Hybrid extended Lagrangian, post-Hartree-Fock
Born-Oppenheimer ab initio molecular dynamics using fragment-based electronic structure

Junjie Li, Cody Haycraft and Srinivasan S. Iyengar *

Department of Chemistry and Department of Physics,
Indiana University, 800 E. Kirkwood Ave, Bloomington, IN-47405

(Dated: May 6, 2016)

* Email: iyengar@indiana.edu
Abstract

We present a hybrid ab initio molecular dynamics scheme that includes both DFT and Hartree-Fock based extended Lagrangian and converged post-Hartree-Fock Born-Oppenheimer components, combined within the framework of molecular fragmentation based electronic structure. The specific fragmentation algorithm used here is derived from ONIOM, but includes multiple, overlapping “model” systems. The interaction between the various overlapping model systems is approximated by invoking the principle of inclusion-exclusion at the chosen higher level of theory, and within a “real” calculation performed at the chosen lower level of theory. Furthermore, here, the lower level electronic structure of the full system is propagated through an extended Lagrangian formalism, whereas the fragments, treated using post-Hartree-Fock electronic structure theories, are computed using the normal converged Born-Oppenheimer treatment. This conservative dynamical approach largely reduces the computational cost to approximate on-the-fly dynamics using post-Hartree-Fock electronic structure techniques, and can be very beneficial for large systems where SCF convergence may be challenging and time-consuming. Benchmarks are provided for medium-sized protonated water clusters, $H_9O_4^+$ and $H_{13}O_6^+$, and poly-peptide fragments including a proline tripeptide fragment, and alanine decamer. Structural features are in excellent agreement between the hybrid approach using MP2:B3LYP fragment based electronic structure and BOMD using MP2 for the full system. Vibrational properties derived from dynamical correlation functions do show a small red-shift for the extended Lagrangian treatments, especially at higher frequencies. Strategies are discussed to improve this red-shift. The computational methodology works in parallel using both MPI and OpenMP, and shows good scaling with processor number. The timing benchmarks are provided for the alanine decamer. A powerful feature of the computational implementation is the fact that it is completely decoupled from the electronic structure package being employed and thus allows for an integrated approach that may include several different packages. These computational aspects will be further probed in future publications.
I. INTRODUCTION

Since the seminal work of Karplus\textsuperscript{1} and Leforestier\textsuperscript{2}, and the extended Lagrangian generalization by Car and Parrinello\textsuperscript{3}, \textit{ab initio} molecular dynamics (AIMD) has become a central tool for studying reactive events, vibrational properties beyond the harmonic approximation and for biochemical modeling when combined with hybrid techniques. Here, electronic structure and gradients are simultaneously computed in-step with nuclear motion. Due to this, AIMD is in general much more resource intensive as compared to both classical, force field based, molecular dynamics methods as well as commonly used electronic structure applications such as geometry optimization and frequency calculations. This limitation has thus far restricted AIMD mostly to DFT based application for medium sized systems. Improvements in electronic structure methodology, certainly, have critical impact on AIMD and one of the ways to reduce the cost of the on-the-fly electronic structure calculations is through the utilization of the recently popularized fragment-based electronic structures schemes\textsuperscript{4–25} where a molecular system is divided into multiple regions and the total energy of the system is assembled from the energy of individual pieces. Detailed discussion of these methods can be found in several recently published review articles\textsuperscript{26–30}.

To explore the application of fragment-based electronic structure in AIMD and potentially \textit{ab initio} quantum nuclear dynamics\textsuperscript{31}, we recently\textsuperscript{32} (a) considered a generalization to the ONIOM\textsuperscript{5,13,33–41} energy extrapolation scheme using the set-theoretic principle of inclusion and exclusion\textsuperscript{42} to treat multiple overlapping model systems\textsuperscript{32} (abbreviated in this publication as the PIE-ONIOM energy expression), and (b) applied this method to obtain accurate potential energy surfaces, and to compute \textit{ab initio} dynamical trajectories with correlation functions that agree well with higher levels of theory such as MP2 and Coupled Cluster. Specifically, mean absolute errors in potential energy surfaces are (a) less than 0.05 kcal/mol when the PIE-ONIOM results that combine B3LYP and MP2, are compared with MP2 calculations on the full system, and (b) less than 0.07 kcal/mol when a similar hybrid calculation is compared with CCSD(T) on the full system. This data was obtained in Ref. 32 by considering 693 different geometries of protonated water clusters sampled along normal mode coordinates and the associated computational effort is a very small fraction of the computational cost incurred in the full calculations. The AIMD trajectories computed using this approach reproduce vibrational density of states from regular MP2 calculations.
FIG. 1: Figure 1 and Figure 2 briefly summarize the developments from Ref. 32. Here, the vibrational ground state ($\rho(R)$) weighted mean absolute error in potential surface is provided for $H_9O^-_4$. The error is computed between the fragment based calculations and full CCSD(T) calculations as $\int dR \left\{ \rho(R) \left| E_{\text{CCSD(T)}} - E_{\text{CCSD(T):B3LYP}} \right| \right\}$. The statistics is obtained from 693 different geometries chosen along normal mode coordinates where the horizontal axis in the figure represents mode frequencies. Thus, the figure depicts the error in the critical regions of the surface sampled by the ground state wavepacket and is clearly very small. When the full surface is included (that is without vibrational ground state weighting) the error rises to the order of 0.1 kcal/mol. All calculations use 6-31+G(d,p) basis function.

FIG. 2: The solvated Zundel cation, $H_{13}O_6^+$. The vibrational density of states are computed from the fragment based AIMD trajectories (frag-BOMD MP2:B3LYP) and these agree quite well with full MP2 based AIMD. All calculations use the 6-31+G(d,p) Gaussian basis set.

as well and these results are summarized in Figures 1 and 2. As seen in Figure 1(b), and noted in the caption, the potential energy surface errors are indeed extremely small and vibrational density of states peaks agree quite well when the hybrid approach is compared to higher level theory.

Despite the preliminary success of our approach in Ref. 32, several challenges remain:
During dynamical simulations, the molecular structural variation can lead to changes in the fragmentation topologies that must be handled carefully to maintain conservative dynamics. The computational expense may be dominated by the lower level calculation on the full system, since our approach is based on ONIOM. This second hurdle is also seen in other hybrid fragmentation schemes, such as the Generalized Energy Based Fragmentation (GEBF)\(^1\), Molecular Tailoring Approach (MTA)\(^1^{4}\), the eXtended ONIOM method\(^1^{5}\) and Molecule-in-Molecule (MIM)\(^1^{6}\). Analytical gradient for the embedded charges can be expensive\(^4^{3}\,4^{4}\). In the current publication, we propose a solution to the second problem listed above by constructing a hybrid Extended Lagrangian/Born-Oppenheimer dynamics scheme. Thus the paper is organized as follows: Section II introduces the new formalism with facilitating technical discussions provided in Appendix A. The approach is then used in Section III to compute structural properties, dynamical trajectories and correlation functions for protonated water clusters and polypeptide fragments. In all cases, the trajectories are compared with MP2 based AIMD trajectories and the results are in good agreement. We also demonstrate numerical benchmarks in Section VI that show a large reduction in computational effort. Furthermore, the algorithm is written in parallel using MPI (also demonstrated in Section VI) and through a C++ driver that calls external electronic structure packages. Concluding remarks are made in Section VII.

II. HYBRID EXTENDED LAGRANGIAN BORN-OPPENHEIMER DYNAMICS USING FRAGMENT-BASED ELECTRONIC STRUCTURE FROM PIE-ONIOM FOR POST-HARTREE-FOCK TREATMENT

Classical ground state ab initio molecular dynamics primarily has two flavors, one is the traditional Born-Oppenheimer molecular dynamics (BOMD) where the electronic energy is converged; the other is the extended Lagrangian molecular dynamics (ELMD)\(^3\,4^{5}\,5^{7}\) formalism where the electronic degrees of freedom, either single particle density matrix or molecular orbital coefficients, are assigned fictitious inertia and propagated simultaneously with the classical nuclei through a simple adjustment of time-scales\(^5^{8}\). The first ELMD method, the Car–Parrinello molecular dynamics (CPMD)\(^3\,4^{5}\,4^{7}\) approach was influenced by earlier work from Anderson\(^5^{9}\), Parrinello and Rahman\(^4^{5}\,6^{0}\), and was developed in 1985. In CPMD, the wavefunction is expanded in plane-wave basis and pseudo-potentials are used to describe
the core electrons. In the 2001, Atom-centered Density Matrix Propagation (ADMP)\textsuperscript{48–54}, an ELMD formalism that employs the single particle electronic density matrix written in an all-electron atom-centered Gaussian basis function, was developed. The ADMP method has been demonstrated through applications to vibrational spectroscopy\textsuperscript{61–65,65,66}, biological ion-channel studies\textsuperscript{54} and the study of atmospheric reactions where pump-probe techniques have been incorporated into the dynamics\textsuperscript{67–69}. Critically, the ADMP formalism has had an important role in the prediction of the amphiphilic nature of the hydrated proton\textsuperscript{70,71} that was later confirmed through several experiments\textsuperscript{70–76}. The ADMP method has many attractive features. Systems can be simulated by accurately treating all electrons or by using pseudo-potentials. Through the use of smaller values for the tensorial fictitious mass, large time-steps can be employed and lighter atoms such as hydrogens are routinely used. A wide variety of exchange-correlation functionals can be utilized, including hybrid density functionals such as B3LYP. Atom centered functions can be used with the appropriate physical boundary conditions for molecules, polymers, surfaces and solids, without the need to treat replicated images to impose 3d periodicity. Consequently, charged systems and QM/MM models of biological systems\textsuperscript{54} can be treated as easily as neutral molecules. Deviations from the Born-Oppenheimer surface and the mixing of fictitious and real kinetic energies can also be rigorously controlled on-the-fly in ADMP. ADMP trajectories of the order of picoseconds show stable dynamics, and the adiabaticity can be controlled effectively in these systems without thermostats. The important conceptual and computational differences between ADMP and other Gaussian basis-set based implementations\textsuperscript{77–80} of the Car-Parrinello formalism have been discussed in detail in Refs. 50.

The elimination of SCF in extended Lagrangian MD is a computational advantage, especially for large molecules and systems where SCF convergence may be challenging. However, thus far, a critical restriction for ELMD (Car Parrinello as well as ADMP) has been that it can only be applied to single particle formalisms such as DFT and Hartree-Fock. Here we combine ELMD and post-Hartree-Fock-based BOMD through PIE-ONIOM and allow extended Lagrangian (Car-Parrinello) techniques to become applicable to higher level electronic structure methods, such as MP2 and potentially CCSD, and also improve the computational efficiency of fragment based AIMD. But before we embark into a discussion of this dynamical formalism, we must first summarize the PIE-ONIOM energy extrapolation
equations. The overall PIE-ONIOM molecular energy expression\textsuperscript{32} is given by

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

where, the extrapolation term is defined as:

\[
S(i) = E_{\text{high}}(i) - E_{\text{low}}(i)
\]

when two levels of theory are used. (This approach is easily generalized to more levels\textsuperscript{32}, but this generalization is not considered here.) In Eq. (1), \(S(i \cap j)\) represents the extrapolation term for the derivative fragment \((i \cap j)\), formed from the intersection of the \(i\)-th and \(j\)-th primitive fragments. Every fragment is treated at two levels of theory as written in Eq. (2), and the entire system (0th fragment) is only considered at the lowest level of theory (with energy \(E_{\text{low}}(0)\)). This energy extrapolation scheme is pictorially demonstrated in Figure 3. A powerful feature of the algorithm is the fact that, the identity of the overlapping model systems, or fragments, are efficiently evaluated either (i) through an automated bit-manipulation algorithm presented in Ref. 32 that recomputes fragments at each instant in time during dynamics, or (ii) through user specification. In the bit-manipulation algorithm, fragments are represented as bits in an integer, and overlapping regions are computed using binary AND operations. The use of bitwise arithmetic improves efficiency, but a large integer library may be required, depending on the number of fragments involved and the platform being used for the calculation. One such library is available at \url{http://www.boost.org/}.

Clearly, for a large enough system the computational expenses may be dominated by the lower level calculation, \(E_{\text{low}}(0)\). Thus, in the hybrid extended Lagrangian, Born-Oppenheimer scheme introduced here, it is electronic degrees of freedom of the entire system, \(E_{\text{low}}(0)\), that we will propagate using an extended Lagrangian technique. The rest of the fragments, where more accurate calculations maybe needed, will be treated using converged Born-Oppenheimer treatments using post-HF electronic structure. While we choose to demonstrate this approach using the ADMP Lagrangian\textsuperscript{48,49} as the choice for ELMD, extensions to employ CPMD (that is, using molecular orbital coefficients instead of the single
\[ L = \frac{1}{2} \text{Tr}[V^T M V] + \frac{1}{2} \text{Tr} \left[ \left( \frac{1}{4} \mu_{\text{low,0}}^{1/4} W_{\text{low,0}} \right) \right]^2 - E^{\text{PIE-ONIOM}}(R, P_{\text{low,0}}) - \text{Tr}[\Lambda_{\text{low,0}}(P_{\text{low,0}}^2 - P_{\text{low,0}})] \] (3)

In this equation, \( R \), \( V \) and \( M \) are the nuclear geometry, velocity, and mass respectively, and the single particle density matrix \( P_{\text{low,0}} \) that determines \( E_{\text{low}}(0) \) has fictitious inertia tensor \( \mu_{\text{low,0}} \), and velocity \( W_{\text{low,0}} \). The quantity \( \Lambda_{\text{low,0}} \) is a Lagrangian multiplier that maintains the N-representability of \( P_{\text{low,0}} \). The energy functional \( E^{\text{PIE-ONIOM}}(R, P_{\text{low,0}}) \) is identical to that in Eq. (1), but differs with the additional dependence of \( E_{\text{low}}(0) \rightarrow E_{\text{low}}(0)(R, P_{\text{low,0}}) \), since the lower level full system electronic structure is now to be propagated rather than converged. This leads to the equations of motion for nuclei and density matrix:

\[ M \frac{\partial^2 R}{\partial t^2} = - \frac{\partial E^{\text{PIE-ONIOM}}(R, P_{\text{low,0}})}{\partial R} \bigg|_{P_{\text{low,0}}} \] (4)
and

\[ \frac{1}{2} \frac{\partial^2 P_{\text{low},0}}{\partial t^2} \mu_{\text{low},0} \frac{1}{2} \mu_{\text{low},0} = - \frac{\partial E^{\text{PIE-ONIOM}}(R, P_{\text{low},0})}{\partial P_{\text{low},0}} \bigg|_R - A_{\text{low},0} P_{\text{low},0} - P_{\text{low},0} A_{\text{low},0} + A_{\text{low},0} \]  

(5)

where

\[ \frac{\partial E^{\text{PIE-ONIOM}}(R, P_{\text{low},0})}{\partial R} \bigg|_{P_{\text{low},0}} = \frac{\partial E_{\text{low}}(0)}{\partial R} \bigg|_{P_{\text{low},0}} + \sum_{i=1}^{n} \frac{\partial S(i)}{\partial R} - \sum_{1 \leq i < j \leq n} \frac{\partial S(i \cap j)}{\partial R} + \ldots \]  

(6)

and

\[ \frac{\partial E^{\text{PIE-ONIOM}}(R, P_{\text{low},0})}{\partial P_{\text{low},0}} \bigg|_R = \frac{\partial E_{\text{low}}(0)}{\partial P_{\text{low},0}} \bigg|_R . \]  

(7)

In Eq. (6), the constraint, “|\(P_{\text{low},0}\)”, is only included on the first term on right side since that is the only term that is a function of \(P_{\text{low},0}\). The terms \(\frac{\partial S(i)}{\partial R}\), \(\frac{\partial S(i \cap j)}{\partial R}\), etc., involve standard electronic structure gradients that are already present in most electronic structure packages. This is because, as noted in Eq. (2), these terms involve gradients from higher level electronic structure methods (MP2 for the calculations presented in here) and when DFT is used for the lower level in Eq. (2), the SCF is converged to a high degree of accuracy. In this publication, all SCF calculations involved in \(\frac{\partial S(i)}{\partial R}\), \(\frac{\partial S(i \cap j)}{\partial R}\), are well converged requiring the RMS change in density matrix to be less than \(10^{-8}\) and the maximum change in the density matrix to be less that \(10^{-6}\). These are the default threshold conditions used in the version of the Gaussian quantum chemistry package\(^8\) used here to compute the electronic energy and nuclear gradients for the fragments. The energy change during SCF is not used to test convergence but a \(10^{-N}\) RMS change in density matrix typically corresponds to a \(10^{-2N}\) change in energy in atomic units. While in principle this translates to an energy convergence threshold of \(10^{-16}\) Hartree, in practice we find that SCF convergence is always better than \(10^{-12}\) Hartree in most calculations presented here, with very few at \(10^{-10}\). (A more detailed analysis on this aspect is presented in Section III.) As a result of this, there is practically no drift in these calculations, which appears to be a problem for Born-Oppenheimer-type dynamics calculations in Ref. 82.
Proceeding further with the discussion of the equations of motion, since $P_{\text{low},0}$ is propagated according to Eq. (5), the gradient term $\left. \frac{\partial E_{\text{low}}(0)}{\partial R} \right|_{P_{\text{low},0}}$ in Eq. (6) contains forces from the un converged density matrix, $P_{\text{low},0}$, in addition to the standard Born Oppenheimer gradients and these aspects have been well described in Ref. 48,49,51. We summarize these gradients in Appendix A since the major portion of these forces have already been discussed in other publications noted above. It is only the interpretation of these gradients with respect to the fragment based technology that is new here. The equations of motion are integrated using the velocity Verlet algorithm while the Lagrangian multiplier matrices are determined through an iterative procedure. In the sections below we probe the accuracy and efficiency of this hybrid scheme by studying protonated water clusters and poly-peptides fragments. In all cases consistency of the dynamics is checked by computing the conserved quantities and these conserved quantities (the conjugate Hamiltonian for the Lagrangian in Eq. (3)) are outlined in Appendix B.

But before we proceed further we must note and distinguish our approach from two other recent developments in Born-Oppenheimer and extended Lagrangian dynamics. In Ref. 84, the author has carefully analyzed the time-scale of oscillations of the different constituents of energy in a post-Hartree-Fock (MP2) gradient expression. The author infers through many computational studies that the post-Hartree-Fock portion of the gradient is in fact a slowly varying quantity as a function of nuclear motion, which allows the authors to construct a multiple time-step approach that appears to be quite powerful. This extension is certainly an option available to us here and will be considered in future publications. In Ref. 82, the authors address the critical drift problem in standard Born-Oppenheimer dynamics by propagating a surrogate electronic density through an extended Lagrangian formalism, where such a density is harmonically bound to a converged density. This latter approach has extended Lagrangian and Born-Oppenheimer components, much like the current formalism, but differs in that the ideas in Ref. 82 are thus far implemented for DFT and DFTB, whereas the approach here is specifically geared towards approximating post-Hartree-Fock dynamics. Thus, to differentiate between the two approaches, from hereon, the method developed in this paper is referred to as extended Lagrangian, post-Hartree-Fock Born-Oppenheimer dynamics (EL/(pHF)BO-MD), and the PIE-ONIOM energy functional, Eq. (1), is implied. When the extended Lagrangian treatment is not invoked and Born-Oppenheimer treatment is used for all components of the energy expression in Eq. (1), the associated method is referred to here.
FIG. 4: The protonated water clusters shown in Figures 1(a) and 2(b) are fragmented as depicted here and the electronic energy is assembled using Eq. (1). The proline tri-peptide (Figure (c) is similarly fragmented using overlapping dipeptides (Figure (d)). Dangling bonds are saturated with link atoms (hydrogen) as in standard ONIOM.

as frag-BOMD.

III. ACCURACY AND EFFICIENCY OF THE HYBRID EXTENDED LAGRANGIAN, POST-HARTREE-FOCK BORN-OPPENHEIMER DYNAMICS (EL/(PHF)BO-MD) TREATMENT

We benchmark the dynamics formalism by studying (small and medium sized) protonated water clusters and two different poly-peptide fragments. We choose two challenging protonated water clusters, the Eigen cation ($H_9O_4^+$) and the solvated Zundel cation ($H_{13}O_6^+$), as our benchmark systems with their structures shown in Figure 4(a) and (b). These type of systems have significance in biological\textsuperscript{85-92}, atmospheric\textsuperscript{69,93-95} and condensed phase chemistry, and have been studied extensively by both experiments\textsuperscript{96-104} and
by theories. Here, accurate treatment of hydrogen bonding is crucial to model structural and dynamical behavior. In addition, poly-peptide fragments have recently been studied in some detail using mass spectrometry with a goal to elucidate conformational transformations and kinetics. For example, in Ref. 121,123 the authors observe a step-by-step conformational transition from an all-cis poly-proline helix chain to an all-trans poly-proline helix chain. While our primary benchmarks here, that test the accuracy of EL/(pHF)BO-MD and frag-BOMD through comparison with full system MP2 calculations, are performed for the protonated water clusters stated above, we also present preliminary benchmarks on an all-trans proline tri-peptide fragment (Figure 4(c)) to show the consistency of our approach and its potential for future applications to the study of bio-molecules. In addition we also present an analysis of the computational gain from the fragment-based EL/(pHF)BO-MD and frag-BOMD formalisms by computing trajectories for an all-trans poly-alanine fragment.

At this point it is critical to note that for the peptide systems, chemical bonds intersect the boundary between two fragments and hence link atoms are added to saturate the dangling valencies as in standard ONIOM. The positions of link atoms are uniquely determined based on the connectivity of the system. Specifically, the placement of link atoms is constrained by the positions of the two atoms on either ends of the bond intersecting the fragment boundary according to $r_{\text{link}} = r_{\text{bond}} + g(r_{\text{sub}} - r_{\text{bond}})$, where $r_{\text{bond}}$ is the position of the atom inside the fragment in question, to which the link is bound to substitute the atom belonging to the surrounding region and located at $r_{\text{sub}}$. The quantity $g$ is a scale factor normally set to $R(B - L)/R(B - S)$, where $R(B - L)$ is a typical bond length between the bonded atom within the fragment and the link atom (hydrogen in all cases in this work), and $R(B - S)$ is a typical bond length for the pair of atoms describing the bond being dissected to create the fragment. Careful tests have shown that the ONIOM results are quite insensitive to the value of $g$. It must be stressed that since the link atom positions are fully defined by the coordinates of the full system, as in standard ONIOM, the gradients with respect to link atom coordinates, can be transferred to coordinates of the full system for this fragment based methodology.

All calculations are performed using a C++ driver routine, that is MPI parallelized across nodes and OpenMP parallelized within each node. For the current publication this driver routine employs a development version of the Gaussian quantum chemistry package to com-
compute the electronic energy and nuclear gradients for the fragments and the full system. But the driver routine can be easily generalized to include other electronic structure packages, or, combinations of packages. These computational aspects and associated generalizations will be discussed in future publications.

The dynamical simulations are performed five different ways for each system with all calculations using the 6-31+g(d,p) atom-centered Gaussian basis set. The five different sets of trajectories are: (1) BOMD using pure B3LYP (abbreviated as BOMD B3LYP). (2) ADMP using pure B3LYP (abbreviated as ADMP B3LYP). (3) BOMD using pure MP2 (abbreviated as BOMD MP2). (4) BOMD using two-layer PIE-ONIOM fragment-based electronic structure. MP2 is used for the high level electronic structure while the low level is calculated at B3LYP. This calculation is abbreviated as frag-BOMD MP2:B3LYP, and was fully benchmarked in Ref. 32. Also see Figures 1 and 2 where we present the accuracy of this calculation. (5) Finally, the two-layer PIE-ONIOM fragment-based EL/(pHF)BO-MD scheme proposed in this publication. The methods, MP2 and B3LYP are used for high and low level of theory respectively and the calculation is abbreviated as EL/(pHF)BO-MD MP2:B3LYP.

The stability and consistency of the dynamics trajectories can be gauged from the total energy conservation data for these classical trajectories. The total energy corresponds to the conjugate Hamiltonian in Eq. (B4) and contains the nuclear kinetic energy, the fictitious density matrix kinetic energy, $E^{PIE-ONIOM}$ and any energy penalty that is added as a result of residual loss of N-representability. As shown in Table I, all trajectories have the total energy well conserved to within two or three orders of magnitude less than 1 kcal/mol.
### Table I: Energy conservation properties for dynamical simulations (NVE simulations).

<table>
<thead>
<tr>
<th>System</th>
<th>Level of Theory</th>
<th>Time</th>
<th>Ave Temp (K)</th>
<th>$\Delta H$</th>
<th>Drift $^d$</th>
<th>$\Delta E_{SCF}$ $^e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_9O_4^+$</td>
<td>BOMD B3LYP</td>
<td>20.0ps</td>
<td>150.90K±27.0K</td>
<td>0.022</td>
<td>0.062</td>
<td>1.74x10$^{-11}$</td>
</tr>
<tr>
<td></td>
<td>ADMP B3LYP</td>
<td>20.0ps</td>
<td>140.34K±23.7K</td>
<td>0.021</td>
<td>0.018</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BOMD MP2</td>
<td>20.0ps</td>
<td>152.10K±26.5K</td>
<td>0.002</td>
<td>0.004</td>
<td>6.38x10$^{-12}$</td>
</tr>
<tr>
<td></td>
<td>frag-BOMD MP2:B3LYP</td>
<td>5.2ps</td>
<td>149.64K±32.3K</td>
<td>0.014</td>
<td>0.016</td>
<td>4.08x10$^{-12}$</td>
</tr>
<tr>
<td></td>
<td>EL/(pHF)BO-MD MP2:B3LYP</td>
<td>8.6ps</td>
<td>142.80K±24.7K</td>
<td>0.027</td>
<td>0.055</td>
<td>3.34x10$^{-12}$</td>
</tr>
<tr>
<td>$H_{13}O_6^+$</td>
<td>BOMD B3LYP</td>
<td>20.0ps</td>
<td>151.92K±20.8K</td>
<td>0.010</td>
<td>0.011</td>
<td>2.75x10$^{-11}$</td>
</tr>
<tr>
<td></td>
<td>ADMP B3LYP</td>
<td>13.4ps</td>
<td>154.00K±21.0K</td>
<td>0.040</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BOMD MP2</td>
<td>5.2ps</td>
<td>150.52K±22.6K</td>
<td>0.004</td>
<td>0.001</td>
<td>1.32x10$^{-11}$</td>
</tr>
<tr>
<td></td>
<td>frag-BOMD MP2:B3LYP</td>
<td>6.0ps</td>
<td>151.22K±23.4K</td>
<td>0.042</td>
<td>0.122</td>
<td>6.05x10$^{-12}$</td>
</tr>
<tr>
<td></td>
<td>EL/(pHF)BO-MD MP2:B3LYP</td>
<td>5.0ps</td>
<td>156.44K±20.0K</td>
<td>0.026</td>
<td>0.033</td>
<td>2.93x10$^{-12}$</td>
</tr>
<tr>
<td>Proline-trimer$^f$</td>
<td>ADMP B3LYP</td>
<td>4.33</td>
<td>237.96K±24.8K</td>
<td>0.069</td>
<td>0.148</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BOMD B3LYP</td>
<td>2.18</td>
<td>238.09K±23.7K</td>
<td>0.016</td>
<td>0.011</td>
<td>8.41x10$^{-12}$</td>
</tr>
<tr>
<td></td>
<td>frag-BOMD MP2:B3LYP</td>
<td>2.97</td>
<td>246.30K±23.5K</td>
<td>0.012</td>
<td>0.002</td>
<td>2.96x10$^{-12}$</td>
</tr>
<tr>
<td></td>
<td>EL/(pHF)BO-MD MP2:B3LYP</td>
<td>2.90</td>
<td>246.07K±22.5K</td>
<td>0.047</td>
<td>0.088</td>
<td>1.64x10$^{-11}$</td>
</tr>
</tbody>
</table>

$^a$Water clusters: 6-31+g(d,p) basis is used for B3LYP and MP2. Proline-trimer: 6-31+g(d) basis is used for B3LYP and MP2.

$^b$time step 0.2 fs for all simulations

$^c$Root mean square deviation of the total energy in kcal/mol. The total energy corresponds to the Hamiltonian in Eq. (B4).

$^d$Drift in total energy is computed as a difference between the average total energies for the first and last 100fs of dynamics. (in kcal/mol) In addition, in supporting information, Figure SI-1, we provide the evolution of total energy. This is provided in eV to facilitate comparison with results from other work such as Ref. 82.

$^e$While SCF convergence thresholds are placed on the density matrix (see text for details) rather than energy, in this column we note the average absolute change in energy (in Hartrees) during the last iteration before convergence is achieved for all steps in the trajectory.

$^f$The C-terminus is capped with a carboxylic acid group. See Figure 4(c).

The EL/(pHF)BO-MD scheme produces energy conservation comparable to the BOMD simulations. We also inspect the drift in the conjugate Hamiltonian in Eq. (B4) and as noted in Table I, this quantity is also extremely small and well within the acceptable range$^{82}$. In addition, in Table I, we also present the change in SCF energy in the last iteration before convergence is achieved. This quantity in some sense represents a convergence threshold of the electronic energy. As noted previously, convergence is enforced on the density matrix and not on the SCF energy. In this case, it is required that the RMS change in density matrix be less than 10$^{-8}$ and the maximum change in the density matrix be less that 10$^{-6}$. The corresponding SCF energy convergence is very tight and is noted in Table I.

In addition to gauging conserved quantities, we also compare forces in Figure 5. Specifi-
FIG. 6: Radial distribution function for O-O and O-H distances from various $H_9O^+_4$ trajectories. The statistics is obtained by constructing averages from incremental 2 picosecond snapshots constructed from the full trajectory. That is we compute the RDFs using data from 0-2ps, 0.5-2.5ps, and so on, for all trajectories. The resultant distributions are then averaged here, standard deviations are presented as errorbars in supporting information, Figures SI-2.

Generally, in Figure 5(a) we compute the forces for 693 different geometries chosen along normal mode coordinates where the horizontal axis in the figure represents mode frequencies. The forces are computed at the level of MP2 and also using the MP2:B3LYP fragment method. The difference $\frac{\|F_{MP2:B3LYP} - F_{MP2}\|}{3N_{Atoms}}$ is presented as a function of harmonic frequency. In addition, for the analysis in Figure 5(b) we have retrieved geometries at fixed intervals from the extended Lagrangian Born Oppenheimer trajectory for $H_9O^+_4$ and computed forces using MP2. The vertical axis in Figure 5(b) is the mean absolute deviation between these forces. As can be seen the forces in both figure are well within that expected for chemical accuracy.

IV. STRUCTURAL FEATURES FROM DYNAMICS SIMULATIONS

Next, we examine the structural properties of these trajectories by calculating radial distribution functions for the O-O and O-H distances for the protonated water clusters. From Figure 6 and 7, it is clear that the structural features for the water clusters are captured very well by both EL/(pHF)BO-MD and frag-BOMD fragment based dynamics in comparison with the higher level MP2 calculations. In addition, for the protonated water clusters, these calculations are also in agreement with the plain DFT calculations. Specifically, Figures
FIG. 7: Radial distribution function for O-O and O-H distances from various $H_{13}O_6^+$ trajectories are obtained using the same procedure described in the caption for Figure 6. Errorbars can be found in supporting information, Figures SI-3.

6(a) and 7(a) clearly show the multiple solvation shells, where the width of the distributions indicate fluctuations that arise from temperature. The inner shells have better ordering in all cases, based on width of the distributions, and this is to be expected. The most critical aspect from Figures 6(a) and 7(a) is the fact that the inner shell is tighter in Figure 7(a) with a shorter OO distance. This is due to the strong Zundel-like hydrogen bond present in $H_{13}O_6^+$. Thus, Figure 7(a) shows both Zundel-like and Eigen-like (OO distance $\approx 2.8\AA$) peaks and this is critically reinforced in all simulations. The peaks beyond 3.7Å in Figures 6(a) and 7(a) represent the second solvation shell that is generally not expected to be well-resolved for these smaller water clusters. Nevertheless, the agreement is reasonable here too, but not as good as in the first solvation shell. The OH distributions in Figure 6(b) similarly show an Eigen-like OH distance of approximately 1.5Å which informs an asymmetric hydrogen bond; the two shortest OH distributions in Figure 6(b) are consistent with the OO distribution for the first peak in Figure 6(a). Figure 7(b) shows four regions of OH distribution. The covalent OH bond is strongly noted, and is followed immediately by a shallow hump at a distance slightly greater than 1Å. This distance is again the short-strong Zundel-like hydrogen bond feature. Followed by this we see a peak at a little greater than 1.5Å which is consistent with an Eigen geometry. Thus, the $H_{13}O_6^+$ system is critical in that, it shows both Zundel and Eigen configurations that are now universally acknowledged as
FIG. 8: Vibrational density of states calculated from dynamical trajectories at 150K for (a) $H_9O_4^+$ and (b) $H_{13}O_6^+$. The excellent agreement in frequencies is highlighted using the light colored arrows. In Supporting Information, we provide details on the statistics and compute error bars.

the two central excess proton configurations in water that participate in proton transport\textsuperscript{125}. Importantly all simulations, including EL/(pHF)BO-MD reproduce the structural features present in the MP2 BO dynamics.

This brings in the critical question of computational efficiency of fragment based AIMD methods. We discuss these aspects in Section VI.

V. DYNAMICAL PROPERTIES FROM CORRELATION FUNCTIONS

Vibrational density of states (VDOS) are calculated for each trajectory from the Fourier transform of the velocity auto-correlation function. As we know from previous studies\textsuperscript{61–63,126}, the temperature can play a significant role in VDOS spectra since this controls the extent to which anharmonicity is sampled during these classical nuclear trajectory calculations. Hence we tune the initial conditions such that all trajectories have an average temperature of about 150K. This is also roughly the temperature of argon tagged action spectroscopy\textsuperscript{127} as seen in previous calculations\textsuperscript{62,65}. The VDOS spectra are presented in Figure 8 and Figure 9. To obtain these figures, we compute multiple spectra using 3ps segments from dynamics that are each shifted by 500fs. That is we compute the vibrational density of states using data from 0-3ps, 0.5-3.5ps, and so on, for all trajectories. These spec-
FIG. 9: Vibrational density of states calculated from dynamical trajectories at 150K for (a) $H_9O_4^+$ and (b) $H_{13}O_6^+$. The extended Lagrangian formalism yields a slight red-shift in the higher frequency peaks greater than 2500 cm$^{-1}$ but is in reasonably good agreement in the lower frequency region. The red-shift is reduced as compared to ADMP which does not include the MP2 treatment. In Supporting Information, we provide details on the statistics and compute error bars.

The averages are presented in Figures 8 and 9 and the standard deviations are shown using error bars in supporting information (Figures SI-4 through SI-8). Specifically, in Figure 8 of the main text and Figures SI-4 through SI-6 we compute the vibrational density of states obtained from full MP2 and frag-BOMD MP2:B3LYP and the fact that the frequencies are in excellent agreement all through the spectral range is shown using arrows to guide the eye. In Figure 9 of the main text, and Figures SI-7 and SI-8, we provide the same for the full breadth of simulations. (Also note that to check convergence, we also computed the same using 2 picosecond snapshots. That is, we also computed the vibrational density of states using data from 0-2ps, 0.5-2.5ps, and so on, for all trajectories. No perceivable difference was noted from the spectra presented here.) It is clear from these figures that the DFT results are red-shifted with
respect to those from MP2. This behavior is drastically improved upon in Figure 8 when the MP2 results are compared with those from MP2:B3LYP fragment based BOMD. This is perhaps the most significant message here, that the fragment based approaches are able to provide results in quantitative agreement with higher level electronic structural results. The EL/(pHF)BO-MD result is red-shifted with respect to the MP2 BOMD and MP2:B3LYP fragment based BOMD, but this red-shift is substantially reduced as compared to the respective ADMP red-shifts seen in Figure 9. Furthermore it appears that these red-shifts are more localized to the higher frequencies as note in Figures SI-7 and SI-8. This is to be expected since one of the central ideas behind ADMP and similar extended-Lagrangian formalisms is that the electronic dynamics is mapped onto a set of fictitious variables that obey classical-like dynamics. The mapping is enforced by including an inertia tenor as outlined above and the choice of this inertia tensor does have a role on the higher frequency correlation functions. This aspect has also been noted in previous publications\textsuperscript{50,55}. At the end it appears the structural properties from EL/(pHF)BO-MD are quantitatively accurate and correlation functions are also accurate in lower end of the spectrum. Future work will deal with improvement of these methods in the higher frequency regimes. Fragment based BOMD, on the contrary, has accuracy that is very much comparable to standard full post-Hartree-Fock BOMD (with lesser computational overhead) and hence appears to be well suited for production level calculations.

VI. COMPUTATIONAL EFFICIENCY OF THE EL/(PHF)BO-MD SCHEME

The system studied to demonstrate computational efficiency is the all-trans alanine decamer shown in Figure 10. The full system MP2 timings are prohibitive for this case.

As noted in the histogram plots presented in Figure 11, the frag-BOMD and EL/(pHF)BO-MD simulations contain three separate parts: (i) the DFT energy and gradient calculation for the full system, (ii) the DFT energy and gradient calculations for the fragments, and (iii) MP2 energy and gradient calculations for fragments. The parallelism for the fragment based frag-BOMD simulations is organized such that the full system low-level calculation is performed on a separate node, comprising 16 shared memory processors, in parallel with the fragment high and low level pairs of jobs that are spawned to other similar nodes. The parallelism for the fragment based EL/(pHF)BO-MD is organized in a slightly
FIG. 10: The all-trans alanine decamer. Figure (a): helical view with hydrogen bonds shown. Figures (b) and (c) displays a portion of the polypeptide with fragments labeled. For example the primary fragment shown in Licorice in Figure (b) includes the CO-HN hydrogen bond and it has an overlap with the purple, red, cyan and tan fragments. This is also shown in Figure (c). (The secondary overlapping fragments are automatically generated using a bit-manipulation algorithm presented in Ref. 32.) In this manner critical hydrogen bonds responsible for the helix formation are captured as part of the fragmentation mechanism.

In Figures 11(a) and 11(b) we present the total time per dynamics step (black curve), the
FIG. 11: All calculations are performed on Intel Xeon E5-2650 v2 processors. See text for details.

CPU time required for the full system (low level) DFT calculation (blue histograms) and the average CPU time required for both low and high level calculations for each fragment (red histograms). In both figures the CPU time required for the full system low level (blue) is flat because this calculation is performed on a single node with 16 processors for both EL/(pHF)BO-MD and frag-BOMD. Further reduction in the total time step can be obtained for both EL/(pHF)BO-MD and frag-BOMD by increasing the number of processors for full system low level, but since this is independent of the fragmentation algorithm (and is dependent on the restrictions on implementation of the underlying electronic structure software), we only use a fixed number of processors here. Thus all electronic structure energy and gradient calculations are localized to 16 processors in this benchmark. In general one would expect a similar scaling behavior for any number of processors (assuming the electronic structure method and program used parallelizes appropriately).

The next critical observation in Figure 11(a) is that the full system calculation essentially forces the total CPU time to flatten out at 80 processors and beyond, and the resources required for the full system calculation are nearly a factor of four greater than that for the fragment calculations. This is a major bottleneck for fragment based frag-BOMD when applied to large systems. The EL/(pHF)BO-MD, as seen from the blue histogram in Figure 11(b), reduces the time required for the low level calculation by nearly a factor four and now the full system and fragment calculations require comparable resources. This allows for better load balance and higher code efficiency for larger number of processors and directly translates toward the greater efficiency for the EL/(pHF)BO-MD scheme. In addition, one notices that the total CPU time (black curve) never approaches the blue histogram in Figure
11(b) unlike the case in Figure 11(a). This is partially due to the code inefficiency mentioned above, where the low level calculation is performed before the fragment calculations are spawned. Despite this inefficiency the scaling is better for EL/(pHF)BO-MD and this scaling is expected to further improve in future versions of the computer program. Furthermore, at larger numbers of processors (greater than 200), the low level calculation again uses resources that are nearly a factor of four greater than the fragment calculation. This will require that the lower level calculations be performed over a larger number of processors (that is repartitioning the node allocation) which should further improve the scaling for number of processors greater than 200.

Given that structural features are well preserved for strongly hydrogen bonded systems and for peptides, the EL/(pHF)BO-MD implementation is expected to have an advantage in large scale simulations.

VII. CONCLUDING REMARKS

In this publication, we introduce hybrid extended Lagrangian, post-Hartree-Fock Born-Oppenheimer dynamics approach where the two schemes are combined through a recently developed\textsuperscript{32} fragment-based electronic structure formalism\textsuperscript{32} that employs the set-theoretic principle of inclusion-exclusion\textsuperscript{42} to generalize the well-known ONIOM\textsuperscript{5} approach. Here a large system is first treated using DFT, and then recursively partitioned into overlapping subsections where the electronic description is improved through higher levels of electronic structure treatment; the intersection regions between the overlapping subsections are appropriately corrected as allowed by the principle of inclusion-exclusion in conjunction with ONIOM. The hybrid dynamical method described here propagates the electronic structure for the full system, by an extended Lagrangian scheme whereas the fragments are treated using converged post-HF methods. This significantly reduces the cost for electronic structure calculations on the full system and will be beneficial for ab initio dynamics treatment of large systems where repeated evaluation of energy and gradients of the entire system may be computationally prohibitive.

This approach is benchmarked on a set of well studied protonated water clusters, $H_9O_4^+$ and $H_{13}O_6^+$ and polypeptide fragments: proline tri-peptide and alanine polypeptide. All dynamics trajectories show good energy conservation in the sub-kcal/mol range. The
structural features of the protonated water clusters show that the fragment based hybrid EL/(pHF)BO-MD approach provides results in good agreement with BO dynamics at MP2 level of theory. The analysis of vibrational density of states also indicates large improvements in EL/(pHF)BO-MD over pure extended Lagrangian results at the lower level of theory; furthermore, EL/(pHF)BO-MD provides better agreement with the spectra calculated from higher level Born-Oppenheimer dynamics. In all cases the computational gain in frag-BOMD is orders of magnitude greater than the full MP2 calculations, and EL/(pHF)BO-MD improves further on this performance. The algorithm is written in parallel and potentially allows AIMD treatment of large systems. Two challenges remain in fragment based ab initio molecular dynamics: (a) The topology of fragmentation could change during dynamical simulations, thus an adaptive fragmentation and energy extrapolation scheme needs to be developed in order to study processes involving structural rearrangements, such as proton transfer reactions and conformational changes in poly-peptides. (b) The calculation of forces introduced by electronic embedding could be expensive\textsuperscript{43,44} and computational feasibility studies are required. These aspects will be probed in future publications.

VIII. ACKNOWLEDGMENTS

This research is supported by the National Science Foundation grant NSF CHE-1058949 to SSI. In addition funding from the Indiana University’s Faculty Research Support Program is also acknowledged.

Supporting Information: Supporting information includes additional figures describing the evolution of total energy during dynamics for all simulations, and figures pertaining to the statistics in the radial distribution and vibrational density of states. This information is available free of charge via the Internet at http://pubs.acs.org/

APPENDIX A: NUCLEAR AND DENSITY MATRIX GRADIENTS ASSOCIATED WITH EQUATIONS (4) AND (5)

Since $P_{\text{low},0}$ is propagated according to Eq. (5), the gradient term $\left. \frac{\partial E_{\text{low}}(0)}{\partial R} \right|_{P_{\text{low},0}}$ in Eq. (6) contains forces from the unconverged density matrix in addition to the standard Born Oppenheimer gradients and these aspects have been well described in Ref. 48,51. We provide
the resulting gradient expressions here, starting from Eq. (35) in Ref. 51. Note that this expression is identical to Eq. (14) in Ref. 48. (Also see the discussion following Eq. (14) and Eq. (15) in Ref. 48.)

\[ \frac{\partial E_{\text{low}}(0)}{\partial R_{i,j}}|_{P_{\text{low},0}} = \frac{\partial E_{\text{BO}}^{\text{low}}(0)}{\partial R_{i,j}}|_{P_{\text{low},0}} + \text{Tr} \left[ \tilde{P}_{\text{low},0}, F_{\text{low},0} \right] \left( \tilde{Q}_{\text{low},0} \frac{dU_{\text{low},0}}{dR_{i,j}} U_{\text{low},0}^{-1} - \tilde{P}_{\text{low},0} U_{\text{low},0}^{-1} \frac{dU_{\text{low},0}^T}{dR_{i,j}} \right) \]

(A1)

where, in the first equation \( \tilde{Q}_{\text{low},0} = I - \tilde{P}_{\text{low},0} \) and \( \tilde{P}_{\text{low},0} = 3P_{\text{low},0}^2 - 2P_{\text{low},0}^3 \) is the McWeeny purified density matrix\(^{128} \). The quantity, \( U_{\text{low},0} \) transforms the density matrix from the non-orthogonal basis to an orthogonal basis. That is, the non-orthogonal Gaussian basis overlap matrix, \( S'_{\text{low},0} \) is factorized as \( S'_{\text{low},0} \equiv U_{\text{low},0}^T U_{\text{low},0} \). For Löwdin symmetric orthonormalization, \( U_{\text{low},0} \equiv S'_{\text{low},0}^{1/2} \) and for Cholesky decomposition \( U_{\text{low},0} \) is upper triangular. The quantity \( U_{\text{low},0} \) may also represent a singular value decomposition of \( S'_{\text{low},0} \) and in this case the square root of the singular values (\( S'_{\text{low},0} \) is positive semi-definite) is to be absorbed in to the left and right singular vectors of \( S'_{\text{low},0} \) to define \( U_{\text{low},0} \). The quantity, \( F_{\text{low},0} \), is the Fock matrix for the whole system at the low level of theory in an ortho-normal basis and includes the exchange correlation function for density functional implementation. In the non-orthogonal basis, it has the form,

\[ F'_{\text{low},0}^{\nu,\sigma} = h'_{\nu,\sigma} + G'_{\nu,\sigma}(\tilde{P}'_{\text{low},0})_{\nu,\sigma} + \frac{\partial E_{xc}}{\partial P'_{\text{low},0}^{\nu,\sigma}}, \]  

(A2)

where the matrix elements \( \nu, \sigma \), are listed as subscripts and superscripts to allow clarity. Here the one-electron terms are \( h'_{\nu,\sigma} \), two-electron terms are \( G'_{\nu,\sigma}(\tilde{P}'_{\text{low},0})^{\nu,\sigma} \) and \( E_{xc} \) is the exchange correlation function. In addition, \( P'_{\text{low},0} \) is the density matrix in the non-orthogonal basis. That is, \( P_{\text{low},0} \equiv U_{\text{low},0} P'_{\text{low},0} U_{\text{low},0}^T, h = U_{\text{low},0}^{-1} h' \ U_{\text{low},0}^{-1} \), etc. The term \( \frac{\partial E_{\text{BO}}^{\text{low}}(0)}{\partial R_{i,j}}|_{P_{\text{low},0}} \)
in Eq. (A1) represents the standard Born Oppenheimer forces from Ref. 129:

\[
\frac{\partial E_{\text{low}}^{BO}(0)}{\partial \mathbf{R}_{i,j}} \bigg|_{\mathbf{P}_{\text{low},0}} = \text{Tr} \left[ \frac{d\mathbf{h}'}{d\mathbf{R}_{i,j}} \tilde{\mathbf{P}}_{\text{low},0}' + \frac{1}{2} \frac{\partial G_{\text{xc}}'}{\partial \mathbf{R}_{i,j}} \mathbf{P}_{\text{low},0}' \right] - \text{Tr} \left[ \mathbf{F}_{\text{low},0}' \tilde{\mathbf{P}}_{\text{low},0}' \frac{d\mathbf{S}_{\text{low},0}'}{d\mathbf{R}_{i,j}} \tilde{\mathbf{P}}_{\text{low},0}' \right] + \frac{\partial E_{\text{xc}}}{\partial \mathbf{R}_{i,j}} \bigg|_{\mathbf{P}_{\text{low},0}} + \frac{\partial V_{\text{NN}}}{\partial \mathbf{R}_{i,j}}. \tag{A3}
\]

Equation (A1) has been carefully derived and discussed in Refs. 48, 51 but it is clear that additional terms exist in the forces when the density matrix is not fully converged. We may further simplify Eq. (A1) as follows,

\[
\frac{\partial E_{\text{low}}(0)}{\partial \mathbf{R}_{i,j}} \bigg|_{\mathbf{P}_{\text{low},0}} = \frac{\partial E_{\text{low}}^{BO}(0)}{\partial \mathbf{R}_{i,j}} \bigg|_{\mathbf{P}_{\text{low},0}}
\]

\[
\begin{align*}
&= \text{Tr} \left[ \left[ \mathbf{P}_{\text{low},0}' \mathbf{F}_{\text{low},0}' \left( 1 - \mathbf{P}_{\text{low},0}' \right) \frac{d\mathbf{U}_{\text{low},0}}{d\mathbf{R}_{i,j}} \mathbf{U}_{\text{low},0}^{-1} \mathbf{P}_{\text{low},0}' \mathbf{U}_{\text{low},0}^{-T} \frac{d\mathbf{U}_{\text{low},0}^T}{d\mathbf{R}_{i,j}} \right) \right] \\
&= \text{Tr} \left[ \left[ \mathbf{P}_{\text{low},0}' \mathbf{F}_{\text{low},0}' \left( \frac{d\mathbf{U}_{\text{low},0}}{d\mathbf{R}_{i,j}} \mathbf{U}_{\text{low},0}^{-1} - \mathbf{P}_{\text{low},0}' \mathbf{U}_{\text{low},0}^{-T} \frac{d\mathbf{S}_{\text{low},0}}{d\mathbf{R}_{i,j}} \mathbf{U}_{\text{low},0}^{-1} \right) \right] \right]. \tag{A4}
\end{align*}
\]

It is seen from Eq. (A1) that the difference between forces used as a result of the extended Lagrangian propagation and the standard BOMD forces (that were used in Ref. 32) is proportional to the commutator \([\mathbf{P}_{\text{low},0}', \mathbf{F}_{\text{low},0}']\), which is a measure of the convergence of the electronic wavefunction. This is a critical term that arises here due to the non-orthogonality of the basis and is absent in standard implementations of the Car-Parrinello scheme. This commutator has been shown to be related to the norm of the fictitious inertia tensor of the propagated density matrix\(^{51}\), represented as \(\mu_{\text{low},0}\) in Eq. (3). For the cases when the density matrix is converged, the commutator tends to zero, and could be neglected, and in that case the ADMP gradient reduces to the standard BOMD gradient.

The density gradient expression in Eqs. (5) and (7) has also been extensively discussed in Refs. 48, 49 and the corresponding expression for idempotent density matrices is concisely written as

\[
\frac{\partial E_{\text{PIE-ONIOM}}(\mathbf{R}, \mathbf{P}_{\text{low},0})}{\partial \mathbf{P}_{\text{low},0}} \bigg|_{\mathbf{R}} = \frac{\partial E_{\text{low}}(0)}{\partial \mathbf{P}_{\text{low},0}} \bigg|_{\mathbf{R}} = \left[ \mathbf{F}_{\text{low},0}', \mathbf{P}_{\text{low},0}' \right], \tag{A5}
\]
and again this term tends to zero for the converged (BO) cases.

For all simulations in this study, the fictitious inertia tensor for valence electrons are fixed at the scalar value $\mu_{\text{valence}}=180$ a.u., and core orbitals are weighted based on $\mu_{\text{valence}}$ and corresponding diagonal elements of the Fock matrix that determines $E_{\text{low}}(0)$. This setup of fictitious mass has been widely used for many applications and shown to be accurate in reproducing structural and vibrational features.

**APPENDIX B: CONJUGATE HAMILTONIAN FOR THE LAGRANGIAN IN EQUATION (3) AND ASSOCIATED CONSERVATION PROPERTIES**

The conjugate Hamiltonian for the Lagrangian presented in Eq. (3) is given by the Legendre transform:

$$\mathcal{H} = Tr(W_{\text{low},0} W_{\text{low},0}) + Tr(V^T V) - \mathcal{L},$$

where $W_{\text{low},0}$ and $V$ are the conjugate momenta for $P_{\text{low},0}$ and $R$, respectively, and are given by

$$W_{\text{low},0} = \frac{\partial \mathcal{L}}{\partial W_{\text{low},0}} = \mu_{\text{low},0}^{1/2} W_{\text{low},0} \mu_{\text{low},0}^{1/2}$$

and

$$V = \frac{\partial \mathcal{L}}{\partial V} = MV.$$  

Using Eqs. (B2) and (B3) in Eq. (B1), one obtains the conjugate Hamiltonian as

$$\mathcal{H} = \frac{1}{2} Tr[V^T M^{-1} V] + \frac{1}{2} Tr \left[ \left( \mu_{\text{low},0}^{-1/4} W_{\text{low},0} \mu_{\text{low},0}^{-1/4} \right)^2 \right] + E_{\text{PIE-ONIOM}}(R, P_{\text{low},0})$$

$$+ Tr[\Lambda_{\text{low},0}(P_{\text{low},0}^2 - P_{\text{low},0})]$$

It is the conservation properties of this Hamiltonian that are described in Table I. The fact that such a quantity must be conserved is seen from the total derivative of $\mathcal{H}$ with respect
to $t$, i.e.,

$$\frac{dH}{dt} = \text{Tr} \left[ \frac{\partial H}{\partial \mathbf{P}_{\text{low},0}} \frac{d\mathbf{P}_{\text{low},0}}{dt} + \frac{\partial H}{\partial \mathbf{W}_{\text{low},0}} \frac{d\mathbf{W}_{\text{low},0}}{dt} + \frac{\partial H}{\partial \mathbf{R}} \frac{d\mathbf{R}}{dt} + \frac{\partial H}{\partial \mathbf{V}} \frac{d\mathbf{V}}{dt} \right]$$

$$= \text{Tr} \left[ \left( \frac{\partial E^{\text{PIE-ONIOM}}(\mathbf{R}, \mathbf{P}_{\text{low},0})}{\partial \mathbf{P}_{\text{low},0}} \right)_{\mathbf{R}} + \Lambda_{\text{low},0} \mathbf{P}_{\text{low},0} + \mathbf{P}_{\text{low},0} \Lambda - \Lambda_{\text{low},0} \right] \mathbf{W}_{\text{low},0} \right] + \text{Tr} \left[ \mathbf{W}_{\text{low},0} \frac{d^2 \mathbf{P}_{\text{low},0}}{dt^2} \right] + \text{Tr} \left[ \frac{\partial E^{\text{PIE-ONIOM}}(\mathbf{R}, \mathbf{P}_{\text{low},0})}{\partial \mathbf{R}} \mathbf{V} \right] + \text{Tr} \left[ \mathbf{V} \frac{d^2 \mathbf{R}}{dt^2} \right]$$

(B5)

where the definitions in Eqs. (B2) and (B3) have been used and it is also assumed that $\mu_{\text{low},0}$ is time-independent. Now, using the Lagrange equations of motion, i.e., Eqs. (5) and (4), it follows that

$$\frac{dH}{dt} = 0.$$  \quad \text{(B6)}$

---


8 Zhang, D. W.; Zhang, J. Z. H. Molecular Fractionation with Conjugate Caps for Full Quantum


33 Humbel, S.; Sieber, S.; Morokuma, K. The IMOMO Method: Integration of Different Levels of Molecular Orbital Approximations for Geometry Optimization of Large Systems: Test for


55 Herbert, J. M.; Head-Gordon, M. Curvy-Steps Approach to Constraint-Free Extended-Lagrangian Ab Initio Molecular Dynamics, Using Atom-Centered Basis Functions: Conver-


67 Vimal, D.; Pacheco, A. B.; Iyengar, S. S.; Stevens, P. S. Experimental and Ab Initio Dynamical


78 Hartke, B.; Carter, E. A. Ab Initio Molecular-Dynamics with Correlated Molecular Wavefunc-


87 Decornez, H.; Drukker, K.; Hammes-Schiffer, S. Solvation and Hydrogen-Bonding Effects on


100 Headrick, J. M.; Diken, E. G.; Walters, R. S.; Hammer, N. I.; Christie, R. A.; Cui, J.; Myshakin, E. M.; Duncan, M. A.; Johnson, M. A.; Jordan, K. Spectral Signatures of Hy-


Li, X.; Moore, D. T.; Iyengar, S. S. Insights from First Principles Molecular Dynamics Studies Towards Infra-Red Multiple-Photon and Single-Photon Action Spectroscopy: Case Study of


For Table of content use only