QM/MM Boundaries Across Covalent Bonds: A Frozen Localized Molecular Orbital–Based Approach for the Effective Fragment Potential Method

2 AUTHORS:

Visvaldas Kairys
Vilnius University
47 PUBLICATIONS **1,109** CITATIONS

SEE PROFILE

Jan Halborg Jensen
University of Copenhagen
116 PUBLICATIONS **6,470** CITATIONS

SEE PROFILE

QM/MM Boundaries Across Covalent Bonds: A Frozen Localized Molecular Orbital-Based Approach for the Effective Fragment Potential Method

Visvaldas Kairys[†] and Jan H. Jensen*

Department of Chemistry, University of Iowa, Iowa City, Iowa 52242

Received: March 7, 2000; In Final Form: May 22, 2000

A computational methodology to treat the covalent boundary between QM and MM regions in the Effective Fragment Potential (EFP) method, by defining a buffer region consisting of frozen localized molecular orbitals (LMOs), is introduced. The implementation of energy, gradient, and EFP parameter evaluations in the presence of frozen LMOs is discussed. The magnitude and source of errors introduced by various choices of buffer region is studied for the proton affinities of lysine and the tripeptide glycine-lysine-glycine. It is shown that by reasonable choice of the frozen density buffer region the proton affinity error can be consistently decreased to less than 0.5 kcal/mol compared to the full ab initio calculation.

I. Introduction

Hybrid quantum mechanical/molecular mechanical methods[1] (QM/MM) hold a great promise for molecular modeling since they in principle can deliver "QM accuracy" for many "MM-sized" chemical systems, including reactions in which bonds are broken or formed. The popularity of this approach precludes any comprehensive review here, so we merely mention a few representative example of its versatility. Morokuma's[1a] ONIOM approach has been used to investigate the reaction mechanisms of large organic and organometallic systems, while Gao's[1b] combined use of AM1 and a polarizable force field has proven to be an effective model of discrete solvation. Similarly, Merz and co-workers[1c,d] have investigated a large variety of enzyme mechanisms using a hybrid PM3/MM approach.

In the Effective Fragment Potential (EFP) method[2] the active part of a molecular system is treated with ab initio quantum mechanics while the rest is replaced by EFPs. The EFPs are generated by separate ab initio calculations.

In their present implementation[2] the EFPs simulate the most important nonbonded energy terms: Coulomb interactions, classical many-body induction, and an empirical representation of the short-range energy terms (the internal geometries of the EFPs are frozen and their internal energies can be neglected). The Coulomb term consists of a distributed multipole expansion[3] (charges through octupoles at all atomic centers and bond midpoints), while the induction term consists of dipole polarizability tensors for each valence (localized) molecular orbital. Both expressions can be systematically improved by including higher order terms or more expansion points, but the current form has proved sufficient thus far. The short-range term represents the difference between the SCF interaction energy and the Coulomb plus induction energies and accounts for purely quantum effects such as exchange repulsion and exchange-induction. It is *currently* generated by a fit to an energy surface that describes all the possible intermolecular arrangements of a representative chemical system (e.g., the water dimer for a water-EFP).

The EFP method has successfully been applied to the study of aqueous solvation effects,[4] by using EFPs to represent solvent molecules while the solute molecule (or molecules) is treated with Hartree–Fock theory, as well as to some studies[5] of enzyme active site mechanisms.

Construction of the short-range potential is the most computational and time-demanding step in creating new EFPs. Thus, recent work has focused on replacing this term with separate expressions for the short-range forces that are *free of adjustable empirical parameters.* Thus, a complete EFP can be generated by a *single* ab initio calculation on the corresponding molecule as shown in Scheme 1. The general expression for EFP/EFP exchange repulsion interactions has been published some years ago,[6] while two different general approaches to include charge penetration have been developed very recently.[7,8] Work on expressions for exchange-induction (including EFP/ab initio exchange repulsion), charge transfer, and dispersion is currently in progress.

However, another major improvement is necessary to make the EFP methodology generally applicable to the study of chemistry in large condensed phase systems, namely the ability to place the ab initio/EFP boundary across a covalent bond. This is the subject of the present paper.
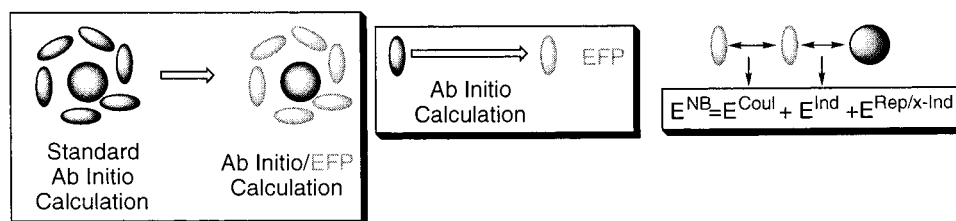
The general problem of covalent boundaries was first dealt with by introducing link atoms (usually hydrogen) that satisfy the valence requirements of the QM system but not contribute to the molecular energy.[9] Gao et al.[10] have reviewed the literature and problems associated with this technique, though subsequent work by Zhang, Lee, and Yang[11] using specifically parametrized boundary atoms gave promising results.

Rivail and co-workers[12] have addressed the boundary problem in a different manner, similar to an earlier idea by Warshel and Levitt,[9a] by representing the boundary-bonds as frozen localized molecular orbitals (LMOs) within their local self-consistent field (LSCF) algorithm. The frozen LMOs are strictly localized, i.e., truncated to span the basis functions of only two atoms, and act as a buffer between the QM and MM regions. A direct comparison to the link atom method by Reuter, Dejagere, Maigret, and Karplus[13] concluded that both methods "give results of similar accuracy and neither one is systematically better than the other" when the QM region is treated with a

* Corresponding author: jan-jensen@uiowa.edu.
† Present address: Center for Advanced Research in Biotechnology, 9600 Gudelsky Dr, Rockville, MD 20850.

QM/MM Boundaries Across Covalent Bonds

J. Phys. Chem. A, Vol. 104, No. 28, 2000 **6657**

SCHEME 1: EFP Method at a Glance

semiempirical wave function. Gao, Amara, Alhambra, and Field,[10] have elaborated on the semiempirical implementation of the LSCF method by specifically parametrizing orbitals (generalized hybrid orbitals) for the linking bonds.

The LSCF method was initially implemented for semiempirical wave functions but was subsequently extended to ab initio wave functions by Assfeld and Rivail.[12d] Very recently, Philipp and Friesner[14] implemented an ab initio/MM version of the frozen LMO method into the program JAGUAR and performed extensive tests on the conformational space of alanine di- and tetra-peptides. In all cases, only RHF wave functions have been considered.

The results of Assfeld and Rivail as well as Philipp and Friesner indicate that the use of frozen LMOs to treat covalent QM/MM boundaries for ab initio wave functions has great promise. Thus, this paper describes the implementation of a frozen LMO buffer method for treating ab initio/EFP covalent boundaries. The following section discusses how the energy and gradients are computed in the presence of frozen LMOs, as well as how the LMOs are truncated. Section III describes how the buffer and EFP regions of a molecule are obtained by two separate ab initio calculations, using the specific example of the amino acid lysine. Finally, we utilize the method to calculate the proton affinity of lysine and the tri-peptide glycine-lysine-glycine, using several choices of buffer region, buffer size, and location. In addition, we explore a few other representations of the electrostatic potential of the MM region, to demonstrate the utility of the EFP/buffer/ab initio method.

II. Theory

A. General Considerations. A buffer region consisting of one or several LMOs is defined as the ab initio/EFP boundary. These LMOs are obtained by an ab initio calculation on all, or a subset, of the system, projected onto the basis functions on the buffer atoms, and subsequently frozen in the EFP calculations. The ab initio/buffer region interactions are calculated by including the exact quantum mechanical Coulomb and exchange operator due to the charge distribution of the buffer region, in the ab initio Hamiltonian. This requires calculation of two-electron integrals over basis functions in the buffer region. Since the buffer MOs are frozen, the changes in induction contributions from the buffer region are neglected during a geometry optimization of the ab initio region. The effect of this approximation on the chemical reaction of interest can be systematically reduced by increasing the size of the ab initio region.

Variational collapse of the ab initio wave function into the buffer and EFP regions is avoided by keeping the ab initio MOs orthogonal to the buffer LMOs by Gramm-Schmidt orthogonalization. The presence of the buffer region provides sufficient separation between the EFP and the ab initio regions so that the remaining interactions can be treated as nonbonded interactions via the EFP terms discussed above. The EFP and buffer regions are always kept in the same relative position, and so

the EFP/buffer interaction energy will remain constant and does not need to be computed.

In the current implementation, the positions of the buffer and EFP regions are frozen in space. Thus, only the derivatives of the total energy with respect to the position of the ab initio atoms need to be computed.

B. Energy. One of the essential parts of our method is freezing selected molecular orbitals during the SCF step. There are two main ways currently in use: (1) Projection operators pioneered by Huzinaga[15] and applied to the QM/MM boundary problem by Assfeld and Rivail.[12d] (2) Methods related to the group function approach due to McWeeny,[16] such as the Reduced Variational Space method of Fink and Stevens[17] and the Constrained Space Orbital Variation method by Bagus and co-workers.[18] We employ the latter approach, and we briefly outline our implementation here.

(1) Construct the AO coefficient matrix $\mathbf{C}$ and identify MO vectors to be frozen. In our case, these are LMOs.

(2) Truncate the frozen LMOs. See below.

(3) LMOs to be frozen are moved to the front of $\mathbf{C}$ and $\mathbf{C}$ is Gramm-Schmidt orthonormalized starting with the frozen orbitals. This ensures that the frozen LMOs are normalized and remain constant during the SCF process.

(4) During each SCF iteration, the Fock matrix in the AO basis is transformed into the MO basis: $\mathbf{C}^{\dagger}\mathbf{F}_{AO}\mathbf{C} = \mathbf{F}_{MO}$.

(5) In the resulting $\mathbf{F}_{MO}$ matrix the nondiagonal elements involving the frozen orbitals are set to zero: $F_{ij} = 0, i \neq j$. This operation ensures the chosen molecular orbitals remain unchanged from iteration to iteration. To simplify implementation into GAMESS the modified MO−Fock matrix is backtransformed to the AO basis.

The SCF iterations are repeated until convergence.

Obviously, since the variational space is reduced, the resulting Hartree−Fock energy is higher than the energy obtained when all the orbitals are optimized.

Step (5) is identical for RHF, ROHF and GVB calculations. In the latter two cases we have restricted our implementation to frozen doubly occupied MOs, since the MOs to be frozen are presumed to be chemically inactive.

In the case of MCSCF molecular orbital freezing is already implemented in GAMESS.[19]

Once the final AO coefficient matrix is obtained, it can be used as a starting point for calculating the dynamic electron correlation energy of the ab initio region, using MP2, CI, or CASPT2. Since the frozen LMOs are chosen to be chemically inactive they can be frozen along with the atomic core MOs in the usual way.[20]

C. Gradients. We wish to optimize the geometry of the region of the molecule described by the nonfrozen MOs, in the presence of the frozen MOs. Thus, we require the analytical derivative of the energy with respect to the coordinate of select atoms. Freezing MOs introduces a few complications, which we discuss here (using the notation of Yamaguchi et al.[21]).

For RHF, the derivative of the electronic energy with respect to atom a can be written as

$$\frac{\partial E_{\text{elec}}}{\partial a} = E^a + 4\sum_i^{\text{all}}\sum_j^{\text{nf}} U_{ij}^a F_{ij} \tag{C1}$$

Here nf denotes nonfrozen occupied MOs, E^a denotes the derivative with respect to basis functions centered on a and also includes the Hellmann−Feynman force, while U_{ij}^a is the orbital response function defined by

$$\frac{\partial C_{\mu i}}{\partial a} = \sum_j^{\text{all}} U_{ji}^a C_{\mu j} \tag{C2}$$

Thus, the latter term accounts for the change in all the nonfrozen MOs as atom a is moved. It is important to note that the sum over j includes frozen MOs since the nonfrozen MOs are made orthogonal to the frozen MOs. So the usual simplification for converged RHF MOs (d.o. = doubly occupied),

$$4\sum_i^{\text{all}}\sum_j^{\text{d.o.}} U_{ij}^a F_{ij} = 4\sum_i^{\text{d.o.}} U_{ii}^a F_{ii}$$

$$= -2\sum_i^{\text{d.o.}} S_{ii}^a F_{ii} \tag{C3}$$

cannot be done since $F_{ij} \neq 0$ when i is a nonfrozen MO and j is a frozen MO. Instead we use the fact that, like for ROHF, $F_{ij} = F_{ji}$ so that

$$4\sum_i^{\text{all}}\sum_j^{\text{d.o.}} U_{ij}^a F_{ij} = 2\sum_i^{\text{d.o.}}\sum_j^{\text{d.o.}}(U_{ij}^a + U_{ji}^a)F_{ij}$$

$$= -2\sum_i^{\text{d.o.}}\sum_j^{\text{d.o.}} S_{ij}^a F_{ij} \tag{C4}$$

For ROHF, no complications are introduced by freezing LMOs since it does not affect the symmetry of the Fock matrix.

For GVB and MCSCF even the symmetry of the Fock matrix (i.e., the Lagrangian) is lost when frozen MOs are introduced, and we will discuss the gradient of these wave functions in another paper.

D. LMO Truncation. Though localized, MOs that describe the buffer region have contributions from all basis functions in the molecule. Since the objective of a QM/MM method is to reduce the size of the QM region these "tails" will also have to be removed in the MM region. Tails that extend into the QM region will have to be removed as well since otherwise the frozen density will not remain strictly frozen during the QM region geometry optimization.

In this study, we explore two methods for removing the LMO tails. One is simply to set the unwanted MO coefficients to zero followed by reorthonormalization. In an example of a single truncated molecular orbital, this procedure is equivalent to a multiplication of all orbital coefficients by the same normalization factor. Another method is to project the LMOs onto a smaller basis set.

Our projection procedure is based on the corresponding orbital transformation procedure used by King et al.[22] to minimize the number of integrals between two sets of nonorthogonal MOs. Here we explore the relation of this method to orbital truncation.

Consider a set of molecular orbitals $\{\psi_A\}$ expanded in terms of K *orthogonal* basis functions $\{\phi_i\}$,

$$\psi_A(\mathbf{r}) = \sum_i^K c_{iA}\phi_i(\mathbf{r})$$

where

$$c_{iA} = \langle\phi_i|\psi_A\rangle \tag{D1}$$

We wish to project these orbitals onto $L < K$ orthogonal basis functions $\{\phi_i'\}$

$$\psi_A'(\mathbf{r}) = \sum_i^L c_{iA}'\phi_i'(\mathbf{r})$$

where

$$c_{iA}' = \langle\phi_i'|\psi_A'\rangle \tag{D2}$$

such that

$$\psi_A(\mathbf{r}) \approx \psi_A'(\mathbf{r}) \tag{D3}$$

Thus, substituting eq D3 into eq D2 the new expansion coefficients are given by,

$$c_{iA}' \approx \langle\phi_i'|\psi_A\rangle \equiv D_{iA} \tag{D4}$$

If more than one orbital is projected it is necessary to enforce orthogonality,

$$\langle\psi_A'|\psi_B'\rangle = \sum_i^L\sum_j^L D_{iA}D_{jB}S_{ij}$$

$$= \sum_i^L D_{iA}D_{iB}$$

$$= \delta_{AB}$$

by determining the matrix $\mathbf{X}$ such that

$$\mathbf{X}^\dagger\mathbf{D}^\dagger\mathbf{D}\mathbf{X} = \mathbf{1} \tag{D5}$$
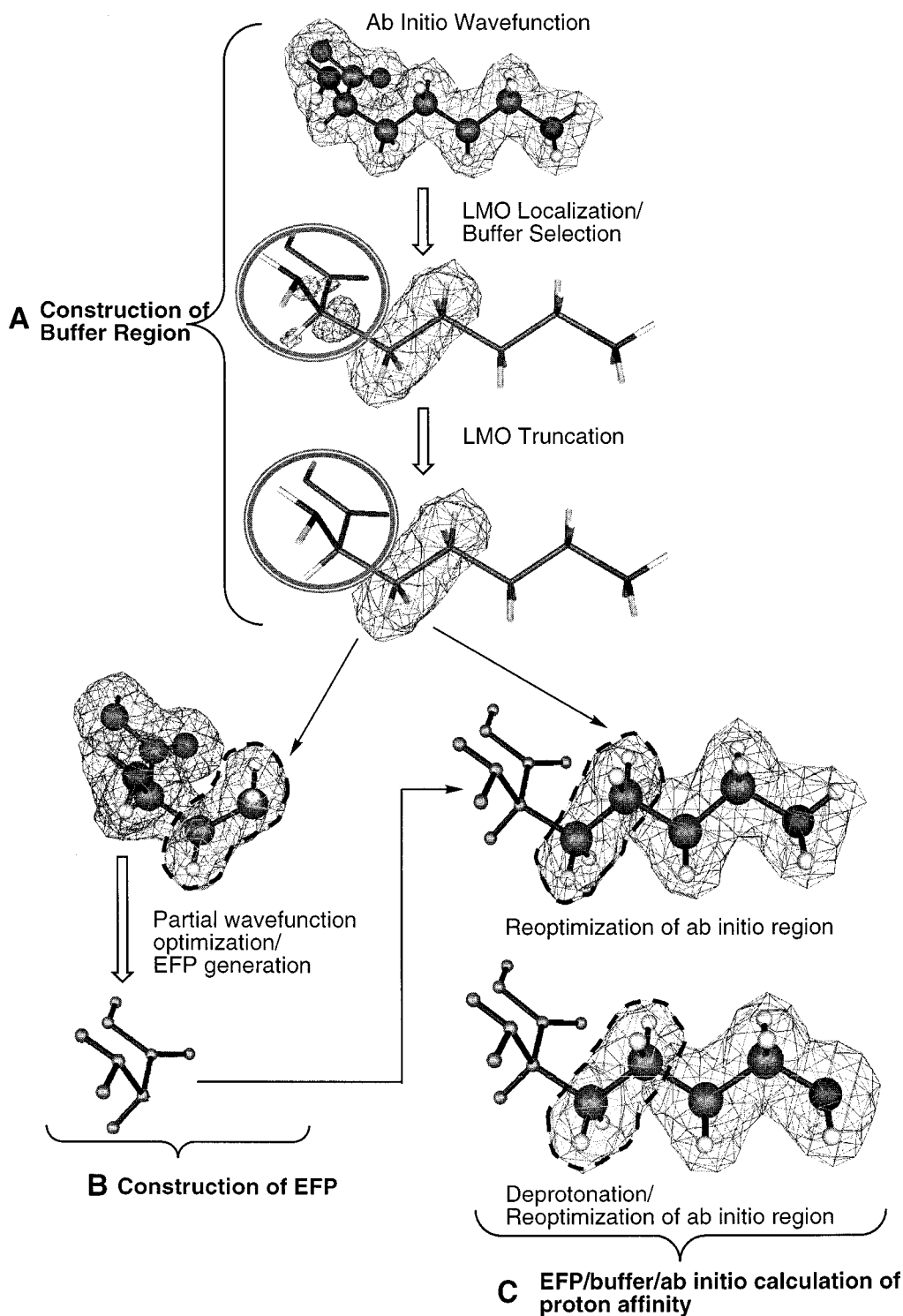
Thus,

$$c_{iA}' = \sum_q^{MO} X_{qA}D_{iq} \tag{D6}$$

This results in the following projection algorithm which transforms orthonormal MO vectors extending over the entire molecule [$\mathbf{C}_{\text{big}}$ where $\psi_A = \sum_\mu^K (\mathbf{C}_{\text{big}})_{\mu A}\chi_\mu$ and χ_μ is an atomic basis function] to orthonormal MO vectors extending over part of the molecule:

(1) Form $\mathbf{S}_{\text{small}}$, a square overlap matrix in the reduced atomic basis set (formed by excluding select atomic centers).

(2) Form an orthonormal set of basis functions $\mathbf{Q}$ by diagonalizing $\mathbf{S}_{\text{small}}$, so that $\phi_i' = \sum_\mu^L Q_{\mu i}\chi_\mu$.

(3) Form the rectangular overlap matrix $\mathbf{S}_{\text{sb}}$ with the rows corresponding to the reduced basis set and columns corresponding to the full basis set.

(4) Form $\mathbf{D} = \mathbf{Q}^\dagger\mathbf{S}_{\text{sb}}\mathbf{C}_{\text{big}}$ [cf. eq D4].

QM/MM Boundaries Across Covalent Bonds

J. Phys. Chem. A, Vol. 104, No. 28, 2000 **6659**

Figure 1. Schematical representation of the steps involved in buffer and EFP generation. See section III for more information.

(5) Diagonalize $\mathbf{D}^\dagger\mathbf{D}$ to obtain a set of eigenvectors $\mathbf{V}$ and eigenvalues Λ [cf. eq D5].

(6) Rearrange the eigenvalues in a descending order.

(7) Form $\mathbf{U} = \mathbf{D}\mathbf{V}\Lambda^{-1/2}\mathbf{V}^\dagger$ [=$\mathbf{D}\mathbf{X}$, cf. eq D6].

(8) Form the projected atomic orbital coefficient matrix $\mathbf{C}_{small} = \mathbf{Q}\mathbf{U}$ of orthonormal MOs.

In step (7), a symmetric transformation is employed, rather than canonical transformation ($\mathbf{U} = \mathbf{D}\mathbf{V}\Lambda^{-1/2}$) as in reference **22**, since the former yields orbitals that are more similar to the original LMOs. Both transformations yield identical frozen densities, but individual MO-dependent properties such as exchange and orthogonality are affected.

III. Computational Methodology

In this paper we demonstrate the utility of our method by calculating the proton affinity of the ϵ-NH$_2$ group in the amino acid lysine and the Gly-Lys-Gly tripeptide. These molecules are small enough to make full ab initio calculations feasible, but large enough to allow for several different choices of buffer region. Figure 1 depicts the general scheme of our method using

TABLE 1: Proton Affinities of Lysine (kcal/mol)[a]

	1 Reference ab initio calculation	2 QM/buffer calculation	3 QM/buffer/EFP calculation
	237.3 0.0	239.7 +2.4	239.5 +2.2
	237.3 0.0	236.9 −0.4	237.8 +0.5
	237.3 0.0	237.1 −0.2	237.2 −0.1
	237.2 0.0	234.6 −2.6	237.0 −0.2
	237.2 0.0	237.2 0.0	Structural Collapse

[a] The upper number is the absolute proton affinity; the lower one is the error relative to the reference ab initio calculation in column 1. The proton affinity of the fully relaxed lysine is 236.6 kcal/mol.

lysine, divided into the following EFP/buffer/ab initio regions $(CO_2H)(NH_2)CH$-/CH_2CH_2/-$CH_2CH_2NH_2$, as an example.

A. Construction of the Buffer Region. The RHF/6-31G* optimized structure of protonated lysine ($LysH^+$) is obtained and the MOs are localized using the Edmiston-Ruedenberg localization scheme.[23] The LMOs that will comprise the buffer are selected and projected, using the corresponding orbital method, so that they only span basis functions on the atoms in the buffer region.

The best source of buffer LMOs is presumably LMOs calculated for the entire molecule. In section IV, we will also consider a case where the buffer was taken from *n*-butane.

B. Construction of the EFP. The density of the molecular region that will be described by the EFP is re-optimized in the presence of the buffer region but in the absence of the ab initio region. The electrostatic potential of the re-optimized density, but not the buffer density, is expanded in terms of multipoles through octupoles centered at all atomic and bond midpoint centers using Stone's Distributed Multipole Analysis.[3] Calculated in this way, these multipoles do not account for polarization of the EFP region due to the ab initio region, so that this effect is not double counted when dipole polarizabilities are added. Furthermore, the multipoles describe a charge distribution with

net integer charge, so that the entire EFP/buffer/ab initio description has a net integer charge as well.

The dipole polarizability due to each LMO in the re-optimized EFP region is calculated analytically. The use of dipole polarizability tensors calculated using finite difference[24] and the complete neglect of this term are also tested.

In this study we do not include any short-range interactions, such as exchange repulsion or charge penetration, between the ab initio and EFP regions. This allows us to focus on the role of the electrostatic potential on the proton affinity.

C. Calculation of the Proton Affinity. The EFP, buffer, and ab initio regions are combined for $LysH^+$ and the geometry of the ab initio region is re-optimized. In a second calculation the proton is removed and the ab initio region geometry is re-optimized. The energy difference between these two systems is taken to be the proton affinity.

Restricted Hartree−Fock and MP2 calculations on lysine and Gly-Lys-Gly were performed using the 6-31G* basis set[25] with a locally modified version of GAMESS.[19] The Edmiston-Ruedenberg procedure was largely used throughout this work to generate localized orbitals.[23] The Foster-Boys localization was also tested.[26] The core orbitals were included into the orbital localization.

QM/MM Boundaries Across Covalent Bonds

J. Phys. Chem. A, Vol. 104, No. 28, 2000 **6661**

IV. Applications

A. Proton Affinity of Lysine. As an initial application of our method, we calculate the proton affinity (PA) of the ϵ-N in the amino acid lysine. The addition or removal of a full charge is a relatively large perturbation on the whole system, and the accurate representation of the environmental effects on this reaction should provide a stringent test of the accuracy of our approach.

Table 1 lists the PAs calculated with buffer regions (constructed as outlined in section III) at increasingly larger distances from the ϵ-N. The buffer regions, in bold and boxed in Table 1, represent LMOs calculated for the RHF/6-31G* optimized structure of protonated lysine, and truncated by projection. The buffer region in the first row of the table, for example, consists of four CH bonds, one CC bond, and two C 1s core MOs. The geometry of the buffer and EFP regions (always to the left of the buffer region) is thus taken from the RHF/6-31G* optimized geometry of protonated lysine.

The target value is the PA calculated using fully relaxed RHF/6-31G* wave functions and geometries of the protonated and unprotonated form for lysine, 236.6 kcal/mol. It is important to separate the error introduced by approximating part of the electronic charge distribution with an EFP and buffer, from the error introduced by the geometric constraints on those regions. This is accomplished by calculating the PA using RHF/6-31G* for the entire molecule, but only partial geometry optimization with the same geometrical constraints as in the EFP calculations. The values are listed in the first column of Table 1, and shows that the geometrical constraints introduce an error of 0.6 kcal/mol. In the subsequent discussion, we take these constrained all-ab initio calculations as our reference for the corresponding EFP calculations.
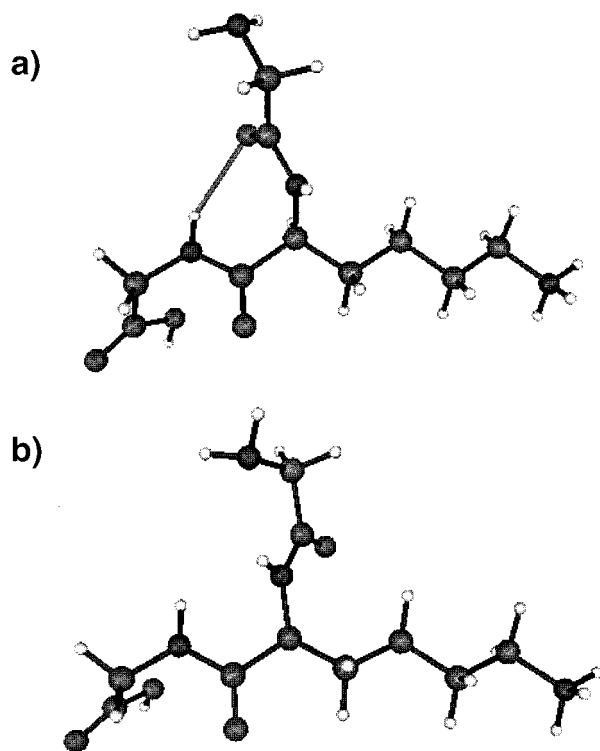
Column 3 of Table 1 lists the results obtained for the EFP/buffer/ab initio calculation. It is evident that the PA converges relatively quickly to within 0.2 kcal/mol of the all-ab initio reference value. Column 2 lists the corresponding PA values without the EFP to isolate the effect of the EFP region of the molecule on the PA, which can be as large as 2.6 kcal/mol for this system. The C_α-C_β bond and the associated CH and core LMOs ($[\alpha\beta]$-buffer) appear to be the optimum choice for the buffer region since this region is relatively nonpolarizable and far from the protonation site.

The last entry in Table 1 indicates a structural collapse of the ab initio region onto the EFP region. This is presumably due to the lack of a repulsive potential combined with a rather unphysical division of the molecule into EFP/buffer/ab initio regions, due to the small size of the molecule. Next, we consider a larger system.

B. Proton Affinity of the Tripeptide Gly-Lys-Gly. Further tests of the EFP/buffer/ab initio method were performed by computing the ϵ-N PAs of two different conformations of the Gly-Lys-Gly tripeptide: one with an intramolecular hydrogen bond and one without (Figure 2). The latter undergoes a larger conformational change in the EFP region and is therefore a more stringent test. The results are summarized in Tables 2 and 3, respectively.

The first columns of both Tables show that the presence of the intermolecular hydrogen bond reduces the effect of conformational rearrangement on the PA by 60−70%. This is a promising result, given the large number of intermolecular hydrogen bonds in proteins.

The second columns of both Tables demonstrate that molecular environment can have a significant effect (up to 7.2 kcal/

Figure 2. Structures of the two Gly-Lys-Gly tripeptide conformations used in this study: (a) one with an intramolecular hydrogen bond and (b) one without.
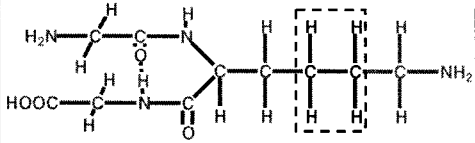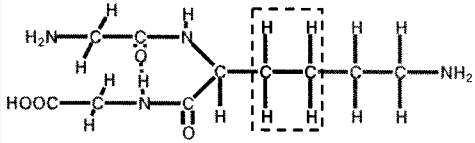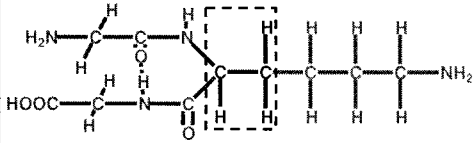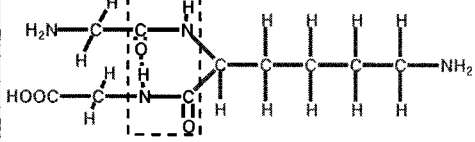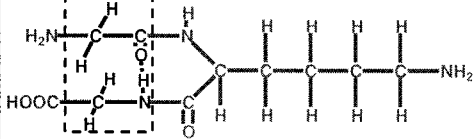
mol) on the PA of lysine. The effect is larger for the hydrogen-bonded conformation, despite the fact that it undergoes a smaller conformational change. The environmental effects are largely captured by the EFP representation, as shown by the data in the last columns. Again, the optimum choice of buffer region is the $[\alpha\beta]$-buffer, which for both conformations reduces the error to below 0.5 kcal/mol relative to the constrained all-ab initio calculation. Moving the buffer region further out on the backbone increases the absolute error to 0.8−2.0 kcal/mol, presumably since that region is more polarizable. Thus, since the EFP region is polarizable, it is not necessarily as worse representation of the charge density than the all-ab initio buffer region.

C. Construction of Buffer Region and EFP Regions. Several relatively arbitrary choices went into the construction of the buffer and EFP region. This is not necessarily a problem provided that the results are insensitive to these choices, and in section we explore this issue for the PA of the hydrogen bonded Gly-Lys-Gly tripeptide calculated with the $[\alpha\beta]$-buffer. The results are displayed in Table 4.

The first row of Table 4 represents the same calculation as the third row in Table 2, and outlines some of the choices made for this particular calculation in the columns. The source of the buffer region is the protonated form of the tripeptide calculated using RHF/6-31G*. The multipoles expansion was performed at each atom and bond midpoint and truncated after octupoles. The dipole polarizability tensors were calculated analytically for each LMO. The LMOs were calculated using the Edmiston-Ruedenberg scheme, and projected using the corresponding orbital method.

Obtaining the buffer region and the EFP from the unprotonated form of the tripeptide has a small (0.2 kcal/mol) effect on the PA as shown in the second row. However, obtaining the buffer LMOs from a calculation of butane, increases the absolute error by 0.4 kcal/mol. Thus, deriving the buffer MOs from a

TABLE 2: Same as In Table 1, for H-bonded Gly-Lys-Gly Tripeptide in kcal/mol[a]

	1 Reference ab initio calculation	2 QM/buffer calculation	3 QM/buffer/EFP calculation
(structure)	231.9 0.0	239.1 +7.2	234.1 +2.2
(structure)	231.8 0.0	237.3 +5.5	232.5 +0.7
(structure)	231.8 0.0	237.8 +6.0	232.2 +0.4
(structure)	231.6 0.0	234.0 +2.4	232.9 +1.3
(structure)	231.4 0.0	233.5 +2.1	230.4 −1.0

[a] The proton affinity of the fully relaxed tripeptide is 231.0 kcal/mol.

system that most closely resembles the system of interest is a worthwhile investment of computer time. The PA is relatively insensitive to numerical calculation of the dipole polarizability tensors and switching the localization scheme to Boys, or truncation by simply zeroing out of unwanted AO coefficients followed by renormalization. The largest change is for the latter scheme and actually reduces the error slightly (by 0.2 kcal/mol). Thus, the methodology used in the previous sections is relatively insensitive to choices of to calculate the buffer and EFP regions, as long as the source of buffer MOs is not too different from the system of interest.

The fact that the dipole polarizabilities can be calculated numerically or analytically without affecting the PA could derive from the fact that they make a negligible contribution to the PA. However, if the polarizabilities are removed completely the absolute error increases by 1.3 kcal/mol as evidenced by the data in the sixth row of Table 4. Furthermore, truncating the multipole expansion after monopoles, increases the absolute error further, to 3.1 kcal/mol. When the multipole expansion is redone using only atomic centers (to approximate current force field representation of the electrostatic potential) the error decreases to 1.1 kcal/mol. This is presumably due to a
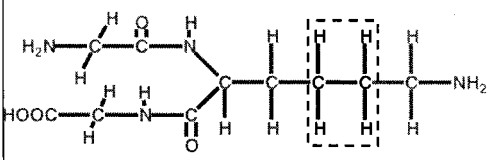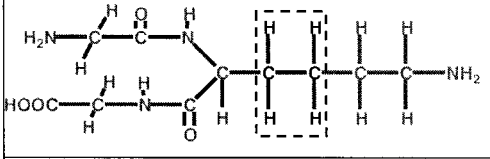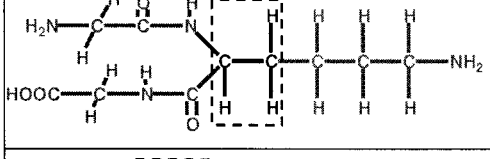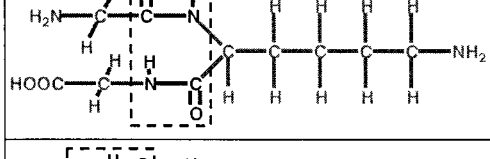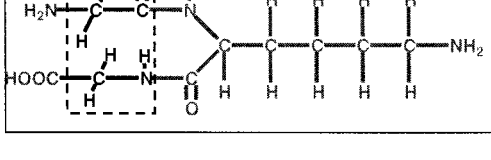
cancellation of errors since the use of more expansion points should give a more accurate representation of the ab initio electrostatic potential.

The best possible atomic centered monopole expansion is presumably that obtained by the Potential Derived Charges (PDCs) method,[27] since they are fit to the electrostatic potential. Row nine of Table 4 shows that these charges actually increase the error further (to 2.4 kcal/mol). These charges are calculated for the entire molecule, and therefore reflect the polarization of the EFP region by the ab initio region. However, they cannot describe the *change* in polarization upon deprotonation, and this is presumably the source of the increase in error. Furthermore, using polarizability tensors in conjunction with the (polarized) PDC charges (row ten) increases the error because the polarization effect is double counted.

The effect of using charges from standard force fields could not be tested for this molecule since charges for the neutral C and N termini are were unavailable. However, we did test a set of empirical charges due to Gasteiger and Marsili,[28] where the charges are assigned based solely on atomic number and connectivity. The resulting PA, however, is now in error by 6.3 kcal/mol (row 11).

QM/MM Boundaries Across Covalent Bonds

J. Phys. Chem. A, Vol. 104, No. 28, 2000 **6663**

TABLE 3: Same as In Tables 1 and 2, but for Non-H-bonded Gly-Lys-Gly Tripeptide[a]

	1 Reference ab initio calculation	2 QM/buffer calculation	3 QM/buffer/EFP calculation
	237.1 0.0	239.5 +2.4	239.3 +2.2
	237.0 0.0	237.4 +0.4	237.7 +0.7
	237.0 0.0	237.7 +0.7	237.3 +0.3
	236.7 0.0	237.9 +1.2	237.5 +0.8
	236.5 0.0	236.6 +0.1	234.5 −2.0

[a] The proton affinity of the fully relaxed tripeptide is 235.2 kcal/mol.

TABLE 4: Comparison of PA Errors Obtained Using Different MM Schemes, for the Hydrogen-Bonded Gly-Lys-Gly Tripeptide Using the [$\alpha\beta$]-Buffer

	source of buffer & geometry[a]	electrostatics	analytical or numeric polarizabilities	truncation[f]	localization[g]	error (kcal/mol)
1	RNH_3^+	$q - \Omega^b$	analytical	projection	Rued.	+0.4
2	RNH_2	$q - \Omega^b$	analytical	projection	Rued.	+0.6
3	butane, $RNH_3^{+\,h}$	$q - \Omega^b$	analytical	projection	Rued.	−0.8
4	RNH_3^+	$q - \Omega^b$	numeric	projection	Boys	+0.3
5	RNH_3^+	$q - \Omega^b$	numeric	zeroing	Boys	+0.2
6	RNH_3^+	$q - \Omega^b$	none	projection	Rued.	−1.3
7	RNH_3^+	q^c	none	projection	Rued.	+3.1
8	RNH_3^+	q^c (no bond midpoints)	none	projection	Rued.	+1.1
9	RNH_3^+	PDC[d]	none	projection	Boys	+2.4
10	RNH_3^+	PDC[d]	analytical	projection	Boys	+4.3
11	RNH_3^+	Gasteiger-Marsili[e]	none	projection	Boys	+6.3
12	RNH_3^+ (MP2)[i]	$q - \Omega^b$	analytical	projection	Rued.	−0.4

[a] The buffer LMOs and MM region geometry can be taken either from protonated or unprotonated form. [b] Charges, dipoles, quadrupoles and octupoles generated by Stone distributed multipole analysis located at atomic centers and bond midpoints. [c] Same as *a*, but only monopoles (charges) are included. [d] Potential determined charges (atomic charges fitted to electrostatic potential) calculated using GAMESS. [e] Gasteiger-Marsili empirical charges. [f] Truncation of LMO tails when forming buffer: projection or plain zeroing out. [g] Localization procedure: Edmiston-Ruedenberg or Boys. [h] The buffer was taken from butane, $CH_3-CH_2CH_2-CH_3$, the ab initio and EFP regions taken from line 1. [i] Single point MP2 calculation at the HF geometry; the frozen orbitals were excluded from the excitation space.

The data discussed in the preceding paragraphs suggest that great care must be taken when constructing a multipole representation of a molecular electrostatic potential. Furthermore, an atom centered monopole expansion may not be sufficient if an accuracy of less than 1 kcal/mol is required.

Finally, the last entry in Table 4 demonstrates that the MP2/6-31G* correlation correction to the PA can be calculated with similar accuracy. As mentioned previously, this is accomplished simply by excluding excitations from the chemically inert buffer region. The result can be compared to all-MP2/6-31G* single

TABLE 5: Comparison of PA Errors Obtained Using Different MM Schemes, for the Hydrogen-Bonded Gly-Lys-Gly Tripeptide Where the Buffer Is a Single $C_\alpha - C_\beta$ Bond[a]

	source of buffer & geometry[b]	electrostatics	analytical or numeric polarizabilities	truncation[e]	localization[f]	error (kcal/mol)
1	RNH_3^+	none	none	zeroing	Boys	+6.5
2	RNH_3^+	$q - \Omega^c$	numeric	zeroing	Boys	−2096[g]
3	RNH_3^+	$q - \Omega^c$	none	zeroing	Boys	+23.9[g]
4	RNH_3^+	q	none	zeroing	Boys	+1.9
5	RNH_3^+	q (no bond midpoints)	none	zeroing	Boys	+0.4
6	RNH_3^+	PDC^d	none	zeroing	Boys	+1.1

[a] Positions of the same atoms as in the $[\alpha\beta]$-buffer case were constrained. [b] The buffer LMOs and MM region geometry can be taken either from protonated or unprotonated form. [c] Charges, dipoles, quadrupoles and octupoles generated by Stone distributed multipole analysis located at atomic centers and bond midpoints. [d] Potential determined charges (atomic charges fitted to electrostatic potential) calculated using GAMESS, charges scaled to make the total charge of the MM region to be equal to 0. [e] Truncation of LMO tails when forming buffer: projection or plain zeroing out. [f] Localization procedure: Edmiston-Ruedenberg or Boys. [g] Error in proton affinity from single point energy calculation using the optimized geometry of row 4, column 3, Table 2.

point calculation of the reference system's PA (230.3 kcal/mol), and the resulting error of 0.4 kcal/mol is essentially equal to that of the underlying RHF calculation (0.3 kcal/mol).

D. Test of Single-LMO Buffer. Previous implementations of the LMO buffer boundary method have all been tested for systems where the buffer is a single CC bond.[12,14] In this section, we test the use of several representations of the Gly-Lys-Gly system used in the previous section but with a buffer consisting only of the truncated LMO connecting C_α and C_β. The results are presented in Table 5.

The first row of Table 5 reflects the intrinsic PA of the buffer/ab initio system and indicates that the EFP region shifts the PA by 6.5 kcal/mol, compared to 6.0 kcal/mol when the $C_\alpha H$ bond is included in the buffer. However, upon inclusion of our default EFP representation, the PA diverges by several thousand kcal/mol. In fact this value could only be obtained by computing *single point energies* of the structures obtained by the larger $[\alpha\beta]$-buffer. Any geometry optimization resulted in structural collapse. The source of the divergence is easily attributed to the polarizability terms by removing them and recalculating the PA-shift (again via single point energy calculations), which drops to 23.9 kcal/mol.

We note that since both values are obtained by single point energy calculations, they likely result from electron density "being pulled" toward the EFP region by the induced dipoles and higher order multipoles. This is evidenced by a very large ($|-21.6|$) Mulliken charge on the C_α buffer-atom compared to the value from the larger buffer region (0.1). Thus, introducing repulsive potentials that only depend on internuclear distances (such as the $1/r^{12}$ atom-pair potentials) is unlikely to prevent these large errors. Rather a potential that keeps the electron density out of the EFP region is needed if dipole polarizabilities and higher order multipoles are to be used in conjunction with a single-LMO buffer.[29]

The 23.9 kcal/mol error can be reduced by eliminating the higher order multipoles and, further, by redoing the multipole expansion only at the atomic centers. The final error of 0.4 kcal/mol is roughly the same as with the larger $[\alpha\beta]$-buffer. However, decreasing the error by using less accurate representations of the electrostatic potential is likely a result of fortuitous error cancellation. Indeed, similar calculations on lysine and the non-hydrogen bonded tri-peptide conformation yield errors of 1.9 and 3.4 kcal/mol, respectively. Finally, the use of PDCs increases the error by 0.6 kcal/mol relative the atom-centered charges from the Distributed Multipole Analysis. As with the $[\alpha\beta]$-buffer this is presumably due to the use of charges that were polarized by the ab initio charge distribution of the

protonated form of the lysine side chain. In the calculations described in this paragraph, the positions of the three hydrogens attached to C_β and C_β were constrained to avoid structural collapse of the ab initio region, while the position of the rest of the ab initio atoms were optimized.

V. Conclusions and Future Directions

A frozen localized molecular orbital-based approach for treating boundaries between molecular regions described by ab initio electronic structure and effective fragment potentials (EFPs) is presented. The approach has been implemented for RHF, ROHF, GVB, and MP2 energies as well as RHF and ROHF gradients.

We test our approach by calculating the proton affinity (PA) of the ϵ-N of lysine and the tripeptide Gly-Lys-Gly using the hybrid EFP/buffer/ab initio method for a variety of buffer sizes and positions. Comparison to all-ab initio calculations reveal that the optimum buffer consists of the bond LMO connecting the C_α and C_β plus the associated CH and core LMOs ($[\alpha\beta]$-buffer). This buffer in combination with an ab initio description of the lysine side chain and an EFP description of the rest, results in PAs that are consistently within 0.4 kcal/mol of the all ab initio reference value. Since the internal geometry of the EFP is fixed, the ab initio reference values are obtained from all ab initio calculations with the same geometrical constraints as in the EFP calculations. For the $[\alpha\beta]$-buffer the PA shift induced by geometrical rearrangements range between 0.6 and 1.8 kcal/mol and is always larger than the errors due to the approximate treatment of part of the charge density.

The EFP representation of the electrostatic potential consists of multipole expansions at each atom and bond midpoints truncated after octupoles, plus induced dipole LMO-polarizabilities centered at the centroid of charge of each LMO. The expansions are obtained from a separate ab initio calculation on a system in which the ab initio region has been removed. This is done in order to avoid a prepolarization of the EFP multipole expansion, which would be double counted by the induced polarizabilities. This representation is shown to be essential to consistently obtain a 0.4 kcal/mol accuracy.

The results are relatively insensitive to the choice of localization procedure, truncation method, analytical versus numerical calculation of the polarizability tensors, and source of buffer LMOs. The latter assumes that the buffer LMOs derive from a source that relatively closely resembles the system of interest. Furthermore, smaller buffer regions such as a single bond-LMO, does not provide adequate separation between the EFP and ab

QM/MM Boundaries Across Covalent Bonds

J. Phys. Chem. A, Vol. 104, No. 28, 2000 **6665**

initio regions and can lead to a large influx of electron density into the buffer region.

It is important to note that the current implementation of the LMO buffer method does not contain any adjustable parameters. Rather the buffer is derived *automatically* from a single ab initio calculation (see Figure 1), and it does not take significantly more human effort to compute the buffer LMOs from a tripeptide than from *n*-butane (for example). Thus, we advocate the construction of a buffer region (and EFP region) for each problem of interest rather than establishing a library of such buffer LMOs.

All molecules in the present study are small enough to allow full ab initio calculations, to gauge the accuracy of the new methodology. This also eased the construction of the EFPs since they could be derived from a single ab initio calculation. We are currently applying the EFP/buffer method to larger proteins (Turkey ovomucoid third domain and α-chymotrypsin) where the EFPs must be constructed from a series of ab initio calculations on smaller overlapping pieces. We will report on these first principles hybrid calculations on proteins in a future paper.

References and Notes

(1) Recent reviews can be found in *The Encyclopedia of Computational Chemistry* (Schleyer, P. v. R. et al., Eds.; J. Wiley & Sons: New York, 1998): (a) Froese, R. D. J.; Morokuma, K. "Hybrid methods", pp 1244−1257. (b) Gao, J. "Hybrid quantum mechanical/molecular mechanical (QM/MM) methods", pp 1257−1263. (c) Merz, K. M., Jr.; Stanton, R. V. "Quantum mechanical/molecular mechanical (QM/MM) coupled potentials", pp 2330−2343. (d) Tomasi, J.; Pomelli, C. S. "Quantum mechanics/molecular mechanics", pp 2343−2350. (e) See also: Monard, G.; Merz, K. M., Jr. *Acc. Chem. Res.* **1999**, *32*, 904.

(2) (a) Jensen, J. H.; Day, P. N.; Gordon, M. S.; Basch, H.; Cohen, D.; Garmer, D. R.; Krauss, M.; Stevens, W. J. In *Modeling the Hydrogen Bond*; Smith, D. A., Ed. ACS Symposium Series 569; American Chemical Society: Washington, DC, 1994; Chapter 9. (b) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Kraus, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.

(3) Stone, A. J *Chem. Phys. Lett.* **1981**, *83*, 233.

(4) (a) Chen, W.; Gordon, M. S. *J. Chem. Phys.* **1996**, *105*, 11081. (b) Krauss, M.; Webb, S. P. *J. Chem. Phys.* **1997**, *107*, 5771. (c) Day, P. N.; Pachter, R. *J. Chem. Phys.* **1997**, *107*, 2990. (d) Merrill, G. N.; Gordon, M. S. *J. Phys. Chem. A* **1998**, *102*, 2650. (e) Webb, S. P.; Gordon, M. S. *J. Phys. Chem. A* 1999 **103**, *1265*. (f) Petersen, C. P.; Gordon, M. S. *J. Phys. Chem. A* **1999**, *103*, 4162. (g) Day, P. N.; Pachter, R.; Gordon, M. S.; Merrill G. N. *J. Chem. Phys.* **2000**, *112*, 2063.

(5) (a) Wladkowski, B. D.; Krauss, M.; Stevens, W. J. *J. Am. Chem. Soc.* **1995**, *117*, 10537. (b) Krauss, M *Computers and Chemistry* **1995**, *19*, 199. (c) Krauss M.; Wladkowski B. D. *Int. J. Quantum Chem.* **1998**, *69*, 11.

(6) (a) Jensen, J. H. *J. Chem. Phys.* **1996**, *104*, 7795. (b) Jensen, J. H.; Gordon, M. S. *Mol. Phys.* **1996**, *89*, 1313. (c) Jensen, J. H.; Gordon, M. S. *J. Chem. Phys.* **1998**, *108*, 4772.

(7) Kairys, V.; Jensen, J. H. *Chem. Phys. Lett.* **1999**, *315*, 140.

(8) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys.* **2000**, *112*, 7300.

(9) (a) Warshell, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227. (b) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986** 7, 718. (c) Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. *Int. J. Quantum Chem.* **1996**, *60*, 1189.

(10) Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, *102*, 4714.

(11) Zhang, Y.; Lee, T.-S.; Yang, W. T. *J. Chem. Phys.* **1999**, *110*, 46.

(12) (a) Théry, V.; Rinaldi, D.; Rivail, J.-L.; Maigret, B.; Ferenczy, G. G. *J. Comput. Chem.* **1994**, *15*, 269. (b) Monard, G.; Loos, M.; Théry, V.; Baka, K.; Rivail, J.-L. *Int. J. Quantum Chem.* **1996**, *58*, 153. (c) Gorb, L. G.; Rivail, J.-L.; Théry, V.; Rinaldi, D. *Int. J. Quantum Chem. Symp.* **1996**, *30*, 1525. (d) Assfeld, X.; Rivail, J.-L. *Chem. Phys. Lett.* **1996**, *263*, 100.

(13) Reuter, N.; Dejaegere, A.; Maigret, B.; Karplus, M. *J. Phys. Chem. A* **2000**, *104*, 1720.

(14) Philipp D. M.; Friesner, R. A. *J. Comput. Chem.* **1999**, *20*, 1468.

(15) Katsuki, S.; Huzinaga, S. *Chem. Phys. Lett.* **1988**, *152*, 203.

(16) McWeeney, S. *Proc. R. Soc.* **1959** *A253*, 52.

(17) Stevens, W. J.; Fink, W. H. *Chem. Phys. Lett.* **1987**, *139*, 15.

(18) Bagus, P. S.; Hermann, K.; Bauschlicher, C. W., Jr. *J. Chem. Phys.* **1984**, *80*, 4378.

(19) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

(20) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. *Chem. Phys. Lett.* **1990**, *166*, 275.

(21) Yamaguchi, Y.; Masamura, Y.; Goddard, J. D.; Schaefer, H. F. *A New Dimension, A New Dimension to Quantum Chemistry*; Oxford University Press: Oxford, 1994.

(22) King, H. F.; Stanton, R. E.; King, H.; Wyatt, R. E.; Parr R. G. *J. Chem. Phys.* **1967**, *47*, 1936.

(23) Edmiston, C.; Ruedenberg, K. *Rev. Mod. Phys.* **1963**, *35*, 457.

(24) Garmer, D. R.; Stevens, W. J. *J. Phys. Chem.* **1989**, *93*, 8263.

(25) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.

(26) Boys, S. F. *Quantum Science of Atoms, Molecules and Solids*; Lowdin, P. O., Ed.; Academic Press: New York, 1966.

(27) (a) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129. (b) Chirlain, L. E.; Francl, M. M. *J. Comput. Chem.* **1987**, *8*, 894. (c) Woods, R. J.; Khalil, M.; Pell, W.; Moffatt, S. H.; Smith, V. H. *J. Comput. Chem.* **1990**, *11*, 297. (d) Breneman, C. M.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361. (e) Merz, K. M. *J. Comput. Chem.* **1992** *13*, 749. (f) Spackman, M. A. *J. Comput. Chem.* **1996**, *17*, 1.

(28) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.

(29) We note that the method by Philipp and Friesner involves the use of certain frozen virtual MOs to restrict the flow of electron density into the MM region. It is possible that this will significantly reduce the errors, but we have not tested this.