JCTC Journal of Chemical Theory and Computation

Using Molecular Similarity to Develop Reliable Models of Chemical Reactions in Complex Environments

Volkan Ediz, Anthony C. Monda, Robert P. Brown, and David J. Yaron*

Department of Chemistry, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213

Received August 10, 2009

Abstract: The use of molecular similarity to develop reliable low-cost quantum mechanical models for use in quantum mechanical/molecular mechanical simulations of chemical reactions is explored, using the $H + HF \rightarrow H_2 + F$ collinear reaction as a test case. The approach first generates detailed quantum chemical data for the reaction center in geometries and electrostatic environments that span those expected to arise during the molecular dynamics simulations. For each geometry and environment, both high- and low-level ab initio calculations are performed. A model is then developed to predict the high-level results using only inputs generated from the low-level theory. The inputs used here are based on principal component analysis of the low-level distributed multipoles, and the model is a simple linear regression. The distributed multipoles are monopoles, dipoles, and quadrupoles at each atomic center, and they summarize the electronic distribution in a manner that is comparable across basis set. The error in the model is dominated by extrapolation from small to large basis sets, with extrapolation from uncorrelated to correlated methods contributing much less error. A single regression can be used to make predictions for a range of reaction-center geometries and environments. For the trial collinear reaction, separate regressions were developed for the transition region and the entrance and exit channels. These models can predict the results of CCSD(T)/cc-pVTZ computations from HF/3-21G distributed multipoles, with an average error for the reaction energy profile of 0.69 kcal/mol.

1. Introduction

Quantum chemistry has made great strides in developing highly accurate methods for computing the properties of small molecules, but application to large molecules remains challenging due to the rapid increase in computational cost with system size. One means of reducing cost is to restrict the full quantum description to a small locus, the reaction center, of a larger system. This approach is used in quantum mechanics/molecular mechanics (QM/MM) methods, where QM is used to describe a handful of atoms in the reaction center, while MM is used for the thousands of atoms in the remainder of the system (e.g., protein and solvent).[1−4] In such simulations, the QM algorithm is typically called millions of times to generate the energy, forces, and charge distribution of the reaction center in the presence of electrostatic interactions with the MM environment.[4] Despite the use of QM methods only in the reaction center, the QM computations are often the bottleneck in such simulations. It is this high cost of QM that this work seeks to substantially reduce, thus further expanding the reach of QM/MM models to the large and complex systems of relevance in biological and materials applications.[4−9] The computational savings of QM/MM stem from the local nature of chemistry, such that QM can be used on only a small locus of the system. Here, we explore the use of molecular similarity to further reduce the computational cost. We begin by performing high-level quantum computations on the reaction center in a range of environments that span those expected to arise in the QM/MM simulation. This data is then mined for a low cost model that can describe the reaction center in similar environments.

The QM portion of a QM/MM simulation is defined by the boundary with the MM region and by the level of electron

* Corresponding author. E-mail: yaron@cmu.edu.

10.1021/ct9004195 CCC: $40.75
Published on Web 11/19/2009

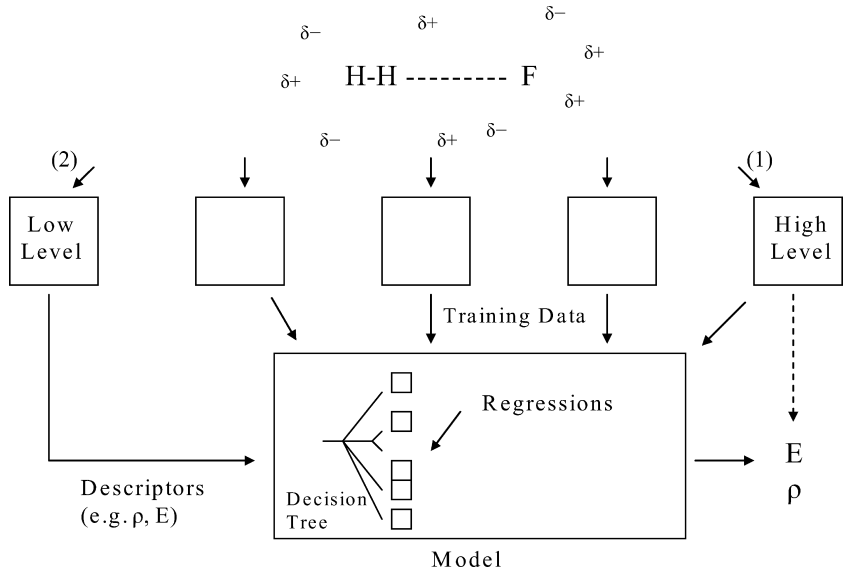

Figure 1. Scheme for the development and the use of a model of the chemical reaction H + HF → H_2 + F. The middle boxes represent various levels of quantum chemical calculations with accuracies and costs that increase from left to right. The bottom box represents a model that maps from the energy (*E*) and the charge distribution (ρ) of a low-level model to that of the high-level model. The model may use different regressions, selected according to region along the reaction coordinate or other criteria.

structure theory used in the reaction center. A wide variety of quantum chemical methods (e.g., semiempirical and ab initio calculations) are now available with differing reliabilities and with CPU times ranging from seconds to days.[10,11] The level of computation needed to achieve a certain target accuracy varies widely, both with system type and with position along the reaction coordinate. A challenging aspect of chemical reactions is the need to obtain an accurate description of the energy and the configuration of the transition-state (TS), since the multireference character of the TS region often requires QM methods with the highest computational cost.[12] The formal scaling of computational effort for ab initio calculations on a N-electron system ranges from O (N^3) for Hartree−Fock theory to O (N^5) for MP2 and O (e^N) for the exact full-configuration interaction (full-CI) solution, making the most accurate and reliable methods difficult to apply in QM/MM simulations where the QM calculations must be called at each time step of a molecular dynamics (MD) trajectory.[13]

Many approaches have been developed to reduce the computational expense of ab initio calculations.[14−24] Generally, these attempts take advantage of two common features of molecular systems: nearsightedness and molecular similarity.

Nearsightedness relates to the local character of the interactions present in a molecule, such that interactions occurring on long length scales can be simplified to interactions between electrostatic multipoles and van der Waals forces.[25] Linear-scaling methods aim to take advantage of this local character wherever possible in a quantum chemical calculation.[26] For instance, divide-and-conquer methods reduce the computational cost of self-consistent field calculations[14,27] or correlated calculations,[28] while fast multipole methods accelerate the computation of Coulomb interactions.[16,29]

Molecular similarity relates to the tendency of molecular fragments, such as functional groups, to behave similarly in different molecules and environments. The assumption of molecular similarity underlies the use of atom- or functional-group-specific parameters in semiempirical quantum chemistry and molecular mechanics.[21,22,30] For instance, in semiempirical methods, the ab initio Hamiltonian is replaced with a simpler model Hamiltonian, which is parametrized either to experimental[21] or ab initio data.[31−34] Similarly, force fields in MM are parametrized using both experimental and theoretical data regarding the structure and the interactions between functional groups. Both methodologies lead to substantially lower computational costs than ab initio calculations, however, with a loss in accuracy that may limit their applicability.[4,35,36]

Here, we explore the use of molecular similarity to develop models for use in QM/MM calculations of chemical reactions that have substantially lower cost than ab initio calculations and that have controllable accuracy and reliability. Our approach first generates detailed quantum chemical data of the reaction center in configurations and electrostatic environments that span those expected to arise during the MD trajectory. For each configuration and environment, both high- and low-level ab initio calculations are performed. These data are then analyzed to develop a low-cost model that can, given only the output of a low-level ab initio calculation, predict the output of the high-level calculation. Development of the model also yields information on the reliability of the mapping from low- to high-level results, including both the expected error of the prediction and the range of configurations and environments over which the assumption of molecular similarity can be expected to hold. Such a model can then be used to perform QM/MM calculations at the cost of the low-level quantum theory, while generating results that approach the accuracy of the high-level theory.

Figure 1 shows a schematic representation of the approach applied to the collinear reaction of H + HF → H_2 + F, the

trial reaction considered in this work. This reaction is sufficiently small that high-level computations can be done quickly, but sufficiently complex that it provides a realistic test of the approach. The transition-state has a substantial multireference character, and its position and energy are sensitive both to the level of ab initio calculation[37−40] and to the environment. Pathway 1 of Figure 1 shows the ideal, but computationally unfeasible, approach of using a high-level quantum chemical method to generate the information (energy of the system, E, and charge distribution of the reaction center, ρ) needed at each time step in an MD simulation. This work explores an alternative approach, pathway 2 of Figure 1, which generates this information from a model that takes as input information obtained from a low-level quantum chemical method. This model is trained on data generated for a set of molecular configurations that vary both the geometry of the reaction center and the electrostatic environment. Figure 1 shows that levels of computation between that of the low- and high-level theories may be useful either as inputs for pathway 2 or as additional information for the model development.

Section 2 describes the development of the model that maps low- to high-level quantum chemical results. The data used to extract the model is discussed in Section 2.1. One important aspect of the model development is determining a set of descriptors that can serve as useful inputs to a regression that maps low- to high-level results. The descriptors used here are based on a principal component analysis of Stone's distributed multipoles,[41] as discussed in Section 2.2. The general form of the regression model is introduced in Section 2.3 with the details of the model development being discussed in Section 3.1. Sections 3.2 and 3.3 present the results, and Section 4 gives a brief summary and future directions.

2. Methods

2.1. Data Generation. To test our model, we chose to study the collinear $H + HF \rightarrow H_2 + F$ reaction. This collinear reaction has only two degrees of freedom and is sufficiently small that calculations may be carried out quickly. Nevertheless, the reaction involves breaking a H−F bond and forming a H−H bond, which is sufficiently complex that accurate prediction of the reaction surface requires both large basis sets and high levels of electron correlation.[37−40] The sensitivity to basis set and correlation have been attributed to the multireference character of the wave function in the vicinity of the transition-state. The best agreement with experiment is obtained using either multireference−configuration interaction (MRCI) calculations with the Davidson correction[38] or coupled-cluster calculations.[37] The coupled-cluster calculations estimate the barrier height of the linear and the bent transition-states as 2.16 and 1.63 kcal/mol, respectively, with an uncertainty of 0.1 kcal/mol. In addition to this sensitivity to the quantum chemical approach, the results below show that the reaction surface is sensitive to the environment, and so this system provides a reasonable test of the ability of the approach to describe reactions in complex environments.

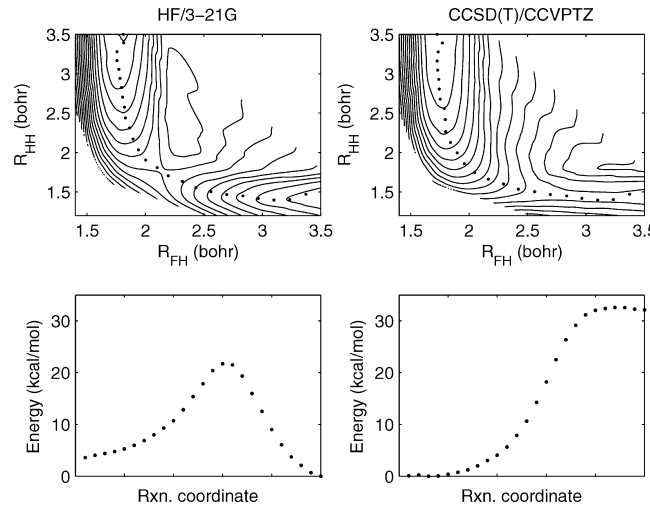

Figure 2. HF/3-21G (top left) versus CCSD (T)/CCVPTZ (top right) surfaces of the gas-phase collinear $H + HF \rightarrow H_2 + F$ trial reaction. The contours are from 0 to 33 kcal/mol in steps of 3 kcal/mol. The dotted lines show reaction paths (top panels) and reaction energy profiles (lower panels).

Table 1. Relative CPU Times for Some Typical QM Methods on C_5H_{12}[10,a]

	3-21G ($N = 69$)	6-31G* ($N = 99$)	6-31+G* ($N = 119$)	6-311++G** ($N = 194$)
HF [$N^{2.7}$]	1	3.8	5	23.1
B3YLP [$N^{\sim 3}$]	2.5	5	7	31
MP2 [$O\,N^4$]	1.4	7.6	10.2	60.8
MP4 [O^3V^4]	29.9	131.5	296.7	4066.2
QCSID(T) [O^3V^4]	63.3	220.9	558.3	8900.3

[a] Columns correspond to basis sets and rows to levels of electron correlation. Asymptotic scalings are in brackets (N = total number of basis functions, O = number of occupied orbitals, V = number of unoccupied orbitals).

Figure 2 shows the substantially different reaction surfaces obtained from a low-level (HF/3-21G) versus a high-level (CCSD (T)/cc-pVTZ) computation. Our goal is to predict the high-level surface from outputs generated by the low-level method. We note that DFT methods may provide low-level models with higher accuracy and, for semilocal DFT, lower computational costs. HF theory is used here as the low-level model to allow a clear test of the ability of the map to include the correlation energy. Table 1 illustrates the computational savings possible from such an approach. This table lists relative CPU times as a function of both the size of the basis set and the level of electron correlation, along with the asymptotic scalings of the correlated methods. The times are listed for a somewhat larger system, C_5H_{12}, than that studied here to better illustrate the potential computational savings from use of this approach.

Data spanning the regions of interest were generated using an automated system that varies both the geometry of the reaction center and the electrostatic environment. The 46 geometries shown as symbols in Figure 3 were chosen to span the relevant region of the potential energy surface (PES) of the collinear reaction. The different symbols in Figure 3 indicate grouping of the geometries into three regions: entrance channel, transition-state, and exit channel. This classification of the PES into three regions is used to test

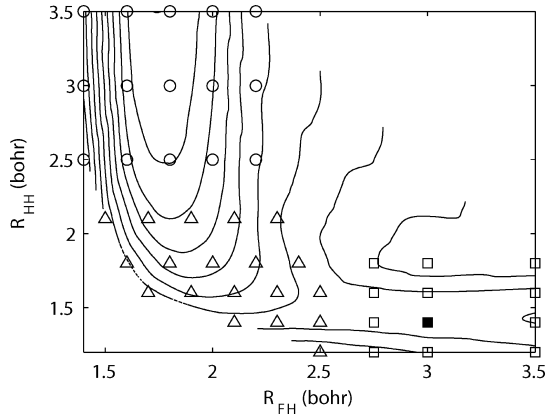

Figure 3. Contour plot of PES of H + HF → H_2 + F reaction in a typical environment. Symbols are geometries where calculations were performed for the transition region (tri) and for the entrance (circ) and exit (square) channel regions. The filled square is the geometry[38] used in the analyses of Tables 2 and 3.

Table 2. Mean Absolute Errors in kcal/mol for Predicting Correlated (QCISD) Self-Energies[a]

	HF	MP2	QCISD
STO-3G	0.12	0.05	0.00
3-21G	0.05	0.05	0.00
6-31G*	0.01	0.00	0.00
6-31++G**	0.02	0.01	0.00
cc-pVTZ	0.03	0.01	0.00

[a] From the output of lower-level computations via eq 2 with $N_{pca} = 10$, for the geometry shown as a filled square in Figure 3.

the ability of a single model to make predictions over a fairly broad region of the configuration space (Section 3.3). For the point shown as a filled square in Figure 3, data were generated using the 15 levels of theory obtained from combining 3 correlation methods (HF, MP2, and QCISD) with 5 basis sets (STO-3G, 3-21G, 6-31G*, 6-31++G**, and cc-pVTZ). These data are used to explore various aspects of mapping low- to high-level theories. For the remainder of the geometries, data were generated using one low-level (HF/3-21G) and one high-level method (CCSD(T)/cc-pVTZ). The data consists of single-point energy calculations carried out using a tight convergence threshold of 1.5×10^{-5} Hartree/Bohr with the GAUSSIAN03 software package.[42] The reaction surfaces are obtained from the 46 points of Figure 3 by triangle-based cubic interpolation. The reaction path and the energy profiles are then obtained from these interpolated surfaces.

For each of the reaction-center geometries of Figure 3, we generate a set of plausible electrostatic environments. In QM/MM calculations, the reaction center experiences the environment only through electrostatic interactions,[13] and so each environment consists of a set of external charges. For a biological reaction, MD trajectories would be run, and snapshots along this trajectory would be selected to yield a set of environments that span those likely to arise in the free energy computations. For our study, we generated a set of 250 random electrostatic environments by placing either a randomly oriented dipole (probability, $p = 0.8$), a single charge ($p = 0.1$), or void ($p = 0.1$) at each of the 8 corners

Table 3. Mean Absolute Error in kcal/mol for Predicting Large Basis (cc-pVTZ) Self-Energies[a]

	STO-3G	3-21G	6-31G*	6-31++G**	cc-pVTZ
HF	0.52	0.18	0.21	0.10	0.00
MP2	0.54	0.36	0.23	0.11	0.00
QCISD	0.57	0.53	0.24	0.12	0.00

[a] From the output of computations performed at the same level of correlation but with smaller basis sets. The model is that of eq 2 with $N_{pca} = 10$. The reaction-center geometry is that of the filled square in Figure 3. For instance, a model mapping descriptors generated at the HF/6-31G* level to energies of a HF/cc-pVTZ has a RMS error of 0.21 kcal/mol.

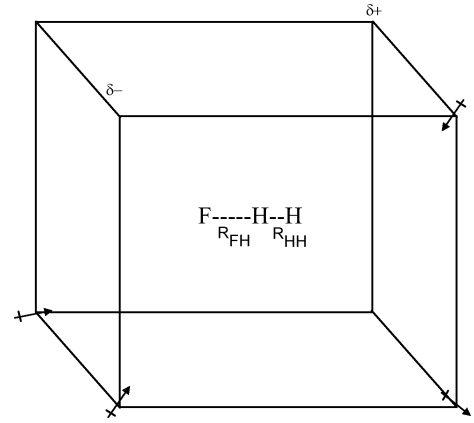

Figure 4. Schematic of electrostatic environments used for the H + HF → H_2 + F collinear reaction. R_{FH} and R_{HH} are the bond distances between the F−H and H−H atoms, respectively.

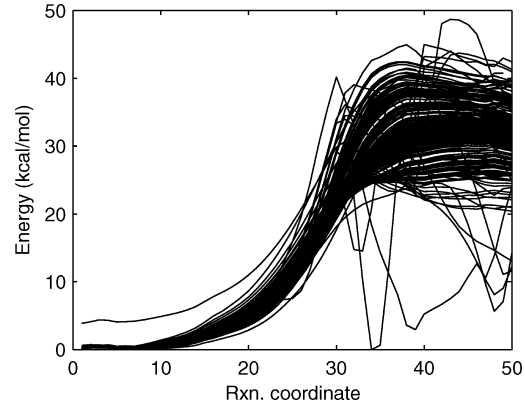

Figure 5. Reaction energy profiles from CCSD(T)/CCVPTZ calculations for the 250 electrostatic environments included in this study.

of a 12 Bohr cube surrounding the reaction center, as shown in Figure 4. The magnitude for the single charges is selected randomly from a uniform distribution ranging from -19.2×10^{-19} to $+19.2 \times 10^{-19}$ Coulomb and for the dipoles from 0 to 4.72 D. These environments were chosen to induce substantial perturbations in the reaction surface. The barrier heights range from 0 to 24 kcal/mol with an average and standard deviation of 1.71 ± 2.92 kcal/mol at CCSD(T)/cc-pVTZ level (see Figure 5). The combination of 250 environments and 46 reaction-center geometries yields a data set with 11 500 entries, where each entry contains the results of both the low-level (HF/3-21G) and high-level (CCSD(T)/cc-pVTZ) computations. Of the 11 500 combinations, 12

failed to converge at the high-level theory within the time feasible given the computational resources, and so were not included in the data set.

Our goal is to create a model that can map from low- to high-level methods by predicting the output of the high-level method, using only inputs generated by the low-level method. Since mapping from low- to high-level methods requires mapping across basis sets, the approach used to generate descriptors must yield results that are comparable across basis sets. For this reason, we use Stone's distributed multipoles.[41] The distributed multipoles are multipoles (monopole, dipole, quadrupole, etc.) centered on each atom and, optionally, on any chosen additional site, such as bond centers. The distributed multipoles reproduce the electrostatic potential surrounding the molecule to chemical accuracy, and so provide an essentially complete representation of the electronic distribution of the molecule.[43] The advantage of this description is that the same number and type of distributed multipoles can be generated for any basis set, as opposed to an object such as the one-electron density matrix, whose size and meaning changes with basis set and cannot be easily compared across basis sets.[44] An additional benefit is that distributed multipoles integrate well in QM/MM since the distributed multipoles are sufficient to compute the electrostatic interaction between the reaction center and the environment.[45,46]

Distributed multipoles are calculated using the distributed multipole Analysis of Gaussian98 wave functions (GDMA) software package.[47] Distributed multipoles up to the sixth order are generated first at just the atom centers, for a total of 108 distributed multipoles, and then at both atom and bond centers, for a total of 180 distributed multipoles.

The mapping from low- to high-level theories is meant to predict the properties of the reaction center in configurations and environments that are similar to those in the original data set. The quantity to be predicted by the model is, therefore, the self-energy of the reaction, which is extracted from the total energy, E_{TOT}, produced by the ab initio calculation as follows:

$$E = E_{TOT} - E_{CHARGES} - E_{INT} \quad (1)$$

where E is the self-energy of the reaction center, $E_{CHARGES}$ is the self-energy of the environment (i.e., the interaction energy between the fixed charges within the environment), and E_{INT} is the interaction energy between the reaction center and the fixed charges of the environment. E_{TOT} and $E_{CHARGES}$ are generated with GAUSSIAN03 by default. E_{INT} is obtained using the ORIENT[48] software package to predict the energy of interaction between the distributed multipoles of the reaction center and the fixed charges of the environment.

2.2. Feature Extraction. Feature extraction methods are useful to discover a minimal set of variables that may be used to describe the electronic structure of the reaction center. In ab initio calculations, a large set of variables is needed to obtain a form for the electronic wave function that is sufficiently flexible that accurate solutions of the Schroedinger equation may be obtained. For mapping low- to high-level quantum chemical results, the variables need only capture the variation in the electronic structure across situations that lie within the target range of validity for the model. The number of variables needed to describe differences among similar molecular structures is likely to be much smaller than the number of variables needed to obtain accurate solutions of the ab initio Hamiltonian.[49,50] Here, feature extraction is used to discover a reduced set of variables that describe variations in the electronic structure across the data set of Section 2.1.

The feature extraction method used here is principal component analysis (PCA), which is a simple and widely used approach.[51] PCA produces an orthogonal linear transformation of the feature space in such a way that the first linear combination of the original features, the first principal component, explains the greatest variance in data, and the second principal component is orthogonal to the first one and explains the greatest remaining variance and so on. Thus, each principal component identifies and ranks the most important features needed to capture the variability in the data. Principal components extracted in this manner define a new feature space that contains the same information as the original feature space but along dimensions that are ranked according to importance. Typically, only a few principal components are sufficient to capture the variability in data.

Here, PCA is applied to the distributed multipoles obtained from the high-level method. This yields principal components that are linear combinations of distributed multipoles. Since the units of the distributed multipoles vary with order (dipole, quadrupole, etc), some scaling approach is needed to make the various orders of the distributed multipoles comparable. A common approach in PCA analysis is the standardization of data by dividing each distributed multipole by the standard deviation of that particular distributed multipole in the input data. This gives each distributed multipole unit variance and so gives equal weight to all distributed multipoles, even those whose variance in the input data is quite small. Here, we instead weight the distributed multipoles according to their interaction with the fixed charges of the environment. This is done by first computing the average interaction energy between the DM with unit magnitude (e.g., the x component of the dipole on the F atom) and the fixed charges of the environment. The distributed multipoles are then divided by this average interaction energy. Figure 6 shows the result of PCA on the distributed multipoles generated at atomic centers from 11 488 calculations (the 46 reaction-center geometries of Figure 3 in the 250 environments of Figure 4). Figure 6 indicates that the number of degrees of freedom needed to capture the variation in the electronic structure of the reaction center is about five for both low- and high-level QM computations. The distributed multipoles from the low-level computations are then projected onto the high-level PCA vectors to give scores, S_i^{LL}, where i labels the principal components in order of importance.

2.3. Model Fitting. Above, we considered the choice of descriptors to be used as input to a model that maps from low- to high-level QM algorithms. The form of the model used here is a simple linear regression:

$$E^{HL} = p^{const} + p^{ener}E^{LL} + \sum_{i=1}^{N_{pca}^{lin}} p_i^{lin} S_i^{LL} + \sum_{i=1}^{N_{pca}^{quad}} p_i^{quad} (S_i^{LL})^2 \quad (2)$$

where p is the model parameter, E is the self-energy of the reaction center (the energy with electrostatic interactions between the reaction center and the environment removed via eq 1), and S_i^{LL} is the projection of the low-level distributed multipoles onto the i^{th} principle component. N_{pca}^{lin} and N_{pca}^{quad} are the number of principal component descriptors included in the model for the linear and quadratic terms, respectively. When N_{pca}^{lin} and N_{pca}^{quad} are equal, they are quoted below as simply N_{pca}.

All results presented below use five-fold cross validation, such that the model is trained on a randomly selected subset of 80% of the data and tested on the remaining 20%. The data is divided randomly into five equally sized subsets, and predictions for each subset are obtained from a model trained to the other four subsets.

3. Results

3.1. Form of the Canonical Model. The model involves choices regarding the number and type of distributed multipoles, the form of the regression in eq 2 (linear versus quadratic terms and inclusion of self-energy from the low-level model), and the number of PCA descriptors included in the regression. The canonical model used in the bulk of this paper includes distributed multipoles up to quadrupoles on each atom center, and includes the low-level energy in the regression along with both linear and quadratic terms for the 10 most important PCA vectors ($N_{pca} = 10$ in eq 2). We next examine the sensitivity of the model predictions to these choices.

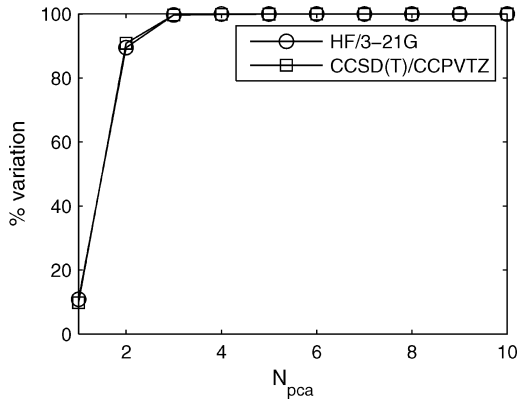

Figure 6. Percent variance of the electronic structure of the reaction center (i.e., the distributed multipoles) explained as a function of the number of principal components retained in the description.

We will initially examine some general aspects of the model fitting, holding the reaction center at the geometry shown as a filled triangle in Figure 3. This point is in the transition-state region of the isolated reaction center, where QM computations are expected to be most challenging.

The sensitivity of the model to the choice of distributed multipoles is shown in Figure 7. The average errors are to be compared to the standard deviation of 1.7 kcal/mol for the self-energy of the QCISD/cc-pVTZ calculations. Figure 7 shows the mean absolute error as a function of the number of principal components included in the regression, N_{pca} of eq 2, for various levels of multipoles and with (right) and without (left) inclusion of distributed multipoles at bond centers. The results in Figure 7 show that inclusion of distributed multipoles at bond centers does little to improve the performance of the model, and so the canonical model includes distributed multipoles only at atomic centers. Figure 7 also shows that the model performance increases significantly with addition of high-order distributed multipoles up to quadrupoles, but inclusion of higher ranks do not significantly decrease the error. Therefore, the canonical model includes only distributed multipoles up to quadrupoles on atomic centers, yielding a set of 27 raw descriptors. Figure 7 also shows that the error in the fit drops rapidly for the first seven principal components and then levels off. (This is a slower convergence than that seen in Figure 6, suggesting that an alternative approach to selecting input variables for the model could be beneficial.) Figure 7 suggests that N_{pca} should be greater than 7, but our final choice of N_{pca} for the canonical model will be based on fits to all reaction-center geometries discussed below.

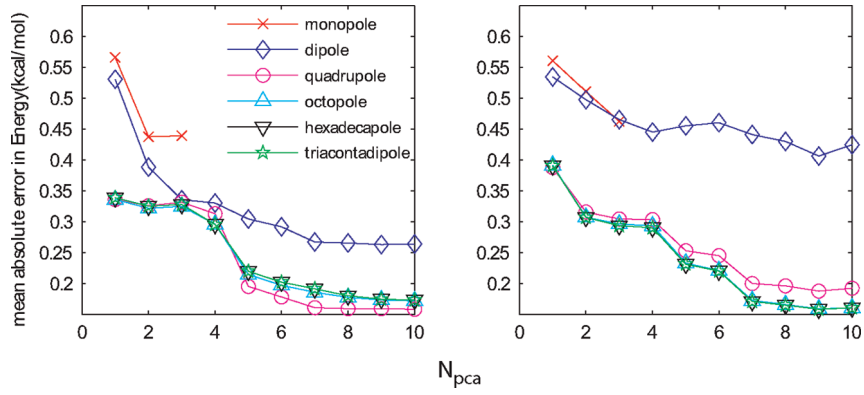

Figure 7. Mean absolute errors in kcal/mol from models predicting the QCISD/cc-pVTZ self-energy from output of HF/3-21G computations at the reaction center geometry, shown as a filled square in Figure 3. The lines show the average error versus the number of principal components included in both the linear quadratic terms of eq 2, when distributed multipoles up to the indicated level are included in the analysis. Distributed multipoles are included only on atoms (left) or on both atoms and bond centers (right).

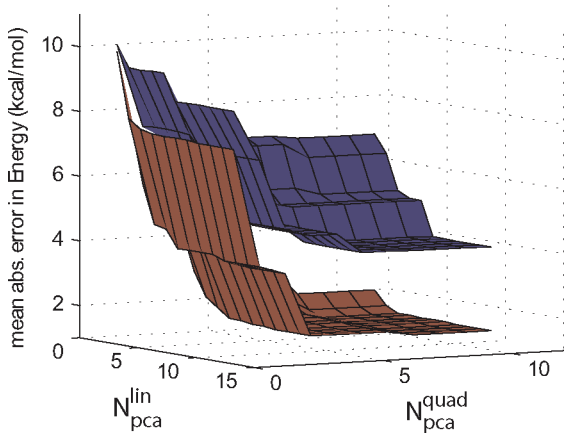

Figure 8. Surface plot of mean absolute error as a function of the number of linear and quadratic terms included in eq 2 for a map from HF/3-21G to CCSD (T)/cc-pVTZ theory, using a single regression for all 46 geometries of Figure 3. Results are shown with (red) and without (blue) including the HF/3-21G self-energy in the model of eq 2.

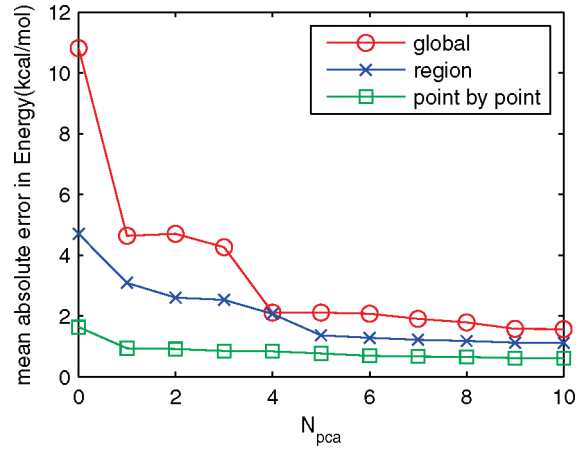

Figure 9. Error versus number of principle components in the linear regression of eq 2 for a model mapping HF/3-21G to CCSD(T)/cc-pVTZ self-energies. Results are shown for global, regional, and point-by-point fits. The total number of fitting parameters in the model of eq 2 is 92 N_{pca} + 1, 6 N_{pca} + 1, and 2 N_{pca} + 1, for point-by-point, regional and global fits respectively.

Next, we examine the model performance when multiple reaction-center geometries are included in a single regression. Figure 8 shows results for the global fit as a function of N_{pca} for the linear and quadratic terms of eq 2. Results are also shown both with and without inclusion of the self-energy from the low-level theory, E_{LL}, in the model of eq 2. Inclusion of E_{LL} substantially improves the quality of the fit, since this allows the model to focus on corrections to the energy arising from use of larger basis sets and from inclusion of electron correlation, as opposed to fitting the energy itself. The error drops smoothly with the addition of parameters to the fit and the addition of linear and quadratic terms leads to roughly equivalent improvements. The canonical model used in the remainder of this paper includes up to the first 10 principal components in both the linear and quadratic terms, at which point the errors are 0.6 kcal/mol for the point-by-point fit and 1.1 kcal/mol for the regional fit (see Supporting Information). Addition of cubic terms was also explored but found not to significantly outperform the model of eq 2 (data not shown).

3.2. Extrapolation Across Basis Sets and Electron Correlation Methods. This section examines the ability of the canonical model to extrapolate along the two dimensions that establish the level of the quantum chemical computation: correlation method and basis set. The analysis is done at the geometry, shown as a filled square in Figure 3.

Table 2 shows the ability to extrapolate across electron-correlation methods for a variety of basis sets. The results suggest that accurate maps can be developed from low to high levels of correlation. This success is consistent with the assumption of density functional theory (DFT), that the correlation energy is a functional of the one-electron density. The distributed multipoles used as inputs to the model capture the electronic distribution and so contain much of the information present in the one-electron density. In previous work, Janesko et al. used feature extraction algorithms to develop models for correlation energy based explicitly on density matrices.[28] That work developed a model to predict the two-electron density matrix, $\rho^{(2)}_{i,j,k,l}$ from the one electron density matrix, $\rho^{(1)}_{i,j}$, thereby predicting the correlation energy from the one-electron density, as in DFT methods. However, the use of density matrices has the disadvantage of making it difficult to develop models that connect across basis sets, and it is for this reason that we have developed the DM approach described here.

Table 3 shows the ability to extrapolate across basis sets, for a variety of levels of correlation. Comparison with Table 2 reveals that most of the error in the model predictions arises from mapping across basis sets. For the HF calculations, substantial improvement is attained by using 3-21G as the low-level theory as opposed to that of STO-3G. This result is consistent with Janesko's work on functional group basis sets derived from PCA of natural orbitals, which found that the intrinsic dimensionality of a functional group is larger than the number of degrees of freedom in a minimal (STO-3G) basis and is roughly equivalent to that of a 3-21G basis set.[50] Note, however, that the accuracy of the basis set extrapolation depends on correlation method, with higher levels of correlation requiring larger basis sets as the low-level input to the model.

3.3. Mapping from Low- To High-Level Potential Energy Surfaces. This section examines the degree to which a single regression can be used to make predictions for different reaction center geometries. Figure 9 shows that using a single global regression for all the reaction-center geometries (points labeled with symbols in Figure 3) yields an error that is substantially larger than that obtained from a point-by-point fit in which a separate regression is performed at each reaction-center geometry. This is expected, since the number of fitting parameters is substantially larger for a point-by-point fit. Also, the success of the regression is related to molecular similarity, and the reaction center changes its character as the reaction progresses. Regression of the three regions (transition-state and entrance and exit channel geometries of Figure 3) yields better results than a global fit, while still using a single regression to make predictions for a range of geometries. Figures 10 and 11 show

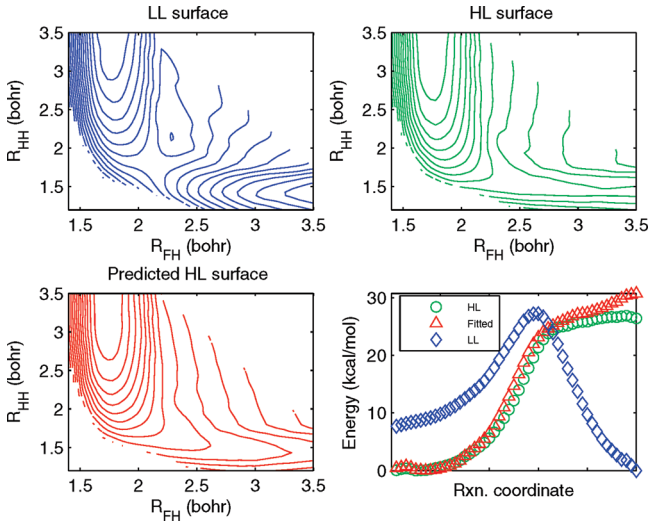

Figure 10. Self-energy obtained from low-level HF/3-21G (left top) and high-level CCSD(T)/cc-pVTZ calculations (right top) and from a global fit from low- to high-level self-energies (left bottom) for the collinear H + HF → H_2 + HF trial reaction in a typical environment. The contours are from 0 to 30 kcal/mol in steps of 3 kcal/mol. The reaction energy profiles are compared in the lower right panel.

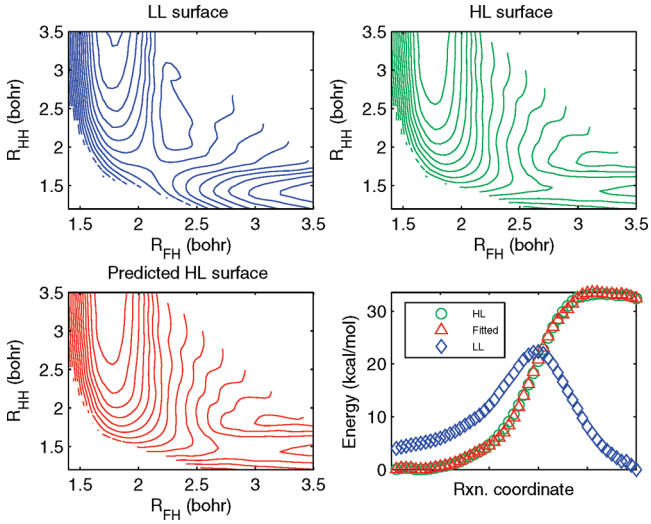

Figure 11. Results of a regional fit from HF/3-21G to CCSD(T)/cc-pVTZ methods for the collinear H + HF → H_2 + HF trial reaction in a typical electrostatic environment. The notation is as in Figure 10.

both a representative fitted reaction surface and path. The regional fits substantially outperform global fits and yield smooth potential energy surfaces. Fits of the remaining 249 environments are summarized in Figure 12, which shows the error in the reaction energy profiles (the lower panels of Figures 10 and 11) averaged over all environments. We note that there are a few environments that induce very large changes in the reaction profiles (see Figure 5), and removal of these extreme environments would further reduce the average error.

The results for the regional fits in Figure 12 indicate that it is possible to develop a single regression that can handle a range of reaction-center geometries. The success of such fits relies on grouping of geometries into sets where

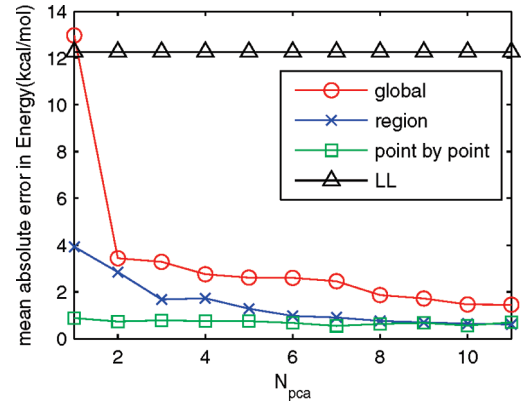

Figure 12. Mean absolute error in the reaction profile energies for a map from HF/3-21G to CCSD(T)/cc-pVTZ theories, averaged over all 250 environments. Results are shown versus number of principle components included in both the linear and quadratic terms of eq 2 for global (green square), regional (blue star) and point-by-point (red circle) fits. The average error between the low- and high-level calculated reaction paths is shown as black triangles for reference.

molecular similarity may be expected to apply. The division into regions, shown in Figure 3, is ad hoc; however, a more systematic approach can be envisioned in which cluster analysis is used to group electronic structures (as opposed to geometries) into similar regions. Such an approach will lead to a model of the type shown in Figure 1, where a decision tree is first used to classify the electronic structure from the low-level theory into a region, and then a region-specific regression is used to map from low-to high-level results. Since the MD algorithm requires forces, smoothing the results from different regions may be necessary, in which case functions that smoothly switch between regressions may be used for cases that lie near boundary regions between clusters. However, no smoothing was used at the boundaries in the current regional and point-by-point fits, and the resulting reaction surfaces and energy profiles are smooth and will lead to smoothly varying forces.

The results presented above consider only models for the self-energy of the reaction center (eq 2). To compute the interaction energy between the reaction center and the environment, the charge distribution of the reaction center is also needed. This interaction can be well computed from the distributed multipoles of the reaction center,[46,46] and so a model that predicts the high-level distributed multipoles from the low-level distributed multipoles would be sufficient for this purpose. A preliminary investigation revealed that for the entire data set r^2 is 0.93 between low-level (HF/3-21G) and high-level (QCISD/6-31++G**) distributed multipoles, as opposed to 0.34 for the correlation between low- and high-level self-energies of the reaction center. This suggests that the prediction of distributed multipoles is a relatively easy task when compared to the prediction of the self-energy, and so predictions of distributed multipoles are not explicitly addressed in this paper.

4. Conclusion

Here, we explore the use of molecular similarity to develop models for use in QM/MM simulations that can, at the cost

of a low-level ab initio calculation, produce results that approach the accuracy of high-level ab initio calculations. Our approach first generates detailed quantum chemical data on the reaction center in geometric configurations and electrostatic environments that span those expected to arise during the MD trajectory. This data includes results obtained from both low- and high-level ab initio methods. This information is then used to develop a low-cost model that can reproduce the output of the high-level theory using only inputs generated from the low-level theory. This approach was tested on the $H + HF \rightarrow H_2 + F$ collinear reaction. This reaction center is sufficiently small that high-level calculations can be performed quickly. Yet despite the small size, the reaction still involves breaking and forming of bonds in a manner that is sensitive to the environment and provides a realistic test of the approach.

The ability to predict high-level results using only descriptors generated from low-level calculations was tested along the two dimensions that define a quantum chemical method: the level of electron correlation and the size of the basis set. The models predict the results of the high-level theory using, as input, distributed multipoles obtained from the low-level method. The distributed multipoles are monopoles, dipoles, and quadrupoles placed on each atomic center, and they summarize the electronic distribution in a manner that is independent of basis set. Including electron correlation, by predicting QCISD results from HF inputs, leads to an average error of less than 0.05 kcal/mol for split-valence basis sets. This relatively low error is consistent with the assumption of DFT, that the correlation energy is a functional of the electronic density. Extrapolating across basis sets is the primary source of error in the models, with the extrapolation from 6-31G* to cc-pVTZ basis sets giving an error of 0.21 kcal/mol within HF theory and 0.24 kcal/mol within QCISD theory. The models used here were parametrized to about 10^4 high-level computations and so will lead to substantial savings for situations, such as MD simulations, where the quantum algorithm is called 10^6 or more times.

An important criterion regarding the applicability of this approach to more complex reaction centers is the extent to which a single model can handle a range of reaction-center geometries. The current study showed that reasonable accuracy can be obtained when configurations of relevance to the collinear trial reaction are broken into three regions: the transition region and the entrance and exit channels. This suggests that regressions can be developed that span fairly large regions of configuration and environment space. The ability of a single regression to describe a range of configurations benefits from the use of the model only to build basis set and correlation corrections onto the energy generated by a low-level method. The energy from the low-level method already contains reasonable estimates to the interactions energies and contains how these vary with geometry.

The use of machine learning to group input configurations into regions, i.e., to develop the optimal decision tree for selecting regressions in Figure 1, should yield even better results than the ad hoc selection of regions used here. This may become especially important for larger reaction centers. Consider, for instance, the important class of biological reactions that involve transfer of a hydrogen atom, a phosphate group or other small molecular fragment between two groups. The dimensions corresponding to fragment transfer will have the greatest total spread. Fluctuations in the orientation of the groups between which the transfer occurs must also be included but with smaller amplitudes, since such motions are often constrained by covalent attachment to the protein backbone.

Inclusion of additional information from the QM methods may also lead to better performing models. In particular, many QM methods have analytical derivative methods that generate forces and higher energy derivatives at little additional cost.[52] These derivatives provide additional information that may aid in the development of the model mapping low- to high-level energies. Analytical gradient information may also allow for a direct and, thus, a more efficient prediction of forces.

The success of the models presented here is encouraging, especially given the simplicity of the methods, i.e., PCA and linear regression, used to discover the models. The use of PCA for feature extraction can be extended both to nonlinear feature extraction methods and to methods that select latent variables based on the importance to the model, as opposed to the variability in the input data. Likewise, the linear regression used here is among the simplest possible means to map between low- and high-level theories, and a wide array of alternative methods from statistics and machine learning can be envisioned.[53]

Acknowledgment. This work was funded by the National Science Foundation (CHE-0719350).

Supporting Information Available: Maps from (HF/3-21G) to another (QCISD/6-31++G**) theory to show that similar results are obtained for maps to different high-level theories. The cost of and the complexity of the high-level calculation can, therefore, be based on the degree of accuracy appropriate for the given application. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lin, H.; Truhlar, D. *Theor. Chim. Acta.* **2007**, *117* (2), 185–199.

(2) Friesner, R.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

(3) Senn, H.; Thiel, W. *Top. Curr. Chem.* **2007**, *268*, 173–290.

(4) Senn, H.; Thiel, W. *Angew. Chem. Int.* **2009**, *48* (7), 1198−1229.

(5) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G.; Xu, D.; Guo, H. *J. Phys. Chem. B* **2006**, *110* (13), 6458–6469.

(6) Xie, W.; Song, L.; Truhlar, D.; Gao, J. *J. Phys. Chem. B* **2008**, *112* (45), 14124–14131.

(7) Vreven, T.; Byun, K.; Komaromi, I.; Dapprich, S.; Montgomery Jr, J.; Morokuma, K.; Frisch, M. *J. Chem. Theor. Comp.* **2006**, *2* (3), 815–826.

(8) Freindorf, M.; Shao, Y.; Furlani, T.; Kong, J. *J. Comput. Chem.* **2005**, *26* (12), 1270−1278.

(9) Crespo, A.; Scherlis, D.; Martí, M.; Ordejon, P.; Roitberg, A.; Estrin, D. *J. Phys. Chem. B* **2003**, *107* (49), 13728–13736.

(10) Foresman, J. B.; Frisch, A. *Exploring Chemistry With Electronic Structure Methods*; Gaussian Inc.: Pittsburgh, PA, 1996; 123−125.

(11) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry Introduction to Advanced Electronic Structure Theory*; Dover Publications: Mineola, N.Y, 1996; 32−45.

(12) Noodleman, L.; Lovell, T.; Han, W.; Li, J.; Himos, F. *Chem. Rev.* **2004**, *104* (2), 459–508.

(13) Gordon, M.; Mullin, J.; Pruitt, S.; Roskop, L.; Slipchenko, L.; Boatz, J. *J. Phys. Chem. B* **2009**, *113* (29), 9646–9663.

(14) Yang, W. *Phys. Rev. Lett.* **1991**, *66* (11), 1438–1441.

(15) Steele, R.; DiStasio Jr, R.; Shao, Y.; Kong, J.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074108.

(16) White, C.; Johnson, B.; Gill, P.; Head-Gordon, M. *Chem. Phys. Lett.* **1996**, *253* (3−4), 268–278.

(17) Wolinski, K.; Pulay, P. *J. Chem. Phys.* **2003**, *118*, 9497.

(18) Lee, M.; Head-Gordon, M. *J. Chem. Phys.* **1997**, *107*, 9085.

(19) Berghold, G.; Parrinello, M.; Hutter, J. *J. Chem. Phys.* **2002**, *116*, 1800.

(20) Lu, W.; Wang, C.; Schmidt, M.; Bytautas, L.; Ho, K.; Ruedenberg, K. *J. Chem. Phys.* **2004**, *120*, 2629.

(21) Dewar, M.; Zoebisch, E.; Healy, E.; Stewart, J. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909.

(22) Ridley, J.; Zerner, M. *Theor. Chim. Acta.* **1973**, *32* (2), 111–134.

(23) Scuseria, G. *J. Phys. Chem. A* **1999**, *103* (25), 4782–4790.

(24) Schutz, M.; Werner, H. *J. Chem. Phys.* **2001**, *114*, 661.

(25) Kohn, W. *Rev. Mod. Phys.* **1999**, *71* (5), 1253–1266.

(26) Goedecker, S. *Rev. Mod. Phys.* **1999**, *71* (4), 1085–1123.

(27) Van der Vaart, A.; Gogonea, V.; Dixon, S.; Merz JR., K. *J. Comput. Chem.* **2000**, *21* (16), 1494−1504.

(28) Janesko, B.; Yaron, D. *J. Chem. Phys.* **2003**, *119*, 1320−1328.

(29) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73* (2), 325–348.

(30) Machida, K. *Principles of Molecular Mechanics*; John Wiley & Sons Inc: New York, 1999, 15−19.

(31) Tangney, P.; Scandolo, S. *J. Chem. Phys.* **2002**, *117*, 8898.

(32) Ercolessi, F.; Adams, J. *Europhys. Lett.* **1994**, *26* (8), 583–588.

(33) Mehl, M.; Papaconstantopoulos, D. *Phys. Rev. B: Condens. Matter* **1996**, *54* (7), 4519–4530.

(34) Tabacchi, G.; Mundy, C.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **2002**, *117*, 1416.

(35) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B.* **2001**, *105* (2), 569–585.

(36) Sastry, K.; Johnson, D.; Thompson, A.; Goldberg, D.; Martinez, T.; Leiding, J.; Owens, J. *Mater. Manuf. Processes* **2007**, *22* (5), 553–561.

(37) Werner, H.; Kallay, M.; Gauss, J. *J. Chem. Phys.* **2008**, *128*, 034305.

(38) Stark, K.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104* (17), 6515–6530.

(39) Gonzalez-Luque, R.; Merchan, M.; Roos, B. O. *Chem. Phys.* **1993**, *171* (1−2), 107–118.

(40) Cardoen, W.; Gdanitz, R.; Simons, J. *J. Phys. Chem. A* **2006**, *110* (2), 564–571.

(41) Stone, A.; Alderton, M. *Mol. Phys.* **2002**, *100* (1), 221–233.

(42) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. *Gaussian 03*, Revision 6.0; Gaussian, Inc.: Wallingford CT, 2004.

(43) Stone, A. *The Theory of Intermolecular Forces*; Clarendon Press: Oxford, U.K., 1997, 105−119.

(44) Stone, A. *J. Chem. Theor. Comp.* **2005**, *1* (6), 1128–1132.

(45) Gordon, M.; Freitag, M.; Bandyopadhyay, P.; Jensen, J.; Kairys, V.; Stevens, W. *J. Phys. Chem. A.* **2001**, *105* (2), 293–307.

(46) Gordon, M.; Slipchenko, L.; Li, H.; Jensen, J. *Ann. Rep. Comp. Chem.* **2007**, 177.

(47) Stone, A. *Distributed Multipole Analysis of Gaussian98 Wavefunctions, Revision 2.1*; University of Cambridge: UK 1999.

(48) Stone, A.; Dullweber, A.; Hodges, M.; Popelier, P.; Wales, D. *ORIENT*, Version 4.6; University of Cambridge: Cambridge, U.K., 1995.

(49) Janesko, B.; Yaron, D. *J. Chem. Phys.* **2004**, *121* (5635).

(50) Janesko, B.; Yaron, D. *J. Chem. Theory Comp.* **2005**, *1* (2), 267–278.

(51) Duda, R.; Hart, P.; Stork, D. *Pattern Classification*; Wiley: New York: 2001, 114−117.

(52) Pulay, P. *Adv. Chem. Phys.* **1987**, *69*, 241–286.

(53) Mitchell, T.; *Machine Learning*; WCB/McGraw Hill: Hightstown, NJ, 1997, 5−14.

CT9004195