terms containing $\mathcal{R}$ + 1 monomers or nodes. In addition, we also have the flexibility in Equation (2) to define a maximum local rank, that is, local values of $\mathcal{R}$. Thus, the availability of larger rank simplexes is dependent on the connectivity of the nodes within the graphical representation of the system. As the number of available edges surrounding any given node within the graph increases, the available faces, tetrahedrons, and other higher order simplexes would increase as well. *Furthermore, the number of available edges surrounding any given node in turn depends on the local proximity of these groups and hence influences the extent to which many-body (bonded and nonbonded) approximations contribute to the energy*. In refs. [39-45], we have shown this adaptive and general form of electronic energy to be a robust and powerful way to estimate higher order correlation effects at the cost of DFT. To further clarify Equations (1), (3), and (7) in Appendix A, we explicitly write out these expressions for a few increasing values of $\mathcal{R}$.

2.2 | Formal scaling of the many-body expansions in Equation (3) facilitated by the graph-theoretic decomposition in Section 2.1

Since the goal of Equation (3) is to reduce the computational effort to allow a calculation of quality previously untenable at a practical cost, we discuss the formal computational complexity associated with Equation (3). In order to quantify the discussion, we will assume that the target molecular system is divided into $\mathcal{N}$ nodes with basis set size of $\mathcal{M}$, which determines the computational complexity for the molecular fragment associated with each node. Thus the size of basis set for the full system is roughly $[\mathcal{N} * \mathcal{M}]$. A computationally sequential (or serial) treatment of Equation (3) leads to the evaluation of each r-body term at two levels of theory and the full system at level, 0. If we further assume that the level, 0 calculation scales as $\mathcal{O}(N^{L_0})$ and level, 1 scales as $\mathcal{O}(N^{L_1})$, then

$$\text{Scaling}_{Eq.(3)}^{\text{serial}} = \mathcal{O}\left([\mathcal{N} * \mathcal{M}]^{L_0}\right) + \sum_{r=0}^{\mathcal{R}} C_r \left\{ \mathcal{O}\left([\mathcal{M} * r]^{L_1}\right) + \mathcal{O}\left([\mathcal{M} * r]^{L_0}\right) \right\} \tag{8}$$

where C_r is the number of r-rank many-body terms considered, which, for a given local spatial envelope, would be less than the ${}_{\mathcal{N}}C_r$ (total) number of r-body interaction terms. Since the polynomial scaling for the higher level of theory is larger than that for the lower level of theory, the r-rank many-body terms for the lower level of theory can be neglected from scaling considerations, leaving two terms that dominate the overall scaling: the full system at the low level, and the summation of the *r* many-body terms at the higher level of theory. Furthermore, only the first term is system-size dependent, and hence the overall scaling of the calculation in Equation (3) is extremely favorable as compared to the full-system level, 1 theory calculations as seen from Figure 2, for both post-HF treatment and basis set extrapolation.

All terms in Equation (8) may, in principle, be computed in a concurrent (or parallel) manner in any order, allowing for connections to (and associated computational improvements from) the well-known chemical abstract machines[87–89] model of concurrency in computer science. While these ideas will be further explored in future studies, our current parallel implementation of Equation (3) uses MPI-parallelism within a C++ module to arrive at an effective computational complexity given by

$$\text{Scaling}_{Eq.(3)}^{\text{parallel}} = \mathcal{O}\left([\mathcal{N} * \mathcal{M}]^{L_0}\right) + \mathcal{O}\left([\mathcal{M} * \mathcal{R}]^{L_1}\right) \tag{9}$$

Now, given that $\mathcal{M}$ is the number of basis functions for a single node and is not system size ($\mathcal{N}$) dependent, the parallel computational implementations of Equation (3) is dominated by the first term, resulting in

$$\text{Scaling}_{Eq.(3)}^{\text{parallel}} \overset{\text{Large } \mathcal{N}}{\longrightarrow} \mathcal{O}\left([\mathcal{N} * \mathcal{M}]^{L_0}\right) \tag{10}$$

A similar argument can be made for basis set extrapolation, where the full system at a smaller basis dominates the calculation, that provides results in agreement with much larger basis sets. The associated scaling curves are shown in Figure 2B, as described in ref. [43]. The basis set extrapolations demonstrated in ref. [43] maintain the same electronic structure methods for both the high and low levels of theory with only the basis set choice changing, as seen from Figure 2B. As we will see later, and as shown in refs. [39-45], Equation (3) does indeed provide a very good estimate of higher level electronic structure energies and gradients, and hence through this approach we achieve high-quality calculations at reduced computational cost. In summary

- In refs. [39-42] we demonstrated DFT scaling while achieving MP2 and CCSD accuracy during AIMD simulations for protonated water clusters and polypeptide fragments. The associated computational scaling is depicted in Figure 2A.
- In ref. [43], DFT-based AIMD trajectories in agreement with a highly diffuse triple-zeta atom-centered Gaussian basis set were obtained at computational costs commensurate with double zeta basis sets. Figure 2B shows the associated scaling curves.
- In ref. [44], we demonstrated MP2-quality potential energy surfaces at DFT costs using multiple graphical descriptions of the system.
- In ref. [45] we adapted this approach to compute condensed-phase electronic-structure-based lattice energies for adsorbates on the surface of water, using higher rung DFT functionals[50,90,91] at much reduced costs. The approach also allows large basis set studies in condensed-matter systems.
- Our studies are performed using an MPI parallelized C++ module with embedded python function calls that computes the energy and forces by spawning out the rank-*r* simplex calculations. There is no restrictions in terms of using any specific electronic structure package during the rank-*r* simplex calculations and hence in our AIMD and potential surface calculations. In fact, in ref. [42] the studies incorporate energies and gradients from multiple electronic structure packages for the rank-*r* simplexes. For example, the current version of our computer program is capable of using Gaussian,[92] Psi4,[93] Orca,[94] Quantum Espresso,[95] and the OpenMX[96] electronics structure packages simultaneously during AIMD and potential surfaces, with our immediate goal being to incorporate condensed-phase versions from packages such as NWChem[97] and Vienna ab-initio simulation package (VASP).[98]

The above set of applications is briefly discussed in Section 3.

2.3 | Born–Oppenheimer and extended Lagrangian, atom-centered density matrix propagation-pHF, AIMD implementations from Equation (1)

The nuclear gradients for the energy in Equation (1) may be written as

$$\frac{\partial E_{\mathcal{R}}^{\text{Embed}}(\mathbf{R})}{\partial \mathbf{R}} = \frac{\partial E^{\text{level,0}}(\mathbf{R})}{\partial \mathbf{R}} + \sum_{r=0}^{\mathcal{R}} (-1)^r \left\{ \sum_{\alpha_r}^{r-\text{rank}} \left(\frac{\partial E^{\text{level,1}}(\alpha_r, r)}{\partial \mathbf{R}_{\alpha_{r,r}}} \right) - \left(\frac{\partial E^{\text{level,0}}(\alpha_r, r)}{\partial \mathbf{R}_{\alpha_{r,r}}} \right) \left(\frac{\partial \mathbf{R}_{\alpha_{r,r}}}{\partial \mathbf{R}} \right) \left[\sum_{m=r}^{\mathcal{R}} (-1)^m p_{\alpha_r}^{r,m} \right] \right\} \tag{11}$$

The nuclear coordinates for the molecular fragment or subsystem representing the α_rth r-rank many-body term $\mathbf{R}_{\alpha_{r,r}}$ may not be entirely a subset of the system coordinate variable $\mathbf{R}$, as it may include link atoms if bonds are broken in the formation of the nodal definitions in the graph, as allowed by ONIOM.[59] If link atoms are included, then $\left[\frac{\partial \mathbf{R}_{\alpha_{r,r}}}{\partial \mathbf{R}}\right]$ is a Jacobian term needed to transform the r-rank many-body gradients back to the full system gradients.[39,99] The formation of these gradients enable the direct use of Born–Oppenheimer MD using Equation (3), as demonstrated in refs. [39-43]. The associated methodology has been referred to as “frag-BOMD” in refs. [39-43]. To further clarify Equation (11) and emphasize the fragment gradient terms that appear in Appendix B, we explicitly write out these expressions for a few increasing values of $\mathcal{R}$.

As the size of the system considered increases, the full system gradients $\left[\frac{\partial E^{\text{level},0}(\mathbf{R})}{\partial \mathbf{R}}\right]$ dominate the overall computational costs for AIMD treatment. This limitation is already clear from Equation (11), in light of Equation (10), where it is shown that the scaling of the extrapolation is constrained only by the full-system lower level calculations in the large system limit. (Also see Figure 2 where scaling curves are presented and the method represented by Equation (3) is constrained *only* by the choice of lower level of theory). It is possible to reduce this complexity further by introducing linear scaling methods[100–105] with SCF parallelism[97,100] for the full-system low-level calculation. But in refs. [40,41,43] *a r-rank many-body extended Lagrangian*[106,107] implementation is introduced where the electronic parameters (such as the single-particle density matrix $\mathbf{P}_{\mathbf{level,0}}$ for single-particle treatment at level, 0) that depict the energy for the full-system low-level calculation, $E^{\text{level},0}$ in Equation (3) are treated as dynamical variables. Specifically, here the electronic parameters $\mathbf{P}_{\mathbf{level,0}}$ that determine $E^{\text{level},0}$ are propagated along with the nuclear degrees of freedom through an adjustment of the relative timescales between the full system, low-level treatment, and nuclear degrees of freedom. This is essentially a Car–Parrinello-like method,[108] but implemented using the atom-centered Gaussian basis functions and single-particle density matrices that determine $E^{\text{level},0}$ and hence follow the atom-centered density matrix propagation (ADMP)[109–112] protocol. One of the critical results of this treatment is the ability to calculate post-HF-based extended Lagrangian (Car–Parrinello-like) trajectories, as the underlying density matrix $\mathbf{P}_{\mathbf{level,0}}$ remains within a single-particle formalism. This methodology is termed “atom-centered density matrix propagation with post-HF accuracy” and abbreviated as ADMP-pHF.[40,41] The associated extended Lagrangian, which serves for post-HF accuracy as well as basis-set-extrapolated dynamics, is given by

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\mathbf{Tr}\left[\mathbf{V}^{\mathbf{T}}\mathbf{M}\mathbf{V}\right] + \frac{1}{2}\mathbf{Tr}\left[\left(\boldsymbol{\mu}_{\text{level},0}^{1/4}\ \mathbf{W}_{\mathbf{level,0}}\ \boldsymbol{\mu}_{\text{level},0}^{1/4}\right)^2\right] - E_{\mathcal{R}}^{\text{embed}}\left(\mathbf{R},\mathbf{P}_{\mathbf{level,0}}\right)\\ -\mathbf{Tr}\left[\boldsymbol{\Lambda}_{\mathbf{level,0}}\left(\mathbf{P}_{\mathbf{level,0}}^{\mathbf{2}} - \mathbf{P}_{\mathbf{level,0}}\right)\right]\end{aligned} \tag{12}$$

Here, the parameters $\mathbf{R}$ and $\mathbf{V}$ represent the classical nuclear positions and velocities, with masses $\mathbf{M}$. The single-particle density matrix $\mathbf{P}_{\mathbf{level,0}}$ represents the full system, at the lower level of theory, and is propagated to determine $E^{\text{level},0}$, which is part of $E_{\mathcal{R}}^{\text{embed}}$ in Equation (12) (Equation (3)). The velocity Verlet113 integration technique is used to evolve the dynamical parameters of the full system {$\mathbf{R}$, $\mathbf{V}$; $\mathbf{P}_{\mathbf{level,0}}$, $\mathbf{W}_{\mathbf{level,0}}$}.

This extended Lagrangian dynamics of the density matrix $\mathbf{P}_{\mathbf{level,0}}$ is tempered by the fictitious velocity $\mathbf{W}_{\mathbf{level,0}}$ (in the spirit of Car–Parrinello[108] and ADMP[109,110]) with fictitious inertia tensor $\boldsymbol{\mu}_{\mathbf{level,0}}$. This fictitious inertia tensor is constructed by fixing the inertia for valence electrons at a specific value μ_{valence}, and weighting the core orbitals using μ_{valence} and appropriate diagonal elements of the single-particle Fock matrix $\mathbf{F}_{\mathbf{level,0}}$, which determines $E_{\text{level},0.}$109,110 Specifically, the elements of $\boldsymbol{\mu}_{\text{level},0}$ are represented using a diagonal scaling tensor $\mathbf{A}_{\mathbf{level,0}}$, with elements defined as

$$\begin{aligned}&{\mathbf{A}_{\mathbf{level,0}}}^{i,i} = \mathbf{I}, && {\mathbf{F}_{\mathbf{level,0}}}^{ii} > -2\ a.u.\\ &{\mathbf{A}_{\mathbf{level,0}}}^{i,i} = \left[2\left|{\mathbf{F}_{\mathbf{level,0}}}^{ii} + 2\right|^{1/2} + 1\right]^2, && {\mathbf{F}_{\mathbf{level,0}}}^{ii} < -2\ a.u.\\ &{\mathbf{A}_{\mathbf{level,0}}}^{i,j} = 0, && i \neq j.\end{aligned} \tag{13}$$

and

$$\boldsymbol{\mu}_{\mathbf{level,0}} \equiv \mu_{\text{valence}}\mathbf{A}_{\mathbf{level,0}} \tag{14}$$

This particular choice of mass tensor has been used for many single-particle ADMP applications[39,99,111,114–126] and also for ADMP-pHF in refs. [40-43].

The choice of the fictitious inertia tensor scaling factor μ_{valence} determines deviations from the Born–Oppenheimer surface. The precise deviations from the Born–Oppenheimer surface have been discussed in Appendix A of ref. [40] (informed by the discussions in refs. [110,112]), and additional forces (which only nominally contribute to the computational expense[110,112]) arise in the extended Lagrangian dynamics. These additional forces modify the level, 0 contributions to the embedding gradient in Equation (11) and are given by

$$\begin{aligned}&\left.\frac{\partial E_{\text{ADMP}-\text{pHF}}^{\text{level},0}\left(\mathbf{R},\mathbf{P}_{\mathbf{level,0}}\right)}{\partial \mathbf{R}_{\mathbf{i,j}}}\right|_{\mathbf{P}_{\text{level},0}} - \left.\frac{\partial E_{BO}^{\text{level},0}(\mathbf{R})}{\partial \mathbf{R}_{\mathbf{i,j}}}\right|_{\mathbf{P}_{\text{level},0}}\\ &= Tr\left[\left[\tilde{\mathbf{P}}_{\text{level},0},\mathbf{F}_{\text{level},0}\right]\left(\frac{d\mathbf{U}_{\text{level},0}}{\mathbf{R}_{\mathbf{i,j}}}\mathbf{U}_{\text{level},0}^{-\mathbf{1}} - \tilde{\mathbf{P}}_{\text{level},0}\mathbf{U}_{\text{level},0}^{-\mathbf{T}}\frac{d\mathbf{S}'_{\text{level},0}}{\mathbf{R}_{\mathbf{i,j}}}\mathbf{U}_{\text{level},0}^{-\mathbf{1}}\right)\right]\end{aligned} \tag{15}$$

where $\tilde{\mathbf{P}}_{\text{level,0}}$ is the McWeeny purified[127] form of the density matrix, that is, $\tilde{\mathbf{P}}_{\text{level,0}} = 3\mathbf{P}^2_{\text{level,0}} - 2\mathbf{P}^3_{\text{level,0}}$. The quantity $\mathbf{S}'_{\text{level,0}}$ is the overlap between the atomic orbital bases used in $E^{\text{level,0}}$ calculation, and $\mathbf{U}_{\text{level,0}}$ is the associated transformation matrix to an orthonormal basis (Löwdin symmetric orthogonalization and Cholesky transformations are used here). That is, $\mathbf{S}'_{\text{level,0}} \equiv \mathbf{U}_{\text{level,0}}{}^{\mathrm{T}}\mathbf{U}_{\text{level,0}}$. The additional nuclear forces on the right side of Equation (15) are proportional to the commutator of the single-particle description of the full system, that is, the associated Fock matrix $\mathbf{F}_{\text{level,0}}$ and the density matrix $\mathbf{P}_{\text{level,0}}$.

The Lagrange multiplier matrix $\mathbf{\Lambda}_{\text{level,0}}$ maintains the N-representability[109,110,127] of $\mathbf{P}_{\text{level,0}}$. This is done (a) by conserving the idempotency of $\mathbf{P}_{\text{level,0}}$ through an iterative procedure,[40,41,110] and (b) by conserving the particle number. Thus, to assemble the full system energy in Equations (12) and (3), we utilize the propagated density matrix $\mathbf{P}_{\text{level,0}}$ to obtain $\frac{\partial E^{\text{level,0}}}{\partial \mathbf{R}}$, and the remaining parts of the embedding energy in Equation (3) are obtained through regular electronic structure calculations including gradients on the fragments with low- and high-level treatments, yielding $\frac{\partial E^{\text{level,1}}(\alpha_r,r)}{\partial \mathbf{R}_{\boldsymbol{\alpha}_\mathbf{r},\mathbf{r}}}$ and $\frac{\partial E^{\text{level,0}}(\alpha_r,r)}{\partial \mathbf{R}_{\boldsymbol{\alpha}_\mathbf{r},\mathbf{r}}}$. The gradients with respect to the density matrix are computed as described in refs. [40,109,110], as

$$\left.\frac{\partial E^{\text{Embed}}_{\mathscr{R}}(\mathbf{R},\mathbf{P}_{\text{level,0}})}{\partial \mathbf{P}_{\text{level,0}}}\right|_{\mathbf{R}} \equiv \left.\frac{\partial E^{\text{level,0}}(\mathbf{R})}{\partial \mathbf{P}_{\text{level,0}}}\right|_{\mathbf{R}} = \left[\left[\tilde{\mathbf{P}}_{\text{level,0}},\mathbf{F}_{\text{level,0}}\right],\tilde{\mathbf{P}}_{\text{level,0}}\right] \tag{16}$$

and these gradients tend to zero as the density matrix converges (as is the case in frag-BOMD).

2.4 | Hypergraph, multi-topology formalism for potential energy surfaces

In the previous sections and in refs. [39-45], we discussed the formation and efficient evaluation of correlation energy and basis-set extrapolation contributions based upon a graphical presentation of many-body theory for molecular systems and condensed-phase ensembles. Equation (3) achieves a perturbative enhancement of high-order correlation and basis-set effects through a facilitating map of the relevant molecular geometry to a graph $\mathcal{G}$, where significant bonded and nonbonded interactions are captured as edges and higher rank simplexes. But since the graphs are defined based on the instantaneous molecular geometry, the connectivities encoded within the graphs (which help capture local and nonlocal correlation and basis-set effects) are likely to change when atoms move, either during dynamics or when potential energy surfaces are computed, presumably to account for quantum nuclear effects. This creates singular hops in potential surfaces when such "graphical hops" occur, and in ref. [44], a new approach is introduced to compute potential surfaces using a weighted set of graphs. Hence, for graphs defined based on (a) a prescribed notion of local spatial interaction or spatial envelope used to define edges surrounding a given node, and (b) a truncation of maximum rank (as in Equation (3)), it is possible to define multiple graphs $\{\mathcal{G}_\alpha\}$ for a given molecular configuration $\mathbf{R}$. While such effects are perhaps most dominant for protonation and deprotonation events, as discussed in ref. [44], it is foreseeable that these could also be critical in large-scale conformational changes.

From this analysis, and the appropriate discussion in ref. [44], it is clear that for a given conformation, multiple implementations of Equation (3) may be considered depending on the graphical representation of the system such that

$$(\mathbf{R},\mathcal{G}_\beta) \mapsto E^{\text{embed}}_{\beta,\mathscr{R}_\beta}(\mathbf{R},\mathcal{G}_\beta) \tag{17}$$

When considering multiple simultaneous graphical representations of a system $\{\mathcal{G}_\beta\}$ with associated maximum ranks $\mathscr{R}_\beta$, the energy of the system becomes a probabilistic sum over the set of energies obtained from the associated set of graphs. The graphs that may be considered are those that represent a certain kind of locality in electronic structure, and thus can be thought of loosely as "valence bond"[128–130] constructs or "diabatic states,"[131–138] which are reminiscent of non-adiabatic electronic structure theory.[139]

Thus, the energy of the system represented by multiple graphs (or a single hypergraph[85,140]) is obtained from a probabilistic sum

$$\langle E(\mathbf{R})\rangle = \sum_\beta \rho_\beta(\mathbf{R}) \cdot E^{\text{embed}}_{\beta,\mathscr{R}_\beta}(\mathbf{R},\mathcal{G}_\beta) \tag{18}$$

where a graph $\mathcal{G}_\beta$ contributes $E^{\text{embed}}_{\beta,\mathscr{R}_\beta}(\mathbf{R})$ to the overall energy based upon the associated probability $\rho_\beta(\mathbf{R})$. In ref. [44], we introduced a graph-theory-based Hamiltonian, where the individual graphs are treated as basis functions, and the eigenstates of such a Hamiltonian are used to compute $\rho_\beta(\mathbf{R})$ in Equation (18). Generalizations to multiple dimensions are currently in progress.

3 | EFFICIENT AND ACCURATE TREATMENT OF GAS- AND CONDENSED-PHASE SYSTEMS USING EQUATION (3)

In Section 3.1 we present a brief discussion of the criteria employed for the generation of the graphical representation, which is comprised of choice of node definition and the spatial envelope used for edge generation. Graphical representations of protonated water clusters, polypeptide chains, and water surfaces with organic adsorbates are used to demonstrate the efficiency and accuracy afforded by Equation (3). Next, the principles outlined in Section 2 are tested by the use of an object-oriented MPI-parallelized C++ module (with added Python functions), which is capable of utilizing multiple electronic structure packages. Here, the simultaneous use of two different electronic structure packages within our calculations, namely Gaussian[92] and Psi4,[93] to compute $E_{\mathcal{R}}^{\mathrm{Embed}}$ and all associated gradients is demonstrated through accurate AIMD and potential surface calculations. To facilitate this discussion, the following nomenclature is used to describe the extrapolations from Equation (3): the extrapolation from method level, 0 to method level, 1 is referred to as $(\mathrm{level}, 1):(\mathrm{level}, 0)$. For example, the notation CCSD:B3LYP would indicate that the full system is treated at the B3LYP level and the embedded graph in Equation (2) is treated at both CCSD (level, 1) and B3LYP (level, 0) levels. First we demonstrate this method to compute post-HF AIMD trajectories (Section 3.2) at DFT cost, and to obtain reduced dimensional potential energy surfaces (Section 3.3), again with post-HF accuracy at DFT cost. Next, hybrid DFT treatment for condensed phase problems is realized at pure DFT functional costs in Section 3.5. Finally, in Section 3.4, large-basis-set DFT accuracy is demonstrated with computational expense corresponding to a much small basis set for both molecular cluster and for condensed-phase calculations.

3.1 | Graph-theoretic representation of molecular and condensed-phase systems

In order to employ Equation (3), the chemical system of interest is first "coarse-grained" into nodes. Following this, the family of two-body interactions between pairs of nodes is included by defining a local spatial envelope that yields graphical edges and, correspondingly, the graphical definition of the system. The coarse-grained nodes are chosen to represent individual chemical units. In the calculations discussed here, and in refs. [39-44], these nodes include amino acid groups, methanol molecules, and water molecules. Critical local interactions between these nodal units are captured through edges and potentially higher order simplexes that form the graphical representation of the system. As noted above, the edges are constructed based upon a spatial envelope, where the envelope may be defined either as node-dependent Cartesian distance cutoffs[39–41,45] or connectivity along a covalent bonding network.[42,43] Such distance- or connectivity-based cutoffs have been traditionally used in many fragmentation-based methods such as FMO[76–78] and EFMO,[79,80] MIM,[81] GMBE,[17,33] HBMI,[82,83] and along with many other methods discussed in refs. [11,13,36]. Once the nodes and edges are defined, higher order *r*-body interactions can be considered within this graph. All simulations shown here only employ two-body (rank-1) interactions.

For the case where covalent bonds are broken in defining the nodes and other higher order simplexes, the resultant dangling valency is saturated by the use of link atoms (principally hydrogen atoms), in a manner consistent with the ONIOM method.[59] (Although it has been shown that other atomic groups can be used based on matching electronegativities,[141] these have not been considered here and will be part of future studies.) As mentioned in the discussion surrounding Equation (11), and discussed in refs. [99,142,143], the forces on these link atoms are transformed back to the full system coordinates using the appropriate Jacobians needed to obtain gradient corrections on the real system.

3.2 | Low-cost post-HF AIMD with Born–Oppenheimer and extended Lagrangian flavors

In refs. [39-43], the effectiveness of Equation (3) for the construction of dynamics trajectories at post-HF accuracy is demonstrated by studying protonated water clusters and polypeptide fragments. As discussed in Section 2.3, there are two classical trajectory methods that have been utilized with our graph-theoretic many-body expansion methodology. Both Born–Oppenheimer and ADMP-pHF trajectories are obtained through the energy and gradient expressions discussed in Section 2.3 (Equations (11), (B1), and (B2) for BOMD and Equations (15) and (16) for ADMP-pHF). The efficiency and accuracy of these trajectories are gauged in refs. [39-43] in multiple ways. First, given the Hamiltonian nature of both simulation methods (Born–Oppenhemier and ADMP-pHF), total energy conservation and the drift in total energy are both probed for a variety of simulations and a variety of systems in refs. [39-43]. It is found that the total energy is conserved to within fractions of kcal/mol for all these studies, and the energy drift is also similarly found to be within the same acceptable range. A second measure of the accuracy of the trajectories is the accuracy of the vibrational density of states obtained through the trajectories. In this case, it was also required to compute classical trajectories at the higher level, namely level, 1, of theory. Because of the cost-prohibitive nature of level, 1 (CCSD), these conclusive studies were conducted only for small systems such as those in Figure 4, where full-system CCSD-based trajectories could be obtained. The vibrational density of states is a spectral (Fourier) representation of the trajectory, which partitions the nuclear velocity contributions as a function of frequency. This description provides the extent of classical inertia that is present in the trajectory at any given frequency and is found from the Fourier transform of the nuclear velocities.[115,116,120,121,144,145] Through the convolution theorem,[145] the Fourier transform of nuclear velocities is also related to the Fourier transform of the velocity–velocity autocorrelation function. The comparison of these spectral features is used to gauge the effectiveness of the low-cost, graph-theory-based many-body expansions used within AIMD as compared to standard AIMD treatments.

The graph-theoretic frag-BOMD trajectories demonstrate great ability to accurately estimate the vibrational density of states and the structural features of the full-system BOMD trajectories at MP2[39,42] and CCSD[41] levels of theory. In Figure 4 we provide one specific illustration of such a comparison with several others in refs. [39-43]. The arrows in Figure 4B demonstrate the quality of the frag-BOMD CCSD:B3LYP

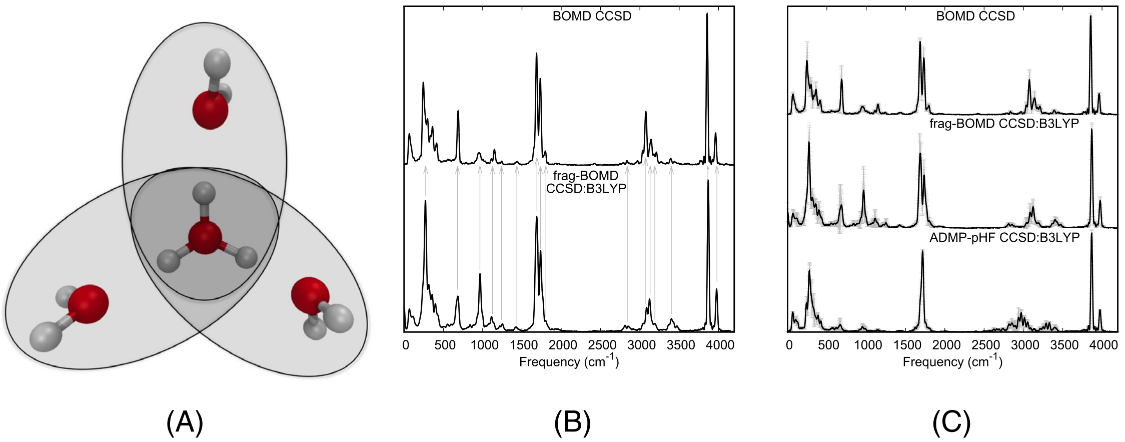

FIGURE 4 Vibrational density of states for the eigen A, protonated water cluster presented at the CCSD level of accuracy. B, Averaged spectra reproduced, with the agreement in frequencies between BOMD CCSD and BOMD CCSD:B3LYP emphasized by arrows. C, Comparison of vibrational density of states from BOMD CCSD, BOMD CCSD:B3LYP, and ADMP-pHF CCSD:B3LYP

trajectory in computing accurate vibrational modes in agreement with full BOMD CCSD. In ref. [41], it was also shown that these trajectories provide structural features in good agreement with full CCSD BOMD trajectories. Similar quality results were also found for frag-BOMD MP2:B3LYP trajectories of water clusters[39] and polypeptide fragments.[42]

An additional, but critical, computational gain from the use of Equation (3) is facilitated through post-HF quality-extended Lagrangian trajectories. But there is a vibrational frequency dependence upon the fictitious electronic parameters in trajectories of extended Lagrangian treatments (μ_{valence} in ADMP and ADMP-pHF). Since the density matrix is propagated along with the nuclear degrees of freedom, there are additional forces (Equation (15)) that deviate from the Born–Oppenheimer surface that can add coherently to the nuclear forces producing a red shift in the vibrational frequency. To remedy this red shift, a uniform scaling factor is introduced in ref. [41] to obtain the observable vibrational frequencies. As demonstrated in ref. [40], this scaling factor is system-independent and has a quadratic dependence on choice of μ_{valence} as seen in Figure 4C, and for the trajectories in Figure 4C the vibrational frequencies are scaled by 1.021, which signifies a 2.1% blue shift to generate good agreement with the BOMD trajectories.[40–43] (A complete study of a range of scaling factors is considered in ref. [41] and obeys the quadratic dependence noted in Figure 5) In summary, graph-based frag-BOMD and ADMP-pHF replicate well the full BOMD results with a significant reduction in computational cost per time step, and because of the quality of constructed trajectories, this graph-based approach offers an attractive method to achieve post-HF accurate BOMD and extended Lagrangian dynamics.

3.3 | Computing reduced-dimensional potential energy surfaces through multiple, simultaneous graphical representations of molecular structure

The molecular graph generation approach is a dynamic phenomenon that, in general, constructs different graphical representations with the change in nuclear configuration. Different graphical representations may occur during potential surface calculations, molecular dynamics, and reaction pathway studies, and may be of critical importance in constructing potential energy surfaces to compute quantum nuclear effects. Examples of such problems include studies involving protonated and hydroxide-rich water clusters, where the excess charge may tend to be delocalized. Similarly, one can also expect these effects to be critical in hydrogen-transfer reactions[146] where quantum nuclear effects[147] (such as tunneling) may have a critical role.

In Figures 6 and 7, we illustrate the dynamic nature of graph-theoretic decomposition as seen during a frag-BOMD simulation performed using the CCSD/6-31+g(d,p):B3LYP/6-31+g(d,p) protocol on a solvated Zundel $[(H_5O_2) \cdot (H_2O)_4]^+$ cation. The singular hops in the potential energy function plotted in Figure 6 are at time steps where the fragmentation topology (or graph description) changes during dynamics. Two of the many molecular graphs seen during this trajectory are shown in Figure 6. The shifted total energies (shown in Figure 7A) and the forces (shown in Figure 7B) are well behaved and do not show significant deviations associated with change in topology despite the fact that there exist numerous jumps in total energy. However, the consequences for the quantum nuclear treatment, which are generally nonlocal, are nontrivial as described in Figure 8. Simple one-dimensional potential surfaces constructed using graphical topologies, found on either side of a hop, provide different surfaces (red and blue) with pronounced difference in quantum nuclear eigen energies.

In ref. [44], we introduced a method to use a set of fragmentation protocols, or a set of graphs, such as those in Figures 7 and 8, together to obtain efficient, high-quality potential surfaces. For a range of potential energy surfaces, represented using oxygen–oxygen donor–acceptor distances in Figure 9B, these high-quality (post-HF) potential surfaces are well reproduced by using multiple graph-theoretic fragmentation protocols, as highlighted in Figure 9C.

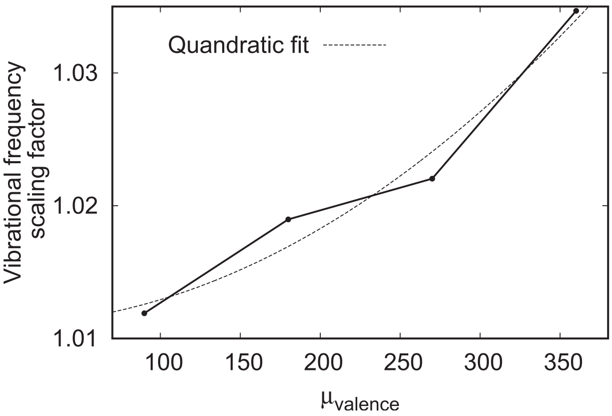

FIGURE 5 Quadratic dependence of the uniform scaling factor on the choice of μ_{valence}. The ADMP-pHF trajectory in Figure 4C is scaled by the scaling factor obtained here, leading to frequencies in agreement with the BOMD trajectories. A full range of scaling factors is considered in ref. [41]

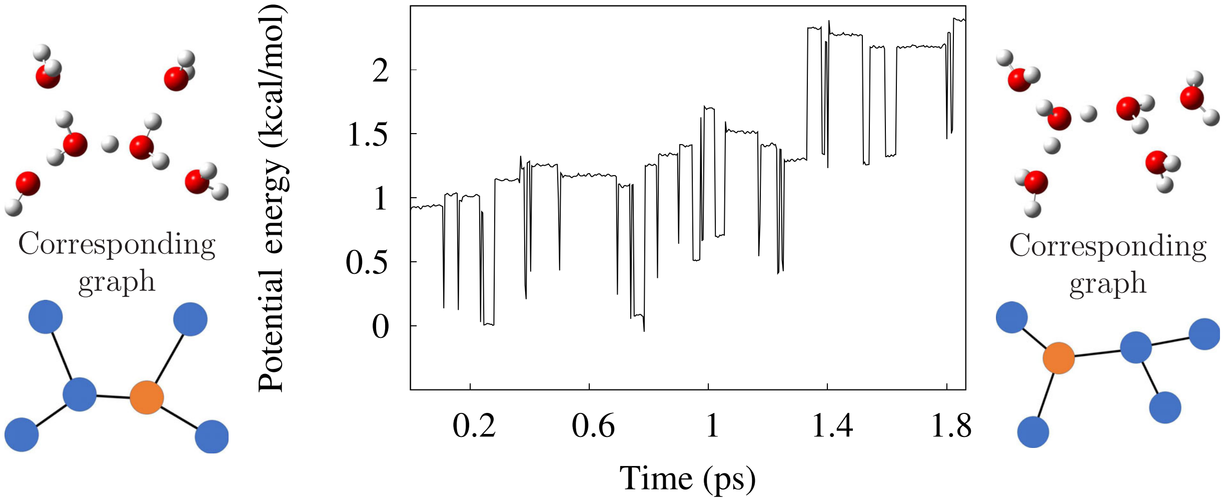

FIGURE 6 Evolution of the electronic potential energy during a frag-BOMD calculation performed at the CCSD/6-31+g(d,p):B3LYP/6-31+g (d,p) level for the solvated Zundel cation. The sudden hops in the figure are due to changes in graph description, two of which are presented on either side with nodes (shown in blue and orange) and edges describing two-body interactions. Note that the graphs on either side are different in their nodal definitions in that a blue node represents H_2O and an orange node indicates H_3O^+. Also see Figure 7

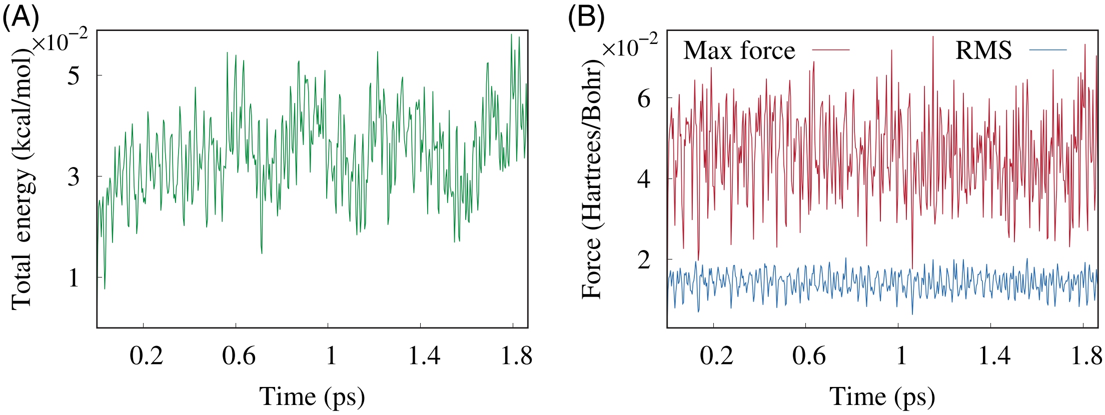

FIGURE 7 As seen in Figure 6, change in graphical topology leads to hops in the total energy during dynamics. However, when, in each case, the total energy is simply shifted by the extent of hop A, the total energy is quite well behaved with deviations of the order of a few hundredth of a kcal/mol. In fact, from A, the RMS deviation in total energy is 0.009 kcal/mol and the associated drift in total energy is 0.019 kcal/mol, both of which are well within acceptable range for AIMD simulations. Similarly, the nuclear forces (maximum and RMS forces) presented in B are also well behaved

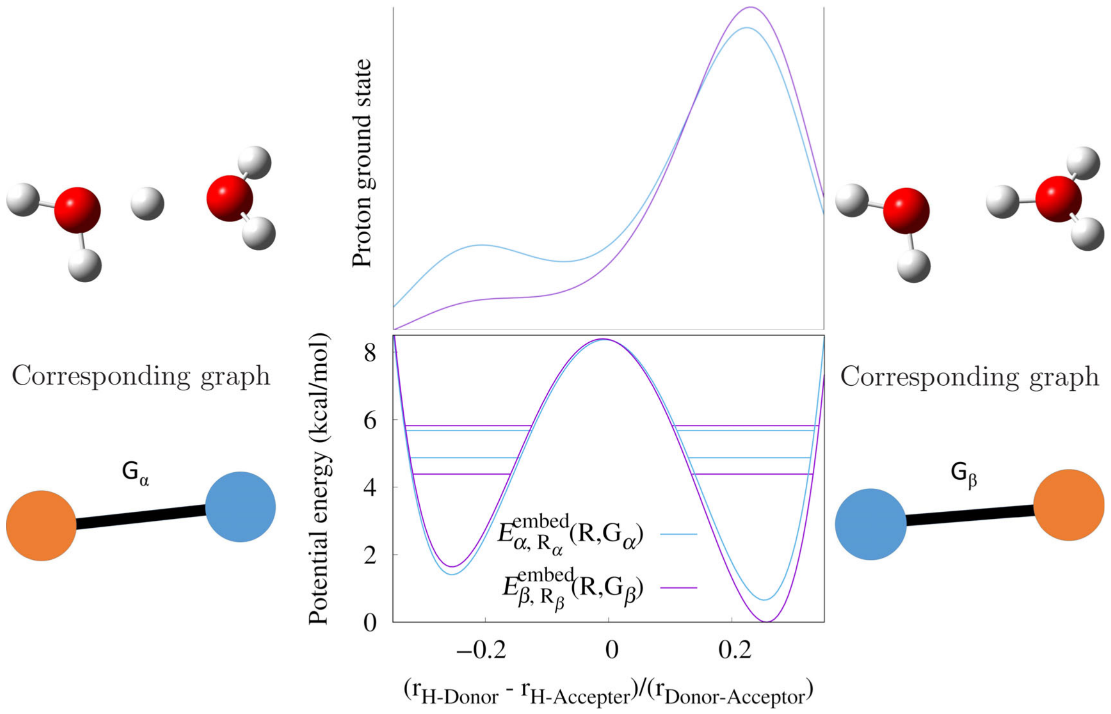

FIGURE 8 Zundel ($H_5O_2^+$) substructure inside a protonated water wire system. The shared proton in the Zundel substructure has a different graph-theory-based potential surface from Equation (3) depending on which water is protonated. (Molecular geometry and graphical representations are on either side, as in Figure 6). This leads to different eigen energies (horizontal lines in the bottom panel) and eigenfunctions (top panel). The methods described in ref. [44] provide a completely general way to treat such problems involving an arbitrary number of graphs, $\left\{\mathcal{G}_\beta, E^{\text{embed}}_{\beta,\mathcal{R}_\beta}(\mathbf{R},\mathcal{G}_\beta); \rho_\beta(\mathbf{R})\right\}$ in Equation (18), in a system-independent manner. A summary of the associated error analysis is presented in Figure 9, where the target level of theory is MP2 (computed at DFT cost)

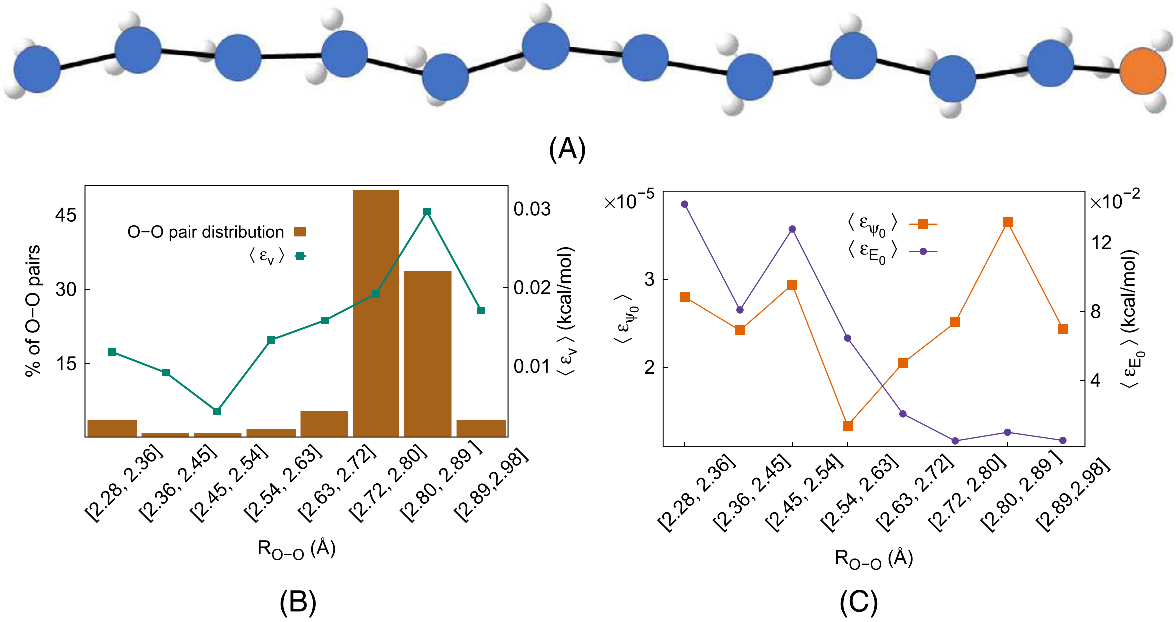

FIGURE 9 Shared proton in the Zundel substructure in a wire–wire (Figure 8) encounters a different potential surface depending upon the geometry of the surrounding water molecules. These potentials are sampled from the distribution shown in B, to conduct an error analysis of the scheme represented by Equation (18) (detailed in ref. [44]) with an MP2 target accuracy. The associated errors in ground state eigen energies and eigenfunctions are shown in C. In all cases, MP2 surfaces are obtained using multiple graphs as highlighted by Equation (18) at DFT cost

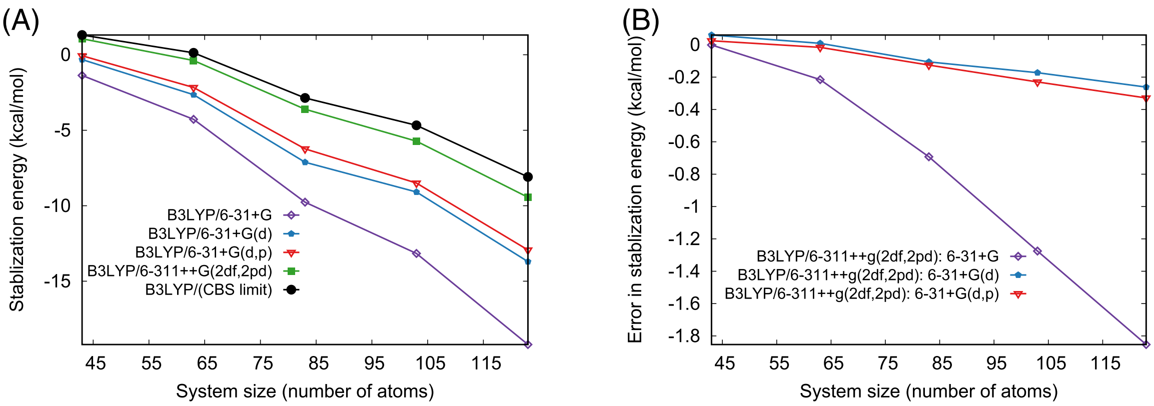

FIGURE 10 Calculated stabilization of polyalanine conformers is highly dependent upon the choice of basis. A, Convergence to the complete basis set limit with the increase of basis set quality. B, Stabilization error of Equation (3) in reference to 6-311++G(2df,2pd) using much smaller basis sets as level, 0. Also see Figure 2B

3.4 | Basis-set-extrapolated AIMD for molecular clusters

Although nonorthogonal atom-centered Gaussian basis sets may have the preferred boundary conditions for studying molecular clusters, these also lead to an expansion of effective functional space of neighboring atoms depending upon interatomic distance and molecular geometry. Because of this effect, as the degrees of freedom evolve in AIMD, the effective linear vector space spanned by the electronic structure changes over the course of the trajectory,[114] which is related to the well-known basis set superposition error (BSSE).[148–150] This error becomes critical when weak interactions are involved, as demonstrated in the conformational energy studies provided in Figure 10A.

Considering the above rationale, in ref. [43], we have used Equation (3) to perturbatively and locally (as dictated by the choice of maximum many-body rank and spatial envelope for graphical edge definition) expand the linear vector basis depicting many-body interactions within molecular systems using the graphical representation discussed above. Similar to the discussions in the previous subsections, the rank-*r* simplexes capture the local treatment of interactions without incurring the global costs of these treatments. Here the many-body terms in Equation (3) act to construct a more extensive local vector space constructed by an extensive basis set, while the full system is considered using only a smaller basis set.

In ref. [43], we demonstrated large, triple-ζ basis DFT accuracy at costs commensurate with smaller basis set treatments, primarily the 6-31 +G(d) basis. The computational scaling gained from this approach was illustrated in Figure 2B, where the costs approach that of the full system at the smaller basis set chosen. Accurate conformational energies were obtained for polypeptide conformers, as illustrated in Figure 10B. We found that the choice of the basis set 6-31+G(d) as level, 0 in Equation (3) is sufficient to produce conformational energies close to the extensive basis 6-311++G(2df,2pd), which approaches the complete basis set limit as seen in Figure 10A. Our ability to achieve conformational energies of substantive basis set quality at lower scaling costs supported the use of these methods for AIMD trajectories using similar conditions as outlined in Section 3.2. These basis-set-extrapolated dynamics calculations demonstrated great agreement with the triple-ζ basis function 6-311++G(2df,2pd) at the computational costs of 6-31+G(d). To illustrate the quality of these trajectories, Figure 11 presents both frag-BOMD and associated extended Lagrangian dynamics using Equation (12) for a polyalanine tetramer strand (Figure 11A). These basis-set-extrapolated trajectories showed reasonable agreement with the large basis trajectory. As was discussed in Section 3.2, the vibrational modes obtained from extrapolated, extended Lagrangian dynamics trajectories were scaled by 1.021. This scaling of frequencies led to similar vibrational modes as in the benchmark trajectory; the basis-set-extrapolated frag-BOMD demonstrated greater agreement with the full system trajectory. With this approach, we were able to achieve basis-set-extrapolated dynamical trajectories for more than 100 atoms at reasonable computational costs.[43]

3.5 | Hybrid functional and large basis set contributions to condensed-phase systems and molecule–surface interactions using the graph-theoretic methodology

Next, when these highly diffuse, nonorthogonal, atom-centered Gaussian basis sets are used in condensed-phase periodic simulations, significant density may be found either (a) outside of the periodic cell, or (b) in regions with limited atomic density, and these lead to instabilities in the SCF procedure.[151,152] Furthermore, from using Ewald summation[153] and from fast multipole-type approximations,[104,105] the long-range portion of the Coulomb repulsion term is computed in reciprocal space (*k*-space) for improved efficiency; the associated integration over *k* cells contributes in a significant manner to the cost of expanding the chosen (diffused and polarized) basis set size. However, as mentioned earlier, larger basis sets are often needed to represent weak interactions[43,150] that are present in catalytic problems.[154–156] We utilized Equation (3) to capture local

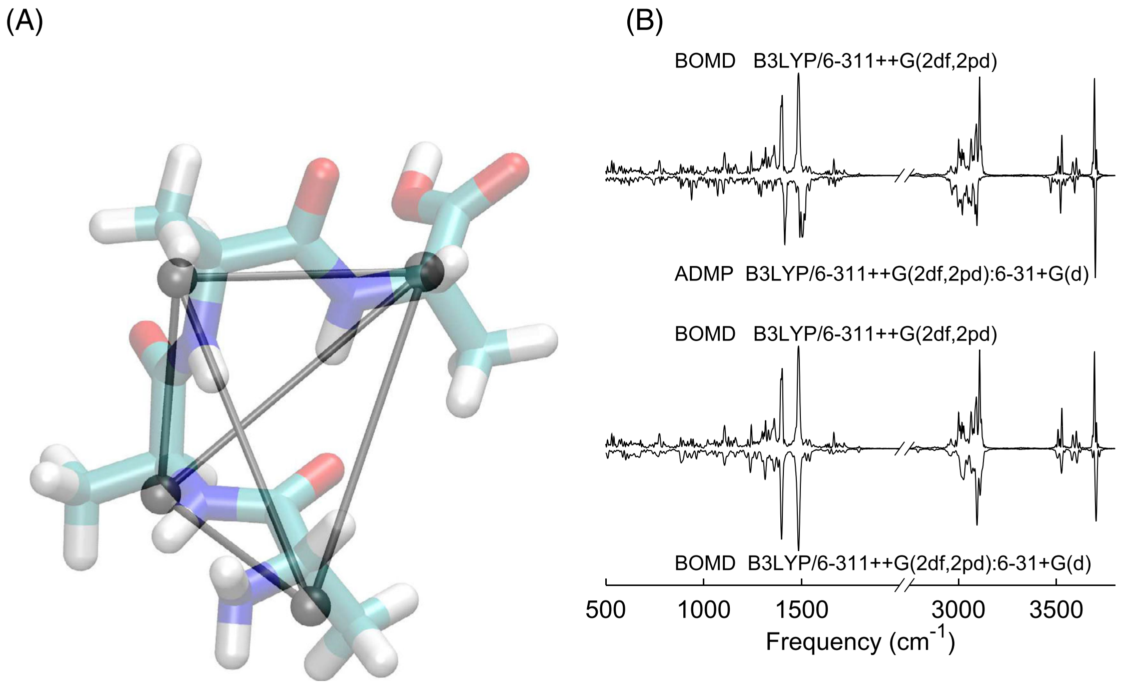

FIGURE 11 Vibrational density of states for a helical conformation of *Ala*$_4$ at B3LYP/6-311++G(2df,2pd) level of accuracy. The full system B3LYP/6-311++G(2df,2pd) trajectories are given as a comparison, the key peaks are well matched from both the use of Equation (3) in constructing basis-set-extrapolated BOMD and through the related extended Lagrangian trajectories arising from Equation (12). The graphical decomposition is shown in A

interactions with an extensive basis set treatment while obtaining the stability and costs of a smaller basis treatment of the periodic wavefunctions. This approach is applied to both bulk systems and films of water, along with organic impurities adsorbed onto films of water. These latter systems act as prototypes for the study of condensed-phase, heterogeneous problems that may be of significance in “on-water” chemical catalysis.[154,156–159] We found great computational cost reduction while obtaining sub kcal/mol accuracy in lattice energy; to illustrate these results, we present this graph-theoretic basis set extrapolation methodology on water films with a methanol adsorbate (Figure 12). We observe a factor of 8 to 14 computational reduction (Figure 12D) when we use this method to achieve accuracy commensurate with that of 6-311++G(df,pd) for a variety of pure and range-separated functionals with similar accuracy independent of the choice of functional. This would seemingly indicate that this basis set treatment is extendable across a variety of different DFT functionals.

Hybrid density functional methods are cost-prohibitive for production-level periodic and condensed phase calculations, such as interactions and adsorption of solutes on surfaces, because of their incorporation of nonlocal HF exchange. Lower-rung DFT methods only utilize pure, gradient-corrected exchange-correlation functional, so they do not incur the same computational costs as the hybrid functionals, but these typically struggle to appropriately capture hydrogen-bonded and other noncovalent interactions.[50,51,160,161] Equation (3) can be used to the reduce complexity by obtaining approximations to higher rung functionals with semi-local treatment of the condensed phase. This is achieved by treating the condensed-phase portion of the calculation, $E^{\mathrm{level},0}$, with periodic pure DFT, while offering perturbative many-body corrections with higher rung DFT using the graph-theoretic interpretation of Equation (3). See Figure 12B. This approach bears similarity to range-specified “screened” exchange functionals,[162–166] but here the nonlocal HF exchange is spatially confined through the locality imposed by the graphical representation. Additionally, here the extent of locality captured by the HF exchange can be tailored adaptively and made spatially dependent through (a) the inclusion of faces and higher order many-body terms as allowed by Equation (3), and (b) the inclusion of a node-dependent spatial envelope that defines edges, and hence the graph. The accuracy of this approach is quantified by the deviation of lattice energy for the adsorbate–water surface structures calculated by Equation (3) compared to (cost-prohibitive) periodic hybrid DFT calculations. Using these studies, we find that hybrid periodic DFT accuracy can be achieved at sub-kcal/mol deviation of lattice energy with computational cost corresponding to pure DFT treatment. This is illustrated in Figure 12E for the adsorption of methanol onto the surface of a water film (Figure 12). For this illustration, a set of pure functionals (BLYP,[167] PBE,[168] and revTPSS[169]) were employed to recover the hybrid functional lattice energies (PBE0[170] and B3LYP[171]). The combinations of PBE0:PBE and B3LYP:revTPSS demonstrated remarkable accuracy, suggesting periodic hybrid DFT calculations can be very well approximated by Equation (3) with the appropriate choice of pure DFT functional. The presence of the impurity on the water surface would suggest a possible need for three-body interactions, which are represented as faces in the graphical representation where the adsorbate interacts with two water molecules

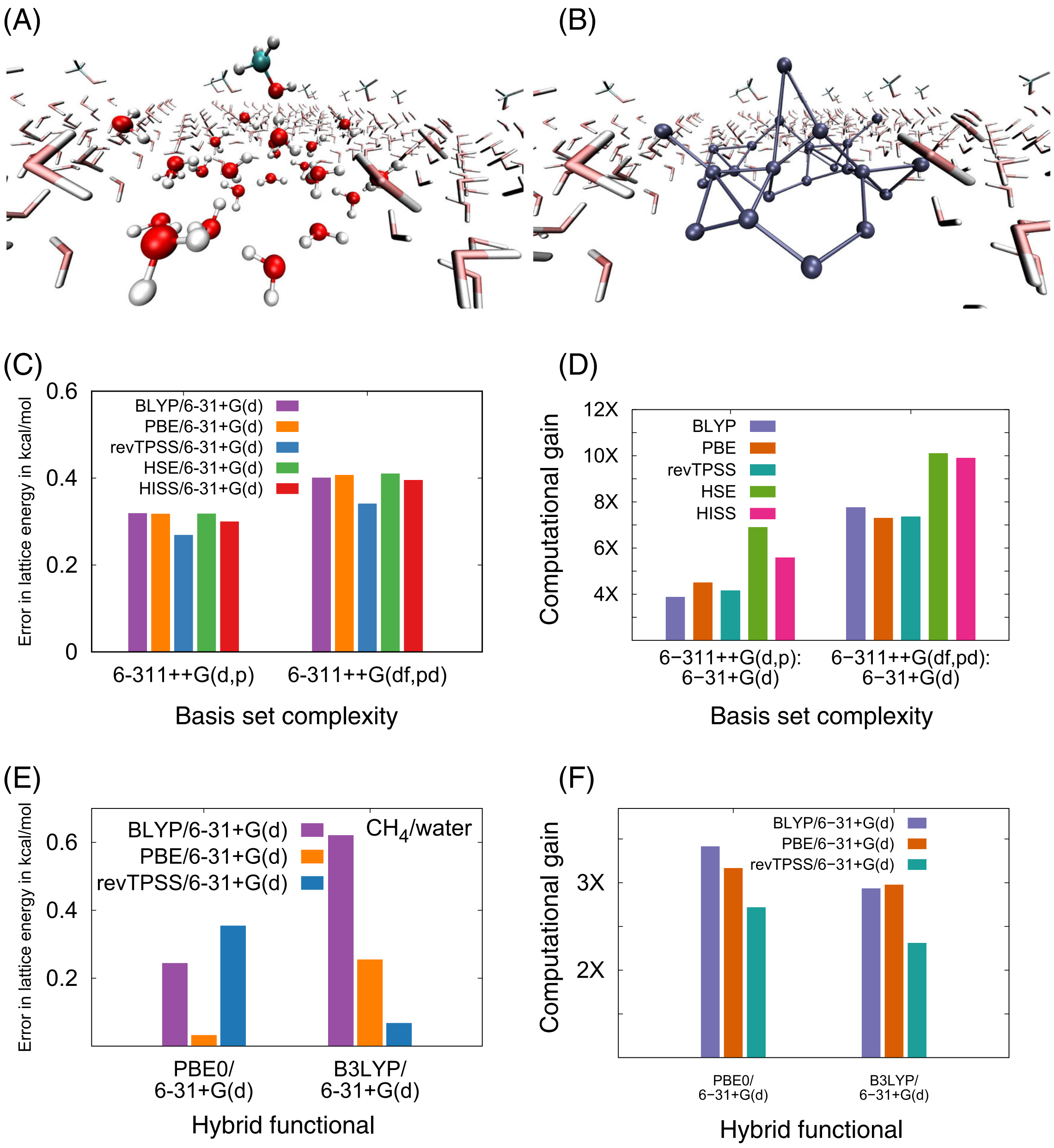

FIGURE 12 Accuracy of the energy expression of Equation (3) evaluated for a film of 24 water molecules with a single methanol molecule on its surface A. Several solutes were considered in ref. [45], but only methanol is highlighted here. The associated graphical decomposition B, provides a detailed many-body treatment of adsorbate/surface interaction. In C, we choose the basis set 6-31+G(d) as the lower level and these are used within the graph-theoretic formalism in Equation (3) to provide results for higher level basis sets presented along the horizontal axis. The graph-theoretic basis set extrapolations consistently shows accuracy to within 0.5 kcal/mol in lattice energy C, while demonstrating a computational gain of roughly a factor of 8 to 10 D. Similarly, E, shows that accurate lattice energies, at hybrid DFT quality, can be obtained using Equation (3) at the lower computational cost shown in F

simultaneously. In the systems we have considered in ref. [45], it was found that these terms offered minimal improvement and these will be further probed in future publications.

4 | CONCLUSION

In this paper, we have discussed a new graph-theoretic embedding procedure with strong connections to many-body theory. The many-body approximation is adaptively computed using rank-*r* graph-theoretic simplexes, each of which is composed from considering a power set of "coarse-grained" local partitions of a chemical system; the higher rank simplexes capture increasingly nonlocal interactions between the "coarse-grained" units that represent nodes in the graphical procedure. The graph-theoretic many-body expression introduced here perturbatively improves upon a full-system calculation conducted at an affordable lower level of theory and basis. The approach introduced here also has strong connections to the ONIOM method and to several molecular fragmentation methods.

This method allows the accuracy of higher quality (post-HF) electronic structure methodologies with lower computational scaling costs (commensurate with DFT). The methods introduced also scale much more favorably with system size as compared to standard post-HF methods. The flexibility of this method was demonstrated through the treatment of protonated water clusters,[39–41,44] condensed-phase water films,[45] interaction of organic adsorbates on the surfaces of water, and the treatment of polypeptide fragments.[42,43] We obtain (a) post-HF accuracy at DFT costs,[39–42,44] (b) periodic hybrid DFT treatment at pure DFT costs,[45] and (c) large triple-ζ basis DFT at small DFT costs.[43,45] The approach was used to efficiently compute post-HF AIMD trajectories within both the Born–Oppenheimer[39,41–43] and extended Lagrangian flavors,[40–43] reduced-dimensional potential energy surfaces for the study of nuclear quantum effects where the surfaces are obtained at post-HF accuracy with DFT cost,[44] and the efficient study of interaction of solutes with surfaces.[45] The cost reduction and accuracy demonstrated render great promise to the methods discussed here.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation grant NSF CHE-1665336 to S. S. I. An XSEDE start-up allocation to S. S. I. is gratefully acknowledged through grant number CHE170081.

AUTHOR CONTRIBUTIONS

Timothy Ricard: Investigation; validation; visualization; writing-original draft; writing-review and editing. **Anup Kumar:** Formal analysis; investigation; methodology; validation. **Srinivasan Iyengar:** Conceptualization; formal analysis; funding acquisition; investigation; methodology; resources; supervision; validation; visualization; writing-original draft; writing-review and editing.

ORCID

Srinivasan S. Iyengar https://orcid.org/0000-0001-6526-2907

REFERENCES

[1] P. Ayala, G. Scuseria, *J. Chem. Phys.* **1999**, *110*, 3660.
[2] M. Schutz, H. Werner, *J. Chem. Phys.* **2001**, *114*, 661.
[3] N. Flocke, R. Bartlett, *J. Chem. Phys.* **2004**, *121*, 10935.
[4] R. A. Distasio Jr., R. P. Steele, Y. M. Rhee, Y. Shao, M. Head-Gordon, *J. Comput. Chem.* **2007**, *28*, 839.
[5] Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, F. Neese, *J. Chem Phys.* **2018**, *148*, 011101.
[6] R. Schutski, J. Zhao, T. M. Henderson, G. E. Scuseria, *J. Chem. Phys.* **2017**, *147*, 184113.
[7] J. Klimes, A. Michaelides, *J. Chem. Phys.* **2012**, *137*, 120901.
[8] P. Mori-Sanchez, A. J. Cohen, W. Yang, *J. Chem. Phys.* **2006**, *125*, 201102.
[9] A. J. Cohen, P. Mori-Sanchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289.
[10] L. W. Chung, H. Hirao, X. Li, K. Morokuma, *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2012**, *2*, 327.
[11] K. Raghavachari, A. Saha, *Chem. Rev.* **2015**, *115*, 5643.
[12] L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, K. Morokuma, *Chem. Rev.* **2015**, *115*, 5678.
[13] M. A. Collins, R. P. A. Bettens, *Chem. Rev.* **2015**, *115*, 5607.
[14] M. S. Gordon, D. G. Fedorov, S. R. Pruitt, L. V. Slipchenko, *Chem. Rev.* **2012**, *112*, 632.
[15] C. Cervinka, G. J. O. Beran, *Chem. Sci.* **2018**, *9*, 4622.
[16] W.-K. Chen, W.-H. Fang, G. Cui, *Phys. Chem. Chem. Phys.* **2019**, *21*, 22695.
[17] K.-Y. Liu, J. M. Herbert, *J. Chem. Theory Comput.* **2020**, *16*, 475.
[18] S. Y. Willow, M. A. Salim, K. S. Kim, S. Hirata, *Sci. Rep.* **2015**, *5*, 14358.
[19] B. Thapa, D. Beckett, K. V. J. Jose, K. Raghavachari, *J. Chem. Theory Comput.* **2018**, *14*, 1383.
[20] C. Konig, O. Christiansen, *J. Chem. Phys.* **2016**, *145*, 064105.
[21] T. M. Sexton, G. S. Tschumper, *Mol. Phys.* **2019**, *117*, 1413.
[22] S. Mande, K. V. J. Jose, *J. Indian Chem. Soc.* **2019**, *96*, 1003.
[23] T. Fang, Y. Li, S. Li, *Wiley Interdiscip. Rev. Comput. Mol. Sci* **2017**, *7*, e1297.
[24] B. W. Hopkins, G. S. Tschumper, *J. Comput. Chem.* **2003**, *24*, 1563.
[25] D. G. Fedorov, K. Kitaura, *J. Phys. Chem. A* **2007**, *111*, 6904.
[26] D. W. Zhang, J. Z. H. Zhang, *J. Chem. Phys.* **2003**, *119*, 3599.
[27] H.-A. Le, H.-J. Tan, J. F. Ouyang, R. P. A. Bettens, *J. Chem. Theory Comput.* **2012**, *8*, 469.
[28] A. J. Varandas, J. N. Murrell, *Faraday Discuss. Chem. Soc.* **1977**, *62*, 92.
[29] K. Nagayoshi, T. Ikeda, K. Kitaura, S. Nagase, *J. Theory Comput. Chem.* **2003**, *02*, 233.
[30] E. E. Dahlke, D. G. Truhlar, *J. Chem. Theory Comput.* **2007**, *3*, 46.
[31] E. E. Dahlke, D. G. Truhlar, *J. Chem. Theory Comput.* **2008**, *4*, 1.
[32] L. D. Jacobson, J. M. Herbert, *J. Chem. Phys.* **2011**, *134*, 094118.

[33] R. M. Richard, J. M. Herbert, *J. Chem. Phys.* **2012**, *137*, 064113.
[34] G. A. Cisneros, K. T. Wikfeldt, L. Ojamäe, J. Lu, Y. Xu, H. Torabifard, A. P. Bartók, G. Csányi, V. Molinero, F. Paesani, *Chem. Rev.* **2016**, *116*, 7501.
[35] Q. Yu, J. M. Bowman, *J. Chem. Phys.* **2017**, *146*, 121102.
[36] J. M. Herbert, *J. Chem. Phys.* **2019**, *151*, 170901.
[37] D. E. Knuth, *The Art of Computer Programming*, Addison-Wesley, Reading, MA **1973**.
[38] U. Manber, *Introduction to Algorithms: A Creative Approach*, Pearson Education, New Jersey **1989**.
[39] J. Li, S. S. Iyengar, *J. Chem. Theory Comput.* **2015**, *11*, 3978.
[40] J. Li, C. Haycraft, S. S. Iyengar, *J. Chem. Theory Comput.* **2016**, *12*, 2493.
[41] C. Haycraft, J. Li, S. S. Iyengar, *J. Chem. Theory Comput.* **2017**, *13*, 1887.
[42] T. C. Ricard, C. Haycraft, S. S. Iyengar, *J. Chem. Theory Comput.* **2018**, *14*, 2852.
[43] T. C. Ricard, S. S. Iyengar, *J. Chem. Theory Comput.* **2018**, *14*, 5535.
[44] A. Kumar, S. S. Iyengar, *J. Chem. Theory Comput.* **2019**, *15*, 5769.
[45] T. C. Ricard, S. S. Iyengar, *J. Chem. Theory Comput.* **2019**.
[46] B. Roos, *Theory and Applications of Computational Chemistry: The First 40 Years*, Elsevier Science, Amsterdam **2005**, p. 725.
[47] K. Raghavachari, J. B. Anderson, *J. Phys. Chem.* **1996**, *100*, 12960.
[48] T. J. Lee, G. E. Scuseria, *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*, Kluwer Academic Publishers, Dordrecht, The Netherlands **1995**, p. 47.
[49] M. Schreiber, M. R. Silva-Junior, S. P. A. Sauer, W. Thiel, *J. Chem. Phys.* **2008**, *128*, 134110.
[50] R. Peverati, D. Truhlar, *Philos. Trans. Royal Soc. A* **2014**, *372*, 10120476.
[51] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289.
[52] R. P. Feynman, J. Hey, R. W. Allen, *Feynman Lectures on Computation*, Addison-Wesley Longman Publishing Co. Inc, Boston, MA **1998**.
[53] M. A. Nielsen, I. L. Chuang, *Quantum computation and quantum information*, Cambridge University Press, Cambridge **2000**.
[54] L. Berman, *Theoretical Comput. Sci.* **1980**, *11*, 71.
[55] R. P. Feynman, *Int. J. Theor. Phys.* **1982**, *21*, 467.
[56] N. DeGregorio, S. S. Iyengar, *J. Chem. Theory Comput.* **2018**, *14*, 30.
[57] N. DeGregorio, S. S. Iyengar, *Faraday Discuss.* **2019**, *221*, 379.
[58] N. DeGregorio, S. S. Iyengar, *J. Chem. Theory Comput.* **2019**, *15*, 2780.
[59] F. Maseras, K. Morokuma, *J. Comput. Chem.* **1995**, *16*, 1170.
[60] T. K. Dey, N. R. Shah, *Comput. Geom.* **1997**, *8*, 267.
[61] W. Guo, A. Wu, X. Xu, *Chem. Phys. Lett.* **2010**, *498*, 203.
[62] N. J. Mayhall, K. Raghavachari, *J. Chem. Theory Comput.* **2011**, *7*, 1336.
[63] N. J. Mayhall, K. Raghavachari, *J. Chem. Theory Comput.* **2012**, *8*, 2669.
[64] S. Li, W. Li, J. Ma, *Accts. Chem. Res.* **2014**, *47*, 2712.
[65] V. Ganesh, R. K. Dongare, P. Balanarayan, S. R. Gadre, *J. Chem. Phys.* **2006**, *125*, 104109.
[66] M. A. Collins, *Phys. Chem. Chem. Phys.* **2012**, *14*, 7744.
[67] M. A. Collins, *J. Phys. Chem. A* **2016**, *120*, 9281.
[68] F. Aurenhammer, *ACM Comput. Surv.* **1991**, *23*, 345.
[69] A. Okabe, B. Boots, K. Sugihara, S. N. Chiu, *Spatial tessellations: concepts and applications of Voronoi diagrams*, Vol. 501, John Wiley & Sons, Hoboken, NJ **2009**.
[70] T. M. Coffey, R. E. Wyatt, W. C. Schieve, *Phys. Rev. Lett.* **2011**, *107*, 230403.
[71] A. Bowyer, *Comp J.* **1981**, *24*, 162.
[72] D. F. Watson, *Comp. J.* **1981**, *24*, 167.
[73] A. Okabe, B. Boots, K. Sugihara, S. N. Chiu, *Spatial Tessellations—Concepts and applications of Voronoi diagrams*, John Wiley and Sons, Hoboken, NJ **2000**.
[74] S. Hert, M. Seel, *CGAL User and Reference Manual*, 4th ed., CGAL Editorial Board, **2018**.
[75] G. Farin, *Comput. Aided Geom. Des.* **1990**, *7*, 281.
[76] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, M. Uebayasi, *Chem. Phys. Lett.* **1999**, *313*, 701.
[77] M. Gordon, J. Mullin, S. Pruitt, L. Roskop, L. Slipchenko, J. Boatz, *J. Phys. Chem. B* **2009**, *113*, 9646.
[78] Y. Komeiji, Y. Mochizuki, T. Nakano, *Chem. Phys. Lett.* **2010**, *484*, 380.
[79] S. R. Pruitt, C. Steinmann, J. H. Jensen, M. S. Gordon, *J. Chem. Theory Comput.* **2013**, *9*, 2235.
[80] C. Steinmann, D. G. Fedorov, J. H. Jensen, *J. Phys. Chem. A* **2010**, *114*, 8705.
[81] A. Saha, K. Raghavachari, *J. Chem. Theory Comput.* **2015**, *11*, 2012.
[82] G. J. O. Beran, *Angew. Chem. Int. Ed.* **2015**, *54*, 396.
[83] W. Sontising, G. J. O. Beran, *Phys. Rev. Mater.* **2019**, *3*, 095002.
[84] G. J. O. Beran, *Chem. Rev.* **2016**, *116*, 5567.
[85] K. R. Mecke, H. Wagner, *J. Stat. Phys.* **1991**, *64*, 843.
[86] S. Hert, M. Seel, *dD convex hulls and delaunay triangulations. CGAL-3.2 User and Reference Manual*, Vol. 5 **2006**, p. 2.
[87] G. Berry, G. Boudol, *Theor. Comput. Sci.* **1992**, *96*, 217.
[88] J.-P. Banǎtre, P. Fradet, D. Le Métayer, *Gamma and the Chemical Reaction Model: Fifteen Years After*, Berlin, Vol. 17 **2001**.
[89] A. Di Pierro, C. Hankin, H. Wiklicky, On a Probabilistic Chemical Abstract Machine and the Expressiveness of Linda Languages, in *Formal Methods for Components and Objects* (Eds: F. S. de Boer, M. M. Bonsangue, S. Graf, W.-P. de Roever), Springer, Berlin, Heidelberg **2006**, p. 388.
[90] N. Mardirossian, M. Head-Gordon, *Mol. Phys.* **2017**, *115*, 2315.
[91] J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, *Phys. Rev. Lett.* **2003**, *91*, 146401.
[92] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci,

C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P. M. W. Gill, B. G. Johnson, W. Chen, M. W. Wong, J. L. Andres, M. Head-Gordon, E. S. Replogle, J. A. Pople, *Gaussian 98*, Gaussian, Inc., Pittsburgh, PA **1998**.

[93] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince III., E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, V. Prakash, H. F. Schaefer III., K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, C. D. Sherrill, *J. Chem. Theor. Comput* **2017**, *13*, 3185.

[94] F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 73.

[95] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, *J. Phys. Condens. Matter* **2009**, *21*, 395502.

[96] Neale MC, Hunter MD, Pritikin JN, Zahery M, Brick TR, Kirkpatrick RM, Estabrook R, Bates TC, Maes HH, Boker SM. *Psychometrika* **2016**, *81*, 535.

[97] M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus, W. de Jong, *Comput. Phys. Commun.* **2010**, *181*, 1477.

[98] G. Kresse, J. Furthmüller, *Phys. Rev. B* **1996**, *54*, 169.

[99] N. Rega, S. S. Iyengar, G. A. Voth, H. B. Schlegel, T. Vreven, M. J. Frisch, *J. Phys. Chem. B* **2004**, *108*, 4210.

[100] F. Neese, F. Wennmohs, A. Hansen, U. Becker, *Chem. Phys.* **2009**, *356*, 98.

[101] J. Kussmann, M. Beer, C. Ochsenfeld, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 614.

[102] G. E. Scuseria, *J. Phys. Chem. A* **1999**, *103*, 4782.

[103] S. Goedecker, *Rev. Mod. Phys.* **1999**, *71*, 1085.

[104] C. A. White, M. Head-Gordon, *J. Chem. Phys.* **1994**, *101*, 6593.

[105] M. C. Strain, G. E. Scuseria, M. J. Frisch, *Science* **1996**, *271*, 51.

[106] H. C. Andersen, *J. Chem. Phys.* **1980**, *72*, 2384.

[107] M. Parrinello, A. Rahman, *Phys. Rev. Lett.* **1980**, *45*, 1196.

[108] R. Car, M. Parrinello, *Phys. Rev. Lett.* **1985**, *55*, 2471.

[109] H. B. Schlegel, J. M. Millam, S. S. Iyengar, G. A. Voth, A. D. Daniels, G. E. Scuseria, M. J. Frisch, *J. Chem. Phys.* **2001**, *114*, 9758.

[110] S. S. Iyengar, H. B. Schlegel, J. M. Millam, G. A. Voth, G. E. Scuseria, M. J. Frisch, *J. Chem. Phys.* **2001**, *115*, 10291.

[111] H. B. Schlegel, S. S. Iyengar, X. Li, J. M. Millam, G. A. Voth, G. E. Scuseria, M. J. Frisch, *J. Chem. Phys.* **2002**, *117*, 8694.

[112] S. S. Iyengar, H. B. Schlegel, G. A. Voth, J. M. Millam, G. E. Scuseria, M. J. Frisch, *Israel J. Chem.* **2002**, *42*, 191.

[113] W. C. Swope, H. C. Andersen, P. H. Berens, K. R. Wilson, *J. Chem. Phys.* **1982**, *76*, 637.

[114] S. S. Iyengar, M. J. Frisch, *J. Chem. Phys.* **2004**, *121*, 5061.

[115] S. S. Iyengar, M. K. Petersen, T. J. F. Day, C. J. Burnham, V. E. Teige, G. A. Voth, *J. Chem. Phys.* **2005**, *123*, 084309.

[116] S. S. Iyengar, *J. Chem. Phys.* **2007**, *126*, 216101.

[117] S. S. Iyengar, T. J. F. Day, G. A. Voth, *Int. J. Mass Spectrom.* **2005**, *241*, 197.

[118] S. S. Iyengar, *J. Chem. Phys.* **2005**, *123*, 084310.

[119] X. Li, V. E. Teige, S. S. Iyengar, *J. Phys. Chem. A* **2007**, *111*, 4815.

[120] X. Li, D. T. Moore, S. S. Iyengar, *J. Chem. Phys.* **2008**, *128*, 184308.

[121] X. Li, J. Oomens, J. R. Eyler, D. T. Moore, S. S. Iyengar, *J. Chem. Phys.* **2010**, *132*, 244301.

[122] P. Phatak, I. Sumner, S. S. Iyengar, *J. Phys. Chem. B* **2012**, *116*, 10145.

[123] P. Phatak, J. Venderley, J. Debrota, J. Li, S. S. Iyengar, *J. Phys. Chem. B* **2015**, *119*, 9532.

[124] D. Vimal, A. B. Pacheco, S. S. Iyengar, P. S. Stevens, *J. Phys. Chem. A* **2008**, *112*, 7227.

[125] A. B. Pacheco, S. M. Dietrick, P. S. Stevens, S. S. Iyengar, *J. Phys. Chem. A* **2012**, *116*, 4108.

[126] S. M. Dietrick, A. B. Pacheco, P. Phatak, P. S. Stevens, S. S. Iyengar, *J. Phys. Chem. A* **2012**, *116*, 399.

[127] R. McWeeny, *Rev. Mod. Phys.* **1960**, *32*, 335.

[128] A. Warshel, R. M. Weiss, *J. Am. Chem. Soc.* **1980**, *102*, 6218.

[129] Y. T. Chang, W. H. Miller, *J. Phys. Chem.* **1990**, *94*, 5884.

[130] T. J. Day, A. V. Soudackov, M. Čuma, U. W. Schmitt, G. A. Voth, *J. Chem. Phys.* **2002**, *117*, 5839.

[131] M. Baer, *Beyond Born-Oppenheimer: Conical Intersections and Electronic Nonadiabatic Coupling Terms*, Wiley, New York **2006**.

[132] G. Hanna, R. Kapral, *Acc. Chem. Res.* **2006**, *39*, 21.

[133] G. A. Worth, L. S. Cederbaum, *Annu. Rev. Phys. Chem.* **2004**, *55*, 127.

[134] A. W. Jasper, C. Zhu, S. Nangia, D. G. Truhlar, *Faraday Discuss.* **2004**, *127*, 1.

[135] B. K. Kendrik, C. A. Mead, D. G. Truhlar, *Chem. Phys.* **2002**, *277*, 31.

[136] A. Kuppermann, in *The Geometric Phase in Reaction Dynamics. In Dynamics of Molecules and Chemical Reactions* (Eds: R. E. Wyatt, J. Z. H. Zhang), Marcel Dekker Inc., New York **1996**, p. 411.

[137] D. R. Yarkony, *Rev. Mod. Phys.* **1996**, *68*, 985.

[138] S. Matsika, D. R. Yarkony, *J. Am. Chem. Soc.* **2003**, *125*, 10672.

[139] D. Coker, Proceedings of the NATO Advanced Study Institute on New Perspectives in Computer Simulation in Chemical Physics, 1993, 315.

[140] Adams, C. C.; Franzosa, R. D. *Introduction to Topology: Pure and Applied*. Pearson Prentice Hall, Upper Saddle River, NJ **2008**.

[141] I. Sumner, S. S. Iyengar, *J. Chem. Phys.* **2008**, *129*, 054109.

[142] D. Bakowies, W. Thiel, *J. Phys. Chem.* **1996**, *100*, 10580.

[143] S. Dapprich, I. Komaromi, K. S. Byun, K. Morokuma, M. J. Frisch, *J. Mol. Struct. Theochem.* **1999**, *462*, 1.

[144] S. M. Dietrick, S. S. Iyengar, *J. Chem. Theory Comp.* **2012**, *8*, 4876.

[145] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, New York **1992**.

[146] N. Isaacs, *Physical Organic Chemistry*, Longman Scientific and Technical, Essex **1995**.

[147] R. Sheridan, in *Reviews of Reactive Intermediate Chemistry* (Eds: M. Platz, R. A. Moss, J. Maitland Jones), Wiley-Interscience, Hoboken, NJ **2007**.

[148] E. Clementi, *J. Chem. Phys.* **1967**, *46*, 3851.
[149] S. F. Boys, F. Bernardi, *Mol. Phys.* **1970**, *19*, 553.
[150] F. B. Van Duijneveldt, J. G. van Duijneveldt-van de Rijdt, J. H. van Lenthe, *Chem. Rev.* **1994**, *94*, 1873.
[151] K. N. Kudin, G. E. Scuseria, *Phys. Rev. B* **2000**, *61*, 16440.
[152] S. Suhai, P. Bagus, J. Ladik, *Chem. Phys.* **1982**, *68*, 467.
[153] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids*, Oxford Science Publications, New York **1987**.
[154] S. Narayan, J. Muldoon, M. Finn, V. Fokin, H. Kolb, K. Sharpless, *Angew. Chem.* **2005**, *44*, 3275.
[155] R. Breslow, U. Maitra, D. Rideout, *Tetrahedron Lett.* **1983**, *24*, 1901.
[156] Y. Jung, R. A. Marcus, *J. Am. Chem. Soc.* **2007**, *129*, 5492.
[157] Y. Jung, R. A. Marcus, *J. Phys. Condens. Matter* **2010**, *22*, 284117.
[158] R. Breslow, *Acc. Chem. Res.* **1991**, *24*, 159.
[159] N. Sahota, D. I. AbuSalim, M. L. Wang, C. J. Brown, Z. Zhang, T. J. El-Baba, S. P. Cook, D. E. Clemmer, *Chem. Sci.* **2019**, *10*, 4822.
[160] M. J. Gillan, D. Alfè, A. Michaelides, *J. Chem. Phys.* **2016**, *144*, 130901.
[161] N. Konstantin, G. E. S. Kudin, R. L. Martin, *Phys. Rev. Lett.* **2002**, *89*, 266402.
[162] J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, J. G. Ángyán, *J. Chem. Phys.* **2006**, *124*, 154709.
[163] B. G. Janesko, T. M. Henderson, G. E. Scuseria, *Phys. Chem. Chem. Phys.* **2009**, *11*, 443.
[164] M. Guidon, J. Hutter, J. VandeVondele, *J. Chem. Theory Comput.* **2009**, *5*, 3010.
[165] M. Lucero, T. Henderson, G. Scuseria, *J. Phys. Condens. Matter.* **2012**, *24*, 145504.
[166] R. Peverati, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2012**, *14*, 16187.
[167] A. D. Becke, *J. Chem. Phys.* **1988**, *38*, 3098.
[168] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
[169] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, J. Sun, *Phys. Rev. Lett.* **2009**, *103*, 026403.
[170] C. Adamo, V. Barone, *J. Chem. Phys.* **1999**, *110*, 6158.
[171] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

How to cite this article: Ricard TC, Kumar A, Iyengar SS. Embedded, graph-theoretically defined many-body approximations for wavefunction-in-DFT and DFT-in-DFT: Applications to gas- and condensed-phase ab initio molecular dynamics, and potential surfaces for quantum nuclear effects. *Int J Quantum Chem*. 2020;1–20. https://doi.org/10.1002/qua.26244

APPENDIX A.: EXPLICIT FORM OF EQUATION (1) FOR A FEW VALUES OF $\mathcal{R}$

The precise fragment energies that arise in the treatment may be understood by explicitly writing out Equations (1), (3), or (7) for a few values of $\mathcal{R}$. For $\mathcal{R} = 1$

$$E^{\text{Embed}}_{\mathcal{R}=1} = E^{\text{level},0} + \sum_{\alpha\in\text{edges}} \Delta E(\alpha,1) - \sum_{\alpha\in\text{nodes}} \Delta E(\alpha,0)\left[p_\alpha^{0,1} - p_\alpha^{0,0}\right] \tag{A1}$$

where $p_\alpha^{0,1}$ is the number of times the node α (one-body term) appears in all edges (or two-body interactions), and $p_\alpha^{0,0}$ is the number of times the node α appears in all nodes and hence $p_\alpha^{0,0} = 1$. Using Equation (4), we may explicitly write Equation (A1) as

$$\begin{aligned} E^{\text{Embed}}_{\mathcal{R}=1} = E^{\text{level},0} &+ \sum_{\alpha\in\text{edges}} \left[E^{\text{level},1}(\alpha,1) - E^{\text{level},0}(\alpha,1)\right] \\ &- \sum_{\alpha\in\text{nodes}} \left[E^{\text{level},1}(\alpha,0) - E^{\text{level},0}(\alpha,0)\right]\left(p_\alpha^{0,1} - 1\right) \end{aligned} \tag{A2}$$

$$\begin{aligned} &= E^{\text{level},0} \\ &+ \left\{ \sum_{\alpha\in\text{nodes}} E^{\text{level},1}(\alpha,0) + \left[\sum_{\alpha\in\text{edges}} E^{\text{level},1}(\alpha,1) - \sum_{\alpha\in\text{nodes}} p_\alpha^{0,1} E^{\text{level},1}(\alpha,0) \right] \right\} \\ &- \left\{ \sum_{\alpha\in\text{nodes}} E^{\text{level},0}(\alpha,0) + \left[\sum_{\alpha\in\text{edges}} E^{\text{level},0}(\alpha,1) - \sum_{\alpha\in\text{nodes}} p_\alpha^{0,1} E^{\text{level},0}(\alpha,0) \right] \right\} \end{aligned} \tag{A3}$$

In Equation (A3), we have clubbed terms belonging to level, 1 and level, 0 separately. The terms inside each of the curly brackets $\{\cdots\}$ are the level, 1 (or level, 0) two-body energies. This may also be clear upon comparison with Equation (1). Equation (A3) also explicates the actual fragments that need to be computed in this formalism. But, as noted earlier, Equation (3) includes many-body contributions to arbitrary orders. Thus, to include three-body interactions, one may use $\mathscr{R}$ = 2 in Equation (3), to obtain

$$\begin{aligned}
E_{\mathscr{R}=2}^{\text{Embed}} &= E^{\text{level},0} + \sum_{r=0}^{\mathscr{R}=2} (-1)^r \left\{ \sum_{\alpha}^{r-\text{rank}} \Delta E(\alpha,r) \left[\sum_{m=r}^{\mathscr{R}=2} (-1)^m p_\alpha^{r,m} \right] \right\} \\
&= E^{\text{level},0} \\
&+ \left\{ \sum_{\alpha\in\text{nodes}} E^{\text{level},1}(\alpha,0) + \left[\sum_{\alpha\in\text{edges}} E^{\text{level},1}(\alpha,1) - \sum_{\alpha\in\text{nodes}} p_\alpha^{0,1} E^{\text{level},1}(\alpha,0) \right] \right. \\
&\quad \left. + \left[\sum_{\alpha\in\text{faces}} E^{\text{level},1}(\alpha,2) - \sum_{\alpha\in\text{edges}} p_\alpha^{1,2} E^{\text{level},1}(\alpha,1) + \sum_{\alpha\in\text{nodes}} p_\alpha^{0,2} E^{\text{level},1}(\alpha,0) \right] \right\} \\
&- \left\{ \sum_{\alpha\in\text{nodes}} E^{\text{level},0}(\alpha,0) + \left[\sum_{\alpha\in\text{edges}} E^{\text{level},0}(\alpha,1) - \sum_{\alpha\in\text{nodes}} p_\alpha^{0,1} E^{\text{level},0}(\alpha,0) \right] \right. \\
&\quad \left. + \left[\sum_{\alpha\in\text{faces}} E^{\text{level},0}(\alpha,2) - \sum_{\alpha\in\text{edges}} p_\alpha^{1,2} E^{\text{level},0}(\alpha,1) + \sum_{\alpha\in\text{nodes}} p_\alpha^{0,2} E^{\text{level},0}(\alpha,0) \right] \right\} \\
&= E^{\text{level},0} + \left\{ E_{1-\text{body}}^{\text{level},1} + E_{2-\text{body,corr.}}^{\text{level},1} + E_{3-\text{body,corr.}}^{\text{level},1} \right\} \\
&- \left\{ E_{1-\text{body}}^{\text{level},0} + E_{2-\text{body,corr.}}^{\text{level},0} + E_{3-\text{body,corr.}}^{\text{level},0} \right\}
\end{aligned} \tag{A4}$$

It is clear from Equations (A3) and (A4) that Equations (1), (3), and (7) provide an ONIOM-type extrapolation approximation ("*high*" minus "*low*" or level, 1 minus level, 0) where the "*high*" and the "*low*" levels are themselves many-body approximations.

APPENDIX B.: EXPLICIT FORM OF EQUATION (11) FOR A FEW VALUES OF $\mathscr{R}$

This section may be considered in parallel with Section A and provides explicit gradient expressions for two- and three-body-based studies. Using $\mathscr{R}$ = 1 in Equation (11)

$$\begin{aligned}
\frac{\partial E_{\mathscr{R}=1}^{\text{Embed}}(\mathbf{R})}{\partial \mathbf{R}} = \frac{\partial E^{\text{level},0}(\mathbf{R})}{\partial \mathbf{R}} &+ \sum_{\alpha\in edges} \left(\frac{\partial E^{\text{level},1}(\alpha,1)}{\partial \mathbf{R}_{\alpha,1}} - \frac{\partial E^{\text{level},0}(\alpha,1)}{\partial \mathbf{R}_{\alpha,1}} \right) \frac{\partial \mathbf{R}_{\alpha,1}}{\partial \mathbf{R}} \\
&- \sum_{\alpha\in nodes} \left(\frac{\partial E^{\text{level},1}(\alpha,0)}{\partial \mathbf{R}_{\alpha,0}} - \frac{\partial E^{\text{level},0}(\alpha,0)}{\partial \mathbf{R}_{\alpha,0}} \right) \frac{\partial \mathbf{R}_{\alpha,0}}{\partial \mathbf{R}} \left[p_\alpha^{0,1} - 1 \right]
\end{aligned} \tag{B1}$$

The similarity in form between Equations (B1) and (A1) may be clear. These gradients are attenuated by a Jacobian factor, $\frac{\partial \mathbf{R}_{\alpha_r,r}}{\partial \mathbf{R}}$ as seen in Equation (A4). Similar to Equation (A3), Equation (B1) may also be presented in many-body form. For $\mathscr{R}$ = 2 in Equation (11), there will be as many terms as in Equation (A4) and these may be explicitly written out by differentiating the energy corresponding to each simplex, using the real-space variables corresponding to that simplex:

$$\begin{aligned}
\frac{\partial E_{\mathscr{R}=2}^{\text{Embed}}(\mathbf{R})}{\partial \mathbf{R}} = \frac{\partial E^{\text{level},0}(\mathbf{R})}{\partial \mathbf{R}} &+ \sum_{\alpha\in triangle} \left(\frac{\partial E^{\text{level},1}(\alpha,2)}{\partial \mathbf{R}_{\alpha,2}} - \frac{\partial E^{\text{level},0}(\alpha,2)}{\partial \mathbf{R}_{\alpha,2}} \right) \frac{\partial \mathbf{R}_{\alpha,2}}{\partial \mathbf{R}} \\
&- \sum_{\alpha\in edges} \left(\frac{\partial E^{\text{level},1}(\alpha,1)}{\partial \mathbf{R}_{\alpha,1}} - \frac{\partial E^{\text{level},0}(\alpha,1)}{\partial \mathbf{R}_{\alpha,1}} \right) \frac{\partial \mathbf{R}_{\alpha,1}}{\partial \mathbf{R}} \left[p_\alpha^{1,2} - 1 \right] \\
&+ \sum_{\alpha\in nodes} \left(\frac{\partial E^{\text{level},1}(\alpha,0)}{\partial \mathbf{R}_{\alpha,0}} - \frac{\partial E^{\text{level},0}(\alpha,0)}{\partial \mathbf{R}_{\alpha,0}} \right) \frac{\partial \mathbf{R}_{\alpha,0}}{\partial \mathbf{R}} \left[p_\alpha^{0,2} - p_\alpha^{0,1} + 1 \right]
\end{aligned} \tag{B2}$$

As can be seen, as the maximum simplex rank $\mathscr{R}$ is increased, the gradient contributions of the lower rank simplexes are additive with the appropriate Jacobian and over-counting $p_\alpha^{r,m}$ corrections. When considering ADMP-based dynamics, the propagation of the full system's density matrix additionally generates additional gradients shown in Equations (15) and (16), which augment the gradients in Equations (B1) and (B2).