An efficient and accurate approach to estimate hybrid functional and large basis set contributions to condensed phase systems and molecule-surface interactions

Timothy C. Ricard and Srinivasan S. Iyengar *

Department of Chemistry and Department of Physics,
Indiana University, 800 E. Kirkwood Ave, Bloomington, IN-47405
(Dated: November 4, 2019)

Abstract

We present a graph theoretic approach to adaptively compute many-body-approximations in an efficient manner to perform accurate hybrid DFT electronic structure calculations for condensed phase systems. The salient features of the approach are ONIOM-like in that, (a) the full system PBC calculation is performed at a lower (pure or semilocal DFT) level of theory, and (b) this approximation is improved through a correction term that captures all many-body interactions up to any given order within a higher (hybrid DFT) level of theory, combined through graph theoretic methods. The approach has similarities in spirit to the range-separated hybrid DFT functionals, but differs in that the range separation strategy is governed by the inherent choice of graphical partitioning of the system. The method is demonstrated for a range of condensed phase problems for both (i) computing hybrid DFT energies for condensed phase systems at pure DFT cost and (ii) also for computing large, triple-zeta, multiply polarized and diffuse atom-centered basis set energies at computational costs commensurate with much smaller sets of basis functions. The methods are demonstrated through calculations on (a) homogeneous water surfaces and heterogeneous surfaces with organic impurities both using two-dimensional periodic boundary conditions, and (b) bulk simulations of water through three-dimensional periodic boundary conditions. A range of structures are considered and in all cases accuracy is obtained in good agreement with higher levels of theory and larger basis sets and computational costs are orders of magnitude lower.

* Email: iyengar@indiana.edu
I. INTRODUCTION

Condensed phase reactive systems are significant to a wide range of chemical and materials problems.\textsuperscript{1–4} There have also been substantial efforts directed towards gauging the reasons behind the relatively high efficiency of organic reactions on the surface of water droplets and interfaces.\textsuperscript{5–10} At the core of all these important studies is the accurate determination of electronic properties for periodic condensed phase systems.\textsuperscript{11–16} There have been several studies\textsuperscript{4,13–18} that attempt to push the envelope of application of forefront electronic structure methods to the condensed phase, including the study of bulk systems,\textsuperscript{19–23} surfaces,\textsuperscript{24–26} and one-dimensional chains.\textsuperscript{27–29} Yet, as of today, there is a disconnect between the quality of electronic structure methods and basis sets that are available for gas-phase, cluster-like and clusters embedded in continuum studies, and those that can be routinely applied to study condensed phase, periodic systems. For example, while there exist serious efforts to provide an hierarchy of accurate density functional approaches for cluster-like studies, with hybrid functionals (the so-called rung-4 functionals\textsuperscript{30–32}) being at the core of many successful studies of chemical interest.\textsuperscript{31} These methods, including the more advanced double-hybrid functionals,\textsuperscript{33–35} become catastrophically expensive when applied to condensed phase systems. Furthermore, there exists a range of high-quality basis functions that are routinely available for the study of isolated systems as well as systems embedded in continuum, yet these become computationally intractable for condensed phase studies. The problem becomes worse for \textit{ab initio} molecular dynamics studies and quantum nuclear studies in the condensed phase\textsuperscript{36} since these calculations quickly become untenable due to scaling costs of both the number and quality of the electronic structure methods including $k$-integration for periodic systems; in fact, these costs are significantly impacted by the use of diffuse and other extended basis functions and non-local exchange-correlation functionals implicit within the higher rung hybrid density functionals.\textsuperscript{16,37–39} As a result, accurate density functionals are very difficult to use for routine periodic system studies, and certainly the accurate modeling of reactivity, adsorption of substrates to surfaces and surface defects may be drastically affected as a result of these limitations.

Apart from computational costs, large diffuse basis functions also suffer from numerical instabilities when $k$-integration\textsuperscript{16} is included. Even-though homogeneous surfaces may not, in general, need large basis functions, surfaces with adsorbates and defects may involve lo-
cal electronic effects that may best be captured through higher quality basis functions. A variety of methods have been developed to alleviate the steep scaling and numerical instabilities in periodic electronic structure calculations, yet condensed phase calculations often involve a trade-off between choice of basis set and computational complexity. While the inclusion of diffuse functions allows a better representation of weak non-bonded interactions, these calculations also often suffer from linear dependencies of basis functions and SCF convergence instabilities when \( k \)-integration is included with accompanying significant increase in numerical cost.

In this publication, we present a new approach that makes it possible to perform accurate electronic structure calculations for extended systems in agreement with hybrid (rung-4) functionals such as B3LYP, CAM-B3LYP, TPSSh, and PBE0 at computational costs that are commensurate with semilocal (rung-2 and rung-3) functionals such as BLYP, PBE, and revTPSS. Furthermore, we also present calculations that are accurate to high quality basis functions for periodic systems such as 6-311++G(2df,pd), that is triple-zeta, doubly polarized and diffused, at numerical effort commensurate with moderate basis sets such as 6-31+G(d), that double-zeta with polarization and diffuse functions. Our approach is based on a graphical decomposition of the molecular structure within a unit cell, where higher level (basis set and density functional) corrections are perturbatively added based on this graphical network to periodic studies at lower levels of theory. This approach is in a similar spirit to ONIOM and to several molecular fragmentation methods previous developed by our group and by other groups. We demonstrate our method for thin films of water/ice, considered due to potential catalytic significance. We consider a range of molecular geometries for each case, and these include both stable (equilibrium) structures as well as those that may be encountered during dynamics studies (and may hence be far from equilibrium) to gauge the accuracy of the proposed methods. At the end, we find the proposed approach to be capable of providing energies in agreement with hybrid functionals (rung-4) and large basis calculations, as noted above, at the accuracy of the order of fractions of kcal/mol at much reduced computational cost.

The paper is organized as follows: In Section II we present our graph-theoretic approach to low-cost condensed phase calculations. In Section III A we discuss the molecular geometries used for our benchmarks that are then discussed in Sections III B and III C. Conclusions are given in Section IV.
FIG. 1: The basic idea of the method arising from Eqs. (1) and (6), is illustrated through a periodic slab of water, with a highlighted cluster. The periodic nature of the problem is captured by $E_{PBC}^{level,0}$. Additional correlation, exchange and large basis set effects are perturbatively included for the highlighted subsystem through $E_\chi$ defined in Eq. (5), using the graphical network depiction of Figure (a) in Figure (b). Although, in this case, the highlighted region coincides with the unit cell, this is not a hard and fast requirement and any subset of critical molecular interactions (that could in principle be larger than the unit cell) may be included as part of $E_\chi$.

II. LOW-COST, ACCURATE TREATMENT OF CONDENSED PHASE ELECTRONIC STRUCTURE THROUGH GRAPH-THEORETIC DECOMPOSITION OF MOLECULAR STRUCTURE

The central hypothesis in this paper is the introduction of long-range, condensed phase, periodic interactions at a lower level of theory (such as semilocal, rung-2, DFT functionals using smaller basis sets), with the perturbative introduction of higher levels of theory (rung-4, hybrid density functionals) and larger basis sets using local many body expansions that are obtained from graph theory. In this manner, local defects, reactive effects, and adsorption properties will be treated at the necessary higher level of electronic structure and larger basis, whereas long-range periodic effects are considered through an ONIOM-like extrapolation procedure. Hence, the central approximation in this work may be captured by an approximation to the overall energy of a periodic system that is given by

$$E_{PBC}^{\text{graph-theoretic}} = E_{PBC}^{level,0} + E_\chi$$

where $E_{PBC}^{level,0}$ refers to the periodic treatment of the system alluded to above, at some lower level of theory and basis together referred to as "level, 0", and $E_\chi$ is the ONIOM-like correction term obtained here from a graph theoretic decomposition of the molecular structure that, as we will show, captures the accuracy at higher levels of theory and basis.
In this paper, the region of chemical interest is represented as a set of geometric nodes, and these nodes are then connected to form a network or a graph. These connections, represented by edges, faces, and other higher order objects referred to as simplexes in the graph, capture local (bonded/non-bonded) as well as non-local (non-bonded) many-body interactions. Such a geometrical decomposition is shown in Figure 1. Furthermore, the nodal structure or graph in Figure 1 has an associated topological invariance known as an Euler characteristic, \( \chi \), defined below in Eq. (2), and guides our definition of \( E_\chi \):

\[
\chi = \eta_0 - \eta_1 + \eta_2 - \cdots + (-1)^r \eta_r + \cdots + (-1)^R \eta_R = \sum_{r=0}^{R} (-1)^r \eta_r.
\]  

(2)

Here \( \eta_r \) is the number of geometric entities (nodes, edges, faces, and higher-order simplexes) of rank-\( r \). That is \( \eta_0 \) is the number of nodes (rank-0 simplexes) in the graph, \( \eta_1 \) is the number of edges (rank-1 simplexes) in the graph, \( \eta_2 \) is the number of faces (rank-2 simplexes) in the graph and so on. \(^{66-73}\) The quantity \( R \) is the largest simplex rank included in the truncated expansion on the right side of Eq. (2), which in general can go up to arbitrary orders.

In this publication, we replace the appearance of each rank-\( r \) simplex in Eq. (2), numbered using the index \( \alpha \), by an energy correction corresponding to the molecular interactions captured by the specific simplex, and thus

\[
\eta_r \to \left\{ \sum_{\alpha} \Delta E(\alpha, r) \left[ \sum_{m=r}^{R} (-1)^m p_{\alpha}^{r,m} \right] \right\}
\]  

(3)

where the summation over \( \alpha \) is over all rank-\( r \) simplexes and \( \Delta E(\alpha, r) \) is an ONIOM-like energy correction or extrapolation term for the \( \alpha \)-th rank-\( r \) simplex:

\[
\Delta E(\alpha, r) = E_{\text{level},1}(\alpha, r) - E_{\text{level},0}(\alpha, r).
\]  

(4)

Thus, in ONIOM-style, the electronic energy for the \( \alpha \)-th rank-\( r \) simplex (representing a molecular fragment in the system) is computed at two levels of theory, namely, “\( \text{level, 1} \)” and “\( \text{level, 0} \)”. (Note that the periodic calculation in Eq. (1) is performed at “\( \text{level, 0} \)”. The square bracketed term in Eq. (3) contains an over counting correction, \(^{56,57}\) where \( p_{\alpha}^{r,m} \) refers to the number of times the \( \alpha \)-th rank-\( r \) simplex appears in all rank-\( m \) simplexes (\( m \geq r \)), with phase \( (-1)^m \). This over-counting correction is along similar lines as those present in
many-body and double many-body expansions\textsuperscript{74–78} and molecular fragmentation\textsuperscript{59,60,63,78–86}.

Correspondingly in Eq. (1) the energy correction $E_\chi$ represents a measure on each simplex. In light of the topological invariant, $\chi$, from Eq. (2) and the energy correction from Eq. (3) the following expression for $E_\chi$ is found:

$$E_\chi \equiv \sum_{r=0}^{R} (-1)^r \left\{ \sum_{\alpha}^{r-\text{rank}} \Delta E(\alpha, r) \left[ \sum_{m=r}^{R} (-1)^m p^{r,m}_\alpha \right] \right\}$$  \hspace{1cm} (5)

The curly bracketed term, $\{\cdots\}$, refers to the energetic replacement for each rank-$r$ simplex, that is, $\eta_r$ in Eq. (2), as per Eq. (3). When this expression is incorporated into the approximation of a periodic system energy given by Eq. (1), the graph-theoretic energy expression\textsuperscript{56} for a periodic system may be written as

$$E_{\text{graph-theoretic}}^{PBC} = E_{PBC}^{\text{level,0}} + \sum_{r=0}^{R} (-1)^r \left\{ \sum_{\alpha}^{r-\text{rank}} \Delta E(\alpha, r) \left[ \sum_{m=r}^{R} (-1)^m p^{r,m}_\alpha \right] \right\}$$  \hspace{1cm} (6)

As noted above, “$\alpha$” represents a simplex\textsuperscript{68–71} of rank-$r$ and $R$ is the largest simplex rank (maximum order of many-body interaction) considered for electronic structure treatment. The availability of larger rank simplexes is highly dependent on the topology of the graph 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. \textit{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 approximations contribute to the energy.}

One may clearly see the connections between Eq. (6) and the well-known many-body expansions (MBE)\textsuperscript{74–78} by simply writing out the appropriate form of Eq. (6) for $R = 1$,

$$E_{PBC,R=1}^{\text{graph-theoretic}} = E_{PBC}^{\text{level,0}} + \sum_{\alpha \in \text{edges}} \Delta E(\alpha, 1) - \sum_{\alpha \in \text{nodes}} \Delta E(\alpha, 0) \left[ p^{0,1}_\alpha - p^{0,0}_\alpha \right]$$  \hspace{1cm} (7)

where now, $p^{0,1}_\alpha$ is the number of times the node $\alpha$ (one-body term) appears in all edges (or two-body interactions) and $p^{0,0}_\alpha$ is the number of times node $\alpha$ appears in all nodes and by
extension $p_{0,0}^\alpha = 1$. Using Eq. (4), may further rewrite Eq. (7) as

$E_{PBC,R=1}^{\text{graph-theoretic}} = E_{PBC}^{\text{level,0}} + \sum_{\alpha \in \text{edges}} \left[ E_{\text{level,1}}^{\alpha} (\alpha, 1) - E_{\text{level,0}}^{\alpha} (\alpha, 1) \right] - \sum_{\alpha \in \text{nodes}} \left[ E_{\text{level,1}}^{\alpha} (\alpha, 0) - E_{\text{level,0}}^{\alpha} (\alpha, 0) \right] (p_{0,1}^\alpha - 1) \quad (8)$

where, as is consistent with the definition in Eq. (3), $p_{0,1}^\alpha$, etc., are the number of times the $\alpha$-th rank-0 simplex occurs in all rank-1 simplexes. Furthermore, in Eq (9), we have clubbed terms belonging to $\text{level,1}$ and $\text{level,0}$ separately. The terms inside each of the curly brackets, $\{ \cdots \}$ are the $\text{level,1}$ (or $\text{level,0}$) two-body energies. These are specifically written as a one-body term which sums over all nodes,

$E_{\text{1-body}}^{\text{level,1}} \equiv \sum_{\alpha \in \text{nodes}} E_{\text{level,1}}^{\alpha} (\alpha, 0) \quad (10)$

and similarly for $\text{level,0}$, and the additional two-body correction term which is captured within the square-bracketed terms, $[\cdots]$, in Eq. (9), that is,

$E_{\text{2-body,corr.}}^{\text{level,1}} \equiv \sum_{\alpha \in \text{edges}} E_{\text{level,1}}^{\alpha} (\alpha, 1) - \sum_{\alpha \in \text{nodes}} p_{0,1}^\alpha E_{\text{level,1}}^{\alpha} (\alpha, 0) \quad (11)$

Thus the two body correction term has the form of energy contribution arising from two-body calculations (edges) minus each one-body contributions times number of times the one-body term (node) appears inside all two-body terms, ($p_{0,1}^\alpha$). Thus, we may rewrite the expression above as

$E_{PBC,R=1}^{\text{graph-theoretic}} = E_{PBC}^{\text{level,0}} + \left\{ E_{\text{1-body}}^{\text{level,1}} + E_{\text{2-body,corr.}}^{\text{level,1}} \right\} - \left\{ E_{\text{1-body}}^{\text{level,0}} + E_{\text{2-body,corr.}}^{\text{level,0}} \right\} \quad (12)$
It is clear that the ONIOM-style extrapolation now applies in Eq. (12) in a many-body context.

Equation (6), includes many-body contributions to arbitrary orders. This can be seen by constructing a similar \( n \)-body analysis for Eq. (6), and for three-body interactions (faces) one may use \( R = 2 \) in Eq. (6), to obtain,

\[
E_{\text{graph-theoretic}}^{P\text{BC}, R=2} = E_{\text{PBC}}^{0} + \sum_{r=0}^{R-2} (-1)^{r} \left\{ \sum_{\alpha} \Delta E_{\alpha}^{r} \left[ \sum_{m=r}^{R-2} (-1)^{m} p_{\alpha}^{r,m} \right] \right\}
\]

\[
= E_{\text{PBC}}^{0} + \left\{ \sum_{\alpha \in \text{nodes}} E^{\text{level},1}_{\alpha,0} + \left[ \sum_{\alpha \in \text{edges}} E^{\text{level},1}_{\alpha,1} - \sum_{\alpha \in \text{nodes}} p_{\alpha}^{0,1} E^{\text{level},1}_{\alpha,0} \right] \right\}
\]

\[
+ \left\{ \sum_{\alpha \in \text{faces}} E^{\text{level},1}_{\alpha,2} - \sum_{\alpha \in \text{edges}} p_{\alpha}^{1,2} E^{\text{level},1}_{\alpha,1} + \sum_{\alpha \in \text{nodes}} p_{\alpha}^{0,2} E^{\text{level},1}_{\alpha,0} \right\}
\]

\[
- \left\{ \sum_{\alpha \in \text{nodes}} E^{\text{level},0}_{\alpha,0} + \left[ \sum_{\alpha \in \text{edges}} E^{\text{level},0}_{\alpha,1} - \sum_{\alpha \in \text{nodes}} p_{\alpha}^{0,1} E^{\text{level},0}_{\alpha,0} \right] \right\}
\]

\[
+ \left\{ \sum_{\alpha \in \text{faces}} E^{\text{level},0}_{\alpha,2} - \sum_{\alpha \in \text{edges}} p_{\alpha}^{1,2} E^{\text{level},0}_{\alpha,1} + \sum_{\alpha \in \text{nodes}} p_{\alpha}^{0,2} E^{\text{level},0}_{\alpha,0} \right\}
\]

\[
= E_{\text{PBC}}^{0} + \left\{ E^{\text{level},1}_{1-\text{body}} + E^{\text{level},1}_{2-\text{body,corr.}} + E^{\text{level},1}_{3-\text{body,corr.}} \right\}
\]

\[
- \left\{ E^{\text{level},0}_{1-\text{body}} + E^{\text{level},0}_{2-\text{body,corr.}} + E^{\text{level},0}_{3-\text{body,corr.}} \right\}
\]

(13)

It is clear from Eqs. (12) and (13) that Eq. (6) provides an ONIOM-type extrapolation approximation ("high" minus "low" or level, 1 minus level, 0) where the "high" and "low" levels are themselves many-body approximations.

\[
E_{\text{PBC}, R}^{\text{graph-theoretic}} = E_{\text{PBC}}^{\text{level},0} + E_{\text{MBE}, R}^{\text{level},1} - E_{\text{MBE}, R}^{\text{level},0}
\]

(14)

where \( E_{\text{MBE}, R}^{\text{level},1} \) (and similarly \( E_{\text{MBE}, R}^{\text{level},0} \)) implies the many-body expansion of order \( R \). Thus, Eq. (6) provide an adaptive recipe to compute interactions to arbitrary order through an efficient graph-theoretic decomposition. The effectiveness of methods involving many body expansions (MBE)\(^{74-78}\) and the related idea of molecular fragmentation\(^{53-63,79-83,87-96}\) has been extensively demonstrated for molecular cluster calculations. These are known to remedy the intractability of higher correlation and extended basis set treatments. Similarly MBE
and fragmentation based methods have been employed for condensed phase calculations. In these methods local quantum mechanical treatments correct long range periodic interactions through embedded charges,\textsuperscript{97–101} intensive molecular mechanics force fields,\textsuperscript{102,103} or condensed phase calculations with HF or DFT treatments.\textsuperscript{104–107} Here we have presented the application of our graph-theoretic interpretation\textsuperscript{56,57} to many-body expansions which aims to reduce the associated costs and instabilities of condensed phase calculations, while presenting a completely adaptive scheme that is suited for \textit{ab initio} dynamics and computing molecular potential surfaces. As noted in Eq. (6), this is done by capturing local interactions within the molecular fragment interactions incorporated into the graphical representation (in this case the unit cell, see Figure 1(a,b)) with higher-order diffuse basis functions, while modeling the bulk properties with lower quality, computationally tractable basis functions. Similarly, we also captured local many-body interactions within the molecular fragments with hybrid and other higher rung DFT functionals, while modeling the long-range, periodic interactions with the more localized “pure” functionals. When considering condensed phase systems, local interactions between the constituents of the unit cell and the periodic, condensed phase are critical aspects to understanding the physics of the system. In this paper, \( R \), the maximum simplex rank, is chosen as one and only nodes and edges are included in the fragmentation work. Hence the equation considered in the results section is Eq. (7) and includes 2-body corrections to exchange, correlation and large basis effects, apart from that already captured within \( E_{PBC}^{\text{level},0} \).

Finally, we also note that there is a very close similarity between the expression in Eq. (6) and our previously developed PIE-ONIOM method\textsuperscript{53–55} approach where the set-theoretic inclusion-exclusion principle\textsuperscript{108} is used to generalize the ONIOM method. The corresponding energy expression for the periodic version is written as

\[
E^{\text{PIE–ONIOM}} = E_{PBC}^{\text{level},0} + \sum_{i=1}^{n} S(i) - \sum_{1 \leq i < j \leq n} S(i \cap j) + \sum_{1 \leq i < j < k \leq n} S(i \cap j \cap k) - \cdots + (-1)^{n-1} \sum_{1 < \cdots < i < \cdots < n} S(1 \cap \cdots \cap i \cap \cdots \cap n) \tag{15}
\]

where

\[
S(i) = E_{PBC}^{\text{level},0}(i) - E_{PBC}^{\text{level},1}(i) \tag{16}
\]
Here, instead of treating the molecular fragmentation process in a graph-theoretic fashion, the molecule is decomposed into overlapping sets and the set-theoretic inclusion exclusion principle\textsuperscript{108} is used in arriving at Eq. (15).\textsuperscript{53–55} Details regarding these connections are discussed in Ref. 56.

III. EFFICIENT EVALUATION OF LARGE BASIS SET APPROXIMATIONS AND RUNG-4 EXCHANGE-CORRELATION FUNCTIONALS FOR CONDENSED PHASE SYSTEMS USING EQ. (7)

The systems considered in this paper are three-dimensional periodic bulk water, two-dimensional periodic thin films of water, and the interaction of adsorbate molecules, H\textsubscript{2}, methane and methanol on the surface of these water films. The accuracy and efficiency of the aforementioned graph-theoretic approximations from Eq. (7) are gauged in obtaining results for such condensed phase problems in agreement with higher levels of DFT (rung-4 functionals) and large, diffused, polarized and split-valence basis-sets.

It has been noted\textsuperscript{109} that exchange and correlation effects, beyond DFT, are critical for the accurate study of water both in the condensed-phase form and in cluster form. Our efforts demonstrated below allow us to efficiently include higher-rung DFT and large basis effects and in future studies include post-Hartree Fock and double hybrid functional\textsuperscript{33–35} effects in condensed phase. All calculations are performed here using the Gaussian series of electronic structure programs.\textsuperscript{110} However, this is not a hard-and-fast requirement and as discussed in Refs. 56 and 57, our implementation of Eq. (7) through an MPI parallelized hybrid C++/Python module is capable of using multiple electronic structure packages within a single AIMD step and single electronic structure calculations. Future versions of this code may allow using plane-wave bases to compute $E_{PBC}^{level,0}$. But here, we use the Gaussian series of electronic structure programs\textsuperscript{110} through atom-centered basis functions, and this allows the flexibility of (a) not requiring periodic boundary conditions to be enforced along the vertical direction for the thin film of water, while still being able to enforce these boundary conditions along the horizontal directions, (b) allowing $k$-point integration capabilities to automatically be included as needed. (See Tables I and II.) Specifically, the $k$-mesh chosen for each system was automatically generated as part of the Gaussian electronic structure calculations.\textsuperscript{110} This was done based on the relative size of the reciprocal space, thus leading to fewer required
TABLE I: Systems and configurations considered as part of benchmark studies

<table>
<thead>
<tr>
<th>Unit Cell Composition</th>
<th>Condensed-phase System</th>
<th>Number of Structures</th>
<th>Energy(^a)</th>
<th>Mean(^b)</th>
<th>(k)(_{\text{graph-theoretic}})</th>
<th>(k)(_{\text{hybrid-DFT}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>((H_2O)_6)</td>
<td>Gas Phase</td>
<td>15</td>
<td>4.3</td>
<td>9.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>((H_2O)_{12})</td>
<td>2D Surface</td>
<td>42</td>
<td>12.9</td>
<td>24.3</td>
<td>71 ± 28</td>
<td>285 ± 108</td>
</tr>
<tr>
<td>((H_2O)_{24})</td>
<td>2D Surface</td>
<td>19</td>
<td>21.0</td>
<td>34.4</td>
<td>43 ± 17</td>
<td>191 ± 54</td>
</tr>
<tr>
<td>((H_2O)_6)</td>
<td>3D Bulk</td>
<td>15</td>
<td>4.8</td>
<td>9.4</td>
<td>228 ± 46</td>
<td>5944 ± 668</td>
</tr>
</tbody>
</table>

\(^a\)Standard deviation of the electronic energies at PBE0/6-31+G(d) (in kcal/mol). Energy distribution shown in Figure 3.

\(^b\)Deviation of the mean energy from the lowest energy configuration within the respective data-sets, at PBE0/6-31+G(d) (in kcal/mol).

\(^c\)The quantity  \(k\)\(_{\text{graph-theoretic}}\) refers to the number of \(k\)\(-\text{space sampling points needed to compute}\ E_{\text{PBC}}^{\text{level,0}}\) in Eqs. (7) and (6). In our study semilocal DFT functionals were used for \(\text{level,0}\) calculations, to extrapolate to \(\text{level,1}\) that was generally hybrid DFT. Hence, in these cases, the choice of semilocal DFT functionals dictate the number of \(k\)-points. Average and Standard deviations are presented.

\(^d\)Number of \(k\)-space sampling points needed during the higher level periodic calculations in section III B. Average and Standard deviations are presented.

FIG. 2: A few structures sampled during the cluster BOMD simulations used to generate the libraries in Tables I and II. The structures shown here are low energy structures seen during dynamics and similar to well-studied hexamer structures.\(^{112,113}\)

\(k\)-points with the increase in lengths of the respective translation vectors.\(^{15,111}\)

A. Construction of structural libraries for condensed phase water systems

To ascertain the accuracy and efficiency of our methods, we considered a range of unit cell and system sizes. In total we considered a total of 180 different structures in the form of thin films of water, bulk water systems obeying three-dimensional periodic boundary conditions, and various adsorbate molecules on the surface of water films. To obtain these structures, our analysis began with the well-studied water hexamer cluster system.\(^{112,113}\) Minimum energy conformers were sampled for this system by creating periodic images of the hexamer
system to provide one family of surface (2 dimensional periodic boundary conditions) and bulk (3 dimensional periodic boundary conditions) data-sets as shown in Table I. Towards this, gas-phase cluster Born-Oppenheimer molecular dynamics trajectories were computed at B3LYP/6-31++g(d,p) level of theory leading to fifteen unique minimum conformers in the gas phase. A few of these structures are shown in Figure 2. The resultant sampled structure population spanned a range of 17 kcal/mol.

The associated stable configurations for condensed phase treatment were obtained through geometry optimization under two-dimensional and three-dimensional periodic boundaries using the BLYP/6-31+G(d) level of theory. Following this, the unit cells were replicated, along the translation vectors, to create larger systems with 12 and 24 water molecules inside the unit cell and these were in turn subject to geometry optimization as outlined above. Thus a library of water surfaces was obtained to conduct our benchmarks as seen from Table I. A few illustrations of the resultant structures are provided in Figure 3. For the pure water systems (those without an adsorbate) the energy distribution of structures is shown in Figure 3(d,e,f) with the energy spread increasing with the unit cell size. The distribution of energies for the configurations represented in Table I are such that the systems with thicker slabs (smaller length translation vectors and hence smaller cross-sectional

FIG. 3: Figures (a)-(c) present sample structures for the three periodic sets of libraries of pure water in Table I. Relative energy distributions for each set of structures are displayed below each sample structure, in Figures (d)-(f).
### TABLE II: Systems and configurations considered for adsorbate systems

<table>
<thead>
<tr>
<th>Unit Cell Composition</th>
<th>Condensed-phase System with adsorbate</th>
<th>Number of Structures</th>
<th>Energy(^a) Spread</th>
<th>Mean(^b) k-graph-theoretic (^c) k-hybrid-DFT (^d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>((H_2O)_{24} \cdot H_2)</td>
<td>2D Surface with adsorbate</td>
<td>37 (13)(^e)</td>
<td>31.0</td>
<td>34.6</td>
</tr>
<tr>
<td>((H_2O)_{24} \cdot CH_4)</td>
<td>2D Surface with adsorbate</td>
<td>45 (13)(^e)</td>
<td>25.2</td>
<td>25.6</td>
</tr>
<tr>
<td>((H_2O)_{24} \cdot CH_3OH)</td>
<td>2D Surface with adsorbate</td>
<td>22 (13)(^e)</td>
<td>32.4</td>
<td>34.5</td>
</tr>
</tbody>
</table>

\(^a\)Standard deviation of the electronic energies at PBE0/6-31+G(d) (in kcal/mol).

\(^b\)Deviation of the mean energy from the lowest energy configuration within the respective data-sets, at PBE0/6-31+G(d) (in kcal/mol).

\(^c\)Number of k-space sampling points needed for \(E_{PBC}^{level,0}\). In our study, semilocal DFT functionals were used for \(level,0\) and hence dictate the number of k-points in this case. Average and Standard deviations are presented. No deviations in k-points are seen here since the surfaces chosen have similar translation vectors. As a result, \(k_{\text{graph-theoretic}}\) is the same for all structures considered in this library.

\(^d\)Number of k-space sampling points needed during the higher level periodic calculations in section III C. Average and Standard deviations are noted. Small deviation in k-points are seen here since the surfaces chosen have similar translation vectors. As a result \(k_{\text{hybrid-DFT}}\) is similar for the structures considered in the library.

\(^e\)Number of structures sampled from AIMD are shown within parenthesis.

Area in two-dimensions) are lower in energy and in turn need more k-points. This is also responsible for the distribution of k-points represented in Table I. The diversity of structures considered in our benchmarks, as signified by Figure 3, aligns well with our eventual goal of constructing accurate AIMD trajectories\(^{53-57}\) and potential energy surfaces,\(^{53}\) for quantum nuclear effects, using this methodology. Furthermore, it is also clear from Table I that far fewer number of k-points are necessary in computing \(E_{PBC}^{level,0}\), when \(level,0\) uses semilocal DFT functionals, as compared to the full system using the higher level of (hybrid) DFT functional. This disparity in required k-points is due to the lattice summations for non-local Hartree-Fock exchange interactions converging slowly with distance within extended systems.\(^{114}\) This also greatly influences the computational advantages of the graph-theoretic method as seen in later in the paper. This is already on display in Table I where the average number of k-points decrease with the number of water molecules. The number of k-points considered significantly increased the overall computational costs of the treatment, especially with hybrid functionals. Thus the fact that the lower energy, thicker films, require larger computational resources (k-points), implies a greater potential computational advantage from the use Eq. 7.
For the heterogeneous adsorbate systems, we positioned the adsorbate molecule (H\textsubscript{2}, CH\textsubscript{4} and CH\textsubscript{3}OH were considered) on two of the previously optimized water-film interface structures already captured as part of Table I. For each case, the adsorbate molecule was initially placed on a grid parallel to the water interface, approximately 4 Å above the surface. The resultant benchmark structures, reflected in Table II, were obtained through geometry optimization using the revTPSS/6-31+G(d) level of theory. Next isochoric, constant unit cell volume, PBC-AIMD trajectories were initiated at the PBE/6-31+g(d) level of theory for each type of system, and thirteen low energy conformations encountered in each dynamics trajectory were selected to expand our structural library. The number of structures sampled from AIMD are indicated within parenthesis in the third column of Table II, whereas the total number of structures displayed outside the parenthesis in the third column. In contrast to the distribution of energies for the “pure” water systems, the systems with adsorbate exhibited a bimodal distribution. This was on account of the presence of both distinctly lower energy configurations obtained from geometry optimization, and structures obtained during the AIMD studies. The structures from AIMD were higher in energy and well separated from the optimized geometries. Sample structures for the adsorbate-on-water system are discussed in Section III C.

**B. Error analysis for energies obtained using Eq. (7) for thin-films, adsorption interfaces and three-dimensional condensed phase systems**

To apply Eq. (7) to the condensed phase systems discussed above, the unit cell used for the periodic calculations is represented as a graph. (See Figures 1 and 4.) There are two specific control parameters in Eq. (6) that influence the construction of the graph, the many-body approximation and the corresponding required set of electronic structure calculations. One pertains to the choice of “R” and the other relates to the spatial envelope that assembles the local family of edges, or two-body interactions, around any given node. The latter defines the spatial neighborhood used to determine the set of \{α\} for each “r” in Eq. (6). (Note that once the nodes and edge are defined, the graph is completely defined and the expression in Eq. (6) can then be truncated at any order, \(R\).) In this paper, \(R\), the maximum simplex rank, is chosen as one and only nodes and edges are included in the fragmentation work. Three-body interactions (\(R=2\)) were found to have minimal effect on the quality of
FIG. 4: A periodic thin film of water with a unit cell of 24 water molecules is represented in Figure (a), where the hydrogen bonds are depicted with dashed lines. The unit cell region is treated as a graph (Figure (b)) where the edges are constructed according to the prescription described in Section III B. This allows the system to be treated by Eq. (1) treating the periodic character of the system ($E_{\text{level},0}^{\text{PBC}}$) by a standard, affordable method while exchange-correlation and/or larger basis set corrections are added by the N-body terms arising from the simplex elements of the graph.

results for the systems considered here and hence are not included in the discussion. Future publications will evaluate this further for more complex non-bonded interactions such as those involving metal centers. Thus, as noted previously, the equation being used here is the truncated two-body version: Eq. (7). To define the spatial envelope (to obtain the set of edges and hence the graph), the following established\textsuperscript{53} protocol is followed. Each water molecule is designated as a node within the graph. Following this, for every water (node) the closest water molecule defines an edge created between the respective water monomer nodes. Additional edges are introduced between water molecules that are within a distance 10% greater than edges that already exist between the water molecules. (The graphical decomposition created for the adsorbate system is slightly different as discussed in Section III C. In that case, a Cartesian distance cutoff was employed for edge generation.) Hydrogen atoms within a fixed distance of 1.4 Å from an oxygen are included within the definition of a node specified by the oxygen atom. In this manner a local interaction neighborhood is created for computing $E_\chi$ in Eq. (6), the part of the formalism that captures the many-body interaction in the system up to order $R$. This neighborhood graph creation algorithm is implemented using Python functions called from our C++ module which allows the automated formation of these graphs and the associated molecular fragments up to order $R$ in Eq. (6). (See Eqs. (12) and (13) regarding connections to many-body expansions.) The graphs thus obtained, are also consistent with the $\eta = 2$-networks from Refs. 56 and 57 where hydrogen bonds act as the connectivity criteria. This connection between hydrogen bonds and edges
can be seen by fact that the hydrogen bonds depicted in Figure 4(a) directly correspond the edges in Figure 4(b).

The use of Eqs. (6) and (7) implies that the extended periodic system is treated at a choice of theory and basis (level, 0, leading to $E_{PBC}^{level,0}$) that bears acceptable cost and numerical stability for the chosen systems; in this paper we chose DFT with semi-local exchange functionals with minimal diffuse and polarized basis functions such as PBE/6-31+G(d) and this choice is responsible for the reported number of $k$-points in Tables I and II. Next, the perturbative correction of Eqs. (6) and (7), in light of Eq. (4), lead to simplexes that are treated at both the desired theory and basis (level, 1), and the periodic, condensed phase form of the cheaper treatment (level, 0), consistent with ONIOM. This aspect may also be clear from Eq. (14). In this paper the desired accuracy (level, 1) was chosen to be (a) hybrid density functionals with a modest basis (Section III B 1), and (b) semilocal-DFT and screened Coulomb hybrid-GGA, density functionals with large basis treatments (Section III B 2). Thus, we present efficient, accurate and stable periodic calculations at higher levels of density functional and basis set approximations.

The accuracy of this method is quantified from errors in lattice energy across a variety of stable and unstable structures noted in Tables I and II. We compute errors for both the energies obtained from Eq. (7) and from using the lower level of theory as an approximation to full periodic calculations at higher levels. These errors are designated as $\epsilon_{\text{Lattice}}$ and computed according to

$$
\epsilon_{\text{Lattice}} = \frac{1}{N_{\text{structures}}} \sum_{i} \left| \frac{1}{N_{\text{monomer}}} \left( E_{PBC}^{level,1}(i) - \sum_{j=1}^{N_{\text{monomer}}} E_{\text{monomer}}^{level,1}(j) \right) - \left( E_{PBC}^{b}(i) - \sum_{j=1}^{N_{\text{monomer}}} E_{\text{monomer}}^{b}(j) \right) \right| \tag{17}
$$

where $N_{\text{monomer}}$ is the number of chemically distinct monomers and the summation over $N_{\text{monomer}}$ is explicitly specified to allow treatment of heterogeneous systems such as those discussed later for the adsorbate-surface interaction problem. The reported error in lattice energy ($\epsilon_{\text{Lattice}}$) is the mean absolute difference between the lattice energies for the full periodic target treatment, $E_{PBC}^{level,1}$, and the method referred to as “b” in Eq. (17), which would be either the graphical extrapolation based on Eq. (7), that is, $E_{\text{graph-theor.}}^{PBC,R=1}$, or the baseline full periodic low level treatment $E_{PBC}^{level,0}$. The latter is used in the benchmark studies.
FIG. 5: Here the accuracy of the energy expression in Eq. (7) is evaluated using Eq. (17) for the films of water (two-dimensional periodic boundary conditions) with a unit cell of size 24 water molecules (Figure 1). In Figure (a) the target (rung-4) functionals are presented along the horizontal axis and levels used to compute $E_{level, 0}$ (which dominates the computational effort) are shown on the legend. Clearly the lattice energy errors are fractions of kcal/mol. The baseline comparisons (Figure (b)) of the chosen lower levels of theory with the high levels of theory show lattice energy error of the order of 1-2 kcals/mol, with some exceptions discussed in the text. Very similar results were recovered for films with unit cell sizes containing 12 water molecules. The computational gain from using Eq. (7) is presented in Figure 6.

1. Efficient and accurate Hybrid (rung-4) DFT for condensed phase

Hybrid density functional methods include a fraction of Hartree-Fock exchange, and are computationally expensive for periodic calculations due to their spatially non-local nature. The lower rungs of semilocal DFT functionals do not suffer from this computational challenge, but are considered to be less accurate in evaluating hydrogen-
FIG. 6: Computational gain from using Eq. (7) for the surface calculations. Computational Gain is defined as ratio of the serialized computation time for calculations performed at the higher level of theory on the systems with a unit cell size of 24 waters and those obtained using Eq. (7). A clear factor of 2X-3X in computational improvement is noted for these isotropic surfaces. The approach is found to be more efficient for the bulk system (Figure 7) and for large basis calculation (Figures 9).

Equation (7) partially alleviates this by obtaining approximations to higher rung functionals at a reduced cost. This approach is employed on the thin-film (Figure 5), bulk water (Figure 7) and the surface adsorption system (Figure 12, in the next subsection). Our methodology shares similar objectives as range-specified “screened” exchange functionals. But here, the extent of locality captured by the Hartree-Fock exchange may be tailored adaptively and made spatially dependent using graph-theory. Specifically this may be done through (a) the inclusion of faces and higher order simplexes as allowed by Eq. (6), and (b) the inclusion of (i) a position, or node dependent spatial envelope that defines the edges, and hence the graph, as outlined at the beginning of Section III B, and (ii) the value of maximum simplex rank, $R$, in Eq. (6).

But before we proceed with our analysis of errors, we first note that Eq. (6) is computed here using two different levels of theory and the goal is to check accuracy and efficiency in computing condensed phase properties at higher levels of theory. The following notation best describes the computational recipes used here: B3LYP:BLYP implies that the $E_{PBC}^{level,0}$ term in Eq. (6) and $E^{level,0}(\alpha, r)$ in Eq. (4) are computed at the BLYP level of theory whereas the target calculation is B3LYP. Thus $E^{level,1}(\alpha, r)$ in Eq. (4) is obtained from B3LYP. Note that $E_{PBC}^{level,0}$, $E^{level,0}(\alpha, r)$ and $E^{level,1}(\alpha, r)$ are required in Eqs. (6) and (7) to obtain the approximate energy at the higher level of theory. Similarly, BLYP/6-
FIG. 7: Similar to Figures 5 and 6, but for bulk water. Figure (a): Each of the target hybrid functionals are extrapolated to within a fraction of a kcal/mol with an appropriately chosen semi-local DFT method. The efficiency of these extrapolations leads to roughly a factor of 15 speed-up, as seen in Figure (c). Baseline comparisons with level, 0 PBC calculations are presented in Figure (b).

311++G(2df,pd):6-31+G(d) accordingly indicates the level of theory and basis sets involved in computing $E^{level,0}_{PBC}$, $E^{level,0}(\alpha, r)$ and $E^{level,1}(\alpha, r)$ and the associated target basis set quality. For the computational implementation utilized here, the bottleneck arises from the $E^{level,0}_{PBC}$ calculation. As a result, to provide a baseline error analysis, we present the mean absolute lattice energy errors between the low (level, 0) and high (level, 1) levels of theory in Figures 5(b) and 7(b). In Figures 5(a) and 7(a), the chosen low levels of theory (BLYP, PBE, and revTPSS) varied in how well they recovered the higher level (TPSSh, PBE0, B3LYP, and CAM-B3LYP), and our next goal is to check if this is improved upon by using the methods presented above. A consistent basis set was used, 6-31+G(d), for the calculations in Figures 5-7, meaning only the density functionals are modified while the basis set treatment remains constant.

Equation (17) is employed to compare the hybrid DFT lattice energy results to that of both the lattice energies from the low level baseline semilocal functionals and from the use of Eq. (7) to extrapolate from those low level functionals. When we consult the baseline results in Figures 5(b) and 7(b), we note that TPSSh lattice results are already well captured by the modified version of its parent 3rd rung functional TPSS. Furthermore, the lattice energies from CAM-B3LYP are surprisingly well captured by the 2nd rung PBE functional. Meanwhile PBE0 and B3LYP are less well represented by these lower rung functionals, but PBE0 appears to be nicely approximated by its parent functional PBE, while B3LYP is moderately matched by revTPSS. The next question pertains to the extent of improvement in accuracy when using Eq. (7). Overall the use of this edge based (two-body) extrapolation...
tion demonstrates significant improvement in obtaining the target hybrid functional lattice energies, with the notable exceptions of TPSSh:revTPSS and CAM-B3LYP:PBE as the low-level calculations reproduce the calculated lattice energies reasonably well. The best extrapolations tested here were PBE0:PBE and B3LYP:revTPSS where the low-level already demonstrated a fair degree of accuracy in replicating the target energies, but with the edge corrections from Eq. (7) remarkable recovery of the lattice energy is achieved in both surface (Figures 5(a)) and bulk water (Figure 7(a)) calculations at much reduced costs. This reduction in cost is quantified in Figures 6 and 7(c) where the ratio of CPU time cost of the target functional over the computational time cost of the use of Eq. (7). For the water film systems the reduction in computational cost by more than a third, while for the three-dimensional bulk system, we see a roughly 15 times reduction in computational times. This manner of reporting does not consider the run-time improvement gained through parallel treatment of the individual component calculations.

In summary we are able to achieve sub-kcal/mol accuracy for higher levels of density functional methods at a fraction of computational cost. It is especially interesting that regardless of choice of lower level of theory, the accuracy in lattice energy is high.

2. Efficient and Accurate basis set extrapolation with periodic systems

When using non-orthogonal atom-centered Gaussian basis sets in condensed phase, periodic simulations, the electronic basis functions are transformed into orthogonal and translationally invariant functions, known as “Bloch functions”. These are also often referred to as “crystalline orbitals”. But when highly diffuse atom-centered basis functions are chosen, significant density may be found either (a) outside of the periodic cell, or (b) in regions with limited atomic density, where the latter leads to instabilities in the SCF procedure. Furthermore, from using Ewald summation and from fast multipole-type approximations, 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, larger basis sets are often needed to represent weak interactions that are present in catalytic problems. To achieve stable and efficient basis set treatments of condensed phase reactive systems, we employ Eq. 7 as done in Ref. 57. By use of the
FIG. 8: Here the accuracy of the energy expression in Eq. (7) is evaluated using Eq. (17) for large basis calculations with two-dimensional periodic boundary conditions. The unit cell size of 24 water molecules. The target basis and the basis-sets used to compute $E_{PBC, R=1}$ are presented along the horizontal axis while the functionals are shown on the legend. Clearly the lattice energy errors from the graph-theoretic formalism (Figure (a)) are fractions of kcal/mol. The baseline comparisons (Figures (b)) for the chosen smaller basis set, 6-31+G(d), with the larger basis sets for the film of water molecules shows a mean absolute error on the order of 1-3 kcal/mol per water. (Similar results were found for the 12-water unit cell.) The computational gain from using Eq. (7) is roughly an order of magnitude and is reported in Figure 9.

graph-theoretic treatment discussed above, we are able to obtain large basis contributions to local interactions, where the word local here is defined from the graphical decomposition perspective in Eq. (6) and is quantified by (a) the size of $R$ and (b) the density of local nodal connectivity through edges. In this study, we only consider up to two-body based local basis-set space expansion (see Eq. (14)) which are represented as edges in the graphical representation of the system. (The effect of three-body interactions in these studies was found to be minimal.) This is done in conjunction with treating the condensed phase portion, $E_{PBC, level, 0}$ in Eq. (6), using less intensive basis sets to reduce cost and increase stability with periodic $k$-cell integration. (See Tables I and II.) Here we provide a numerical tool to compute properties accurate to highly diffuse Pople-type basis sets (such as 6-311++G(2df,2dp)) at costs commensurate with minimal Pople-type basis sets such as 6-31+G(d).\textsuperscript{124} As we demonstrate below, this greatly reduces the computational cost and is hence a key step in pushing the realm of possible accurate condensed phase simulations.

Graph-theory based basis set computations were performed on the thin-films for multiple DFT methods and these are shown in Figure 8. Through computational effort commensu-
FIG. 9: The approximate order of magnitude gain in computational time through the use of Eq. (7) for surface calculations: The gain is computed with respect to the same calculations performed at the higher level of theory on the systems with a unit cell size of 24 water molecules. (Also see Figures 8.)

FIG. 10: Similar to Figures 8 and 9, but for bulk water. The use of Eq. (7), in Figure (a), shows a significant gain in accuracy over the baseline studies shown in Figure (b). While the errors remain under a kcal/mol upon use of Eq. (7), the computational efficiency is substantially greater and shows a 30X-50X reduction in cost as seen in Figure (c).

rate with 6-31+G(d), we are able to obtain accuracy comparable to multiple diffuse basis sets with accuracy in lattice energy in the sub-kcal/mol range. (See Figures 8 and 9.) The lattice energy errors remain consistent across irrespective of choice of DFT functional which is an attestation to the robustness of this approach and shows that the method is transferable across DFT methods. Furthermore, as noted in Figure 9, the graph-theoretic procedure provides a 6X-25X reduction in computational time as compared to PBC calculations using large basis sets on water surfaces. Bulk water systems, shown in Figure 10, have also been
FIG. 11: The adsorbates considered in this study are H$_2$ (Figures (a) and (b)), CH$_4$ (Figures (c) and (d)), and CH$_3$OH (Figures (e) and (f)). The graphical representations, displayed in Figures (b,d,f), are localized to the unit cell in this study (although this is not a requirement) and used for the evaluation of Eq. (7). For the case of methanol (Figures e,f) it is clearly seen that the graphical decomposition considers several non-bonded interactions.

treated in a similar manner. Here again we observe sub-kcal/mol accuracy, with a computational cost reduction of the order of 25X-50X as compared to the large basis full system calculations. In summary, we have demonstrated the accuracy and significant efficiency of our graph-theoretic procedure to capture many-body interactions on surfaces and condensed phase. This is the case for both accurate density functional computations and for large basis set calculations.
FIG. 12: Here the accuracy of the energy expression of Eq. (7) is evaluated using Eq. (17) for a film of water with one solute molecule on its surface. The unit cell consists of 24 water molecules and one solute molecule (H₂, CH₄ or CH₃OH) as in Figure 11. The target (rung-4) functionals are presented along the horizontal axis, while the solute molecule is presented above the histogram plots on the left. The use of Eq. (7) provides accurate lattice energies as seen from the figures on the left. These improvements are clarified by providing baseline comparisons with the respective lower level functionals on the right side.

C. Error analysis for the study of organic impurities on the surface of water

Next we consider organic impurities H₂, methane and methanol, adsorbed on a film of water. The goal here is to gauge the effectiveness of the scheme tested above for the study of heterogeneous systems that may be of significance in chemical catalysis. The structures considered here are obtained from both geometry optimization calculations and from sampling AIMD trajectories of systems where the unit cell is chosen to contain 24 water molecules interacting with a single adsorbate molecule as described in Section III. The graphical representation of the system (Figure 11) is constructed along a similar vein as that described on beginning of Section III B, but in this case, the adsorbed impurity is
FIG. 13: The computational gain through the use of Eq. (7). The gain is computed with respect to the same calculations performed at the higher level of theory on the water films with solutes. The semilocal DFT to hybrid DFT extrapolations demonstrated a gain of about a factor of 2 to 3.

treated as a single additional node that forms edges with all water molecules that are within a distance of 5.5 Å of its geometric center. The graphical representation capturing water-water interactions was maintained along similar lines as those described in Section III B. Figure 11 provides a representative set of graphical decompositions. Due to the presence of the impurity on the surface, graphical faces (representing three-body interactions) involving two water molecules and a solute molecule are possible. However, we find that these $R=2$ terms from Eq. (6) provides only a minor improvement in accuracy over $R=1$ for the impurity on the surface. Hence these terms are not considered in this work, but left for future investigations into systems where these effects may become more relevant.

The set of DFT extrapolations tested in the previous subsection are also considered here for the solute-water-surface interaction, and the results are shown in Figure 12. Very similar trends are observed for the pure water film and bulk systems from Figures 5 and 7. As before, the most beneficial extrapolations are PBE0:PBE and B3LYP:revTPSS. These calculations show an approximate 2X-3X reduction in cost as compared to the full system with increase in efficiency with size of adsorbed solute as seen in Figure 13.

The basis set extrapolations (Figure 14) show lattice energy errors of less than 0.5 kcal/mol for all five functionals, which is a significant improvement when compared with the baseline results shown in Figure 14. As noted above, these results hold across a variety of functionals. A significant cost reduction is seen in Figure 15 with increasing relative efficiency with target basis set size and through consideration of screened hybrid-GGA functionals HISS and HSE.

In summary we have demonstrated that the methods we have developed on pure water
FIG. 14: Similar to Figure 12 but for basis set extrapolation. As before we choose the basis set 6-31+G(d) as the lower level and these are used within the graph-theoretic formalism in Eq. (7) to provide results for higher level basis sets presented along the horizontal axis. The solute molecule for each set of systems is presented above the histograms on the left. The graph-theoretic basis set extrapolations (left) consistently show accuracy to within 0.5 kcal/mol in lattice energy while the smaller basis baseline (right) shows nearly four times the error. The computational gain is large as shown in Figure 15.

FIG. 15: The computational gain through the use of Eq. (7). The gain is computed with respect to the same calculations performed at the higher level of theory on the water films with solutes. Basis set extrapolation demonstrates a gain of roughly a factor of 5-10 as the size of the solute studied increased.
condensed-phase system in section III B are equally applicable to heterogeneous systems such as adsorbates on the surface of water. This application also demonstrates an increase in relative efficiency over the target calculations allowing future studies on complex heterogeneous systems with greater efficiency and stability. Our future goal is to apply these methods to organic reactions on the surface of water.

IV. CONCLUSION

In this paper, we have discussed the implementation of our graph-theoretic extrapolation method for accurate condensed phase electronic structure studies on surfaces and bulk systems. This was achieved by partitioning a chosen cluster within the condensed-phase system into vertices or graphical nodes, that are then connected through edges to form a geometric graph representation of local many-body interactions. Here, the unit cell from the condensed-phase was chosen to be the cluster for graphical partitioning. This graph-theoretic partitioning scheme then provides a perturbative correction to an affordable periodic electronic structure method leading to improved accuracy and efficiency in functional and basis-set extrapolation. Specifically, this paper demonstrates the graphical extrapolation procedure by providing results where (a) semilocal DFT results may be used in conjunction with the above mentioned perturbative corrections based on graph-theory to yield hybrid DFT condensed phase accuracy, at semilocal DFT cost, and (b) modest sized basis set treatment may also be used in conjunction with the same graphical perturbations above leading to large basis set accuracy, at reduced cost. Sub-kcal/mol lattice energy errors are shown to be achievable at significant computational cost reduction for both flavors of extrapolation.

It was also found that the basis set extrapolation results appear to be transferable across DFT functionals with about a tenth of the computational cost and less than 0.5 kcal/mol lattice energy error. For the DFT functional extrapolations, the hybrid functionals PBE0 and B3LYP were well described at costs commensurate with PBE and revTPSS. The approach is demonstrated by studying both homogeneous and heterogeneous systems. Specifically interactions involving water in the condensed phase and those with organic solutes have been studied. The method appears to be promising for the study of organic reactions on the surface of water.
V. ACKNOWLEDGMENTS

This research is supported by the National Science Foundation grant NSF CHE-1665336 to SSI.


53 Li, J.; Iyengar, S. S. Ab initio Molecular Dynamics using Recursive, Spatially Separated, Overlapping Model Subsystems Mixed Within an ONIOM Based Fragmentation Energy Ex-


Beran, G. J. A new era for ab initio molecular crystal lattice energy prediction. Angew. Chem. 2015, 54, 396.


Tsuzuki, S.; Orita, H.; Honda, K.; Mikami, M. First-principles lattice energy calculation of urea and hexamine crystals by a combination of periodic DFT and MP2 two-body interaction


Graph-theoretic method for condensed-phase systems

\[ E_{\text{PBC}}^{\text{Graph}} = E_{\text{PBC}}^{\text{lvl.0}} + \sum_r (-1)^r \left\{ \sum_\alpha \Delta E_{(\alpha, r)} \right\} \sum_m (-1)^m p_{r,m} \]