PLoS one

The Effective Fragment Molecular Orbital Method for Fragments Connected by Covalent Bonds

Casper Steinmann[1]*, Dmitri G. Fedorov[2], Jan H. Jensen[1]

1 Department of Chemistry, University of Copenhagen, Copenhagen, Denmark, 2 NRI, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan

Abstract

We extend the effective fragment molecular orbital method (EFMO) into treating fragments connected by covalent bonds. The accuracy of EFMO is compared to FMO and conventional *ab initio* electronic structure methods for polypeptides including proteins. Errors in energy for RHF and MP2 are within 2 kcal/mol for neutral polypeptides and 6 kcal/mol for charged polypeptides similar to FMO but obtained two to five times faster. For proteins, the errors are also within a few kcal/mol of the FMO results. We developed both the RHF and MP2 gradient for EFMO. Compared to *ab initio*, the EFMO optimized structures had an RMSD of 0.40 and 0.44 Å for RHF and MP2, respectively.

Introduction

The need to study very large systems in an efficient manner has led to the development of many computational schemes trying to cope with the limitation in computational resources. Linear (or nearly linear) scaling methods have long been of particular interest because they allow, within their respective framework [1–11], large systems to be treated by quantum mechanics. In particular, the use of fragments [12,13] is very attractive for doing calculations of large systems.

Recently, we developed the effective fragment molecular orbital (EFMO) method [14], which builds upon the fragment molecular orbital (FMO) method [15–20], and combines it with effective fragment potentials (EFP) [21–23]. EFMO is different from EFP, FMO and FMO/EFP [24,25] in several ways. For instance, the EFPs are computed on-the-fly from gas phase FMO fragment calculations and used for classical interactions of separated dimers and many-body effects. Extending the earlier work [14] limited to molecular clusters at the RHF level, we now present the methodology to treat fragments connected by covalent bonds at the MP2 level.

This article is organized as follows. First, we briefly outline the theoretical background of EFMO. We proceed to discuss the change in methodology needed to include fragmentation across covalent bonds in EFMO, including an overview of how fragment bonds are treated. The addition of correlation in EFMO is also presented here. Second, we benchmark the EFMO energy against *ab initio* calculations on three different sets of polypeptides and compare to FMO. We apply our findings to proteins and protein like structures. The quality of the gradient together with timings are also presented here. Water clusters are also briefly revisited. Finally, we summarize our results and discuss future directions.

Methods

Theoretical Background

In FMO, the total two-body (FMO2) non-correlated energy of a system consisting of N fragments (also called monomers) is given as

$$E^{\mathrm{FMO2}} = \sum_I^N E_I + \sum_{IJ}^N (E_{IJ} - E_I - E_J) \qquad (1)$$

Here E_I (E_{IJ}) is the energy of monomer I (dimer IJ) in the electrostatic potential (ESP) of the other $N-1$ ($N-2$) fragments. The monomers converge in the field of ESP, requiring self-consistent charge (SCC) iterations. Dimers converge in the field of ESP of the $N-2$ monomers.

The total non-correlated EFMO energy of a system of N fragments is

$$E^{\mathrm{EFMO}} = \sum_I^N E_I^0 + \sum_{IJ}^{R_{I,J} \leq R_{\mathrm{resdim}}} (E_{IJ}^0 - E_I^0 - E_J^0 - E_{IJ}^{\mathrm{POL}})$$
$$+ \sum_{IJ}^{R_{I,J} > R_{\mathrm{resdim}}} E_{IJ}^{\mathrm{ES}} + E_{\mathrm{tot}}^{\mathrm{POL}} \qquad (2)$$

where E_I^0 is the gas phase energy of monomer (or fragment) I. E_{IJ}^0 is the gas phase dimer energy of dimer IJ. The second sum in equation 2 is the pairwise correction to the monomer energy and only applies for dimers separated by a distance less than R_{resdim}.

E_{IJ}^{POL} and $E_{\text{tot}}^{\text{POL}}$ are the classical pair polarization energy of dimer IJ and the classical total polarization energy, respectively. The final sum over E_{IJ}^{ES} is the classical electrostatic interaction energy and applies to dimers separated by a distance greater than R_{resdim}. The fragment separation distance $R_{I,J}$ was defined previously [14]. Since EFMO only involves gas phase energy (and gradient) evaluations, only one SCC iteration is required.

In EFMO, the classical terms in the energy expression (equation 2) are calculated from expressions in the EFP perturbation expansion of the interaction energy [21,22]. Based on the converged fragment calculations, EFP parameters are derived on-the-fly completely automatically by computing atom centered monopoles, dipoles, and quadrupoles [26] and dipole polarizability tensors for each electron pair. [27].

The analytical gradient derived previously [14] is reformulated for fragments connected by covalent bonds, and also extended to MP2.

Covalent Bonds

For fragmentation across covalent bonds, no corrections to the basic equation of EFMO is needed. However, the inclusion of fragmentation across bonds requires a change in the methodology. In this paper, we show how fragmentation is carried out on protein backbones, this methodology is transferable to other systems just as FMO was applied to inorganic systems such as zeolites [28] and nanowires [29].

In regular FMO, two different schemes of fragmentation is possible. Common to both is that one specifies pairs of atoms which defines fragment boundaries (Figure 1). Each detached bond is made of a bond attached atom (BAA) and a bond detached atom (BDA). The latter donates an electron to the fragment containing the BAA. One scheme is the hybrid orbital projection (HOP) approach [16], which allows full variational treatment of molecular orbitals (MO) across the bond during the fragment SCF. The other is the adapted frozen orbital (AFO) method [28,29] which freezes the occupied orbital that describes the bond [30]. EFMO uses the latter method, and for completeness we include a discussion of this particular scheme in this work.

In AFO, a model system around the BAA and BDA is constructed (Figure 2). RHF calculations are carried out on this system, followed by an Edminston-Ruedenberg localization [31]. The occupied orbital which has the largest overlap with the BDA and BAA is identified as the special bond orbital (SBO) shown on Figure 3. This orbital, along with several virtual orbitals on the BDA is stored for later use in monomer and dimer SCF calculations.
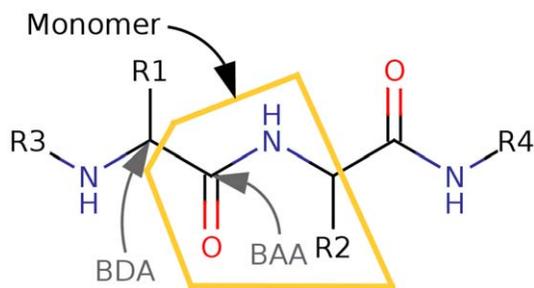
Figure 1. A model of a backbone in a protein. The model has side chains (R1 and R2) as well as the continuation of the backbone (R3 and R4). The bond attached atom (BAA) and the bond detached atom (BDA) face each other across the fragmentation point (marked with the yellow line). One fragment is shown within the yellow box.
doi:10.1371/journal.pone.0041117.g001
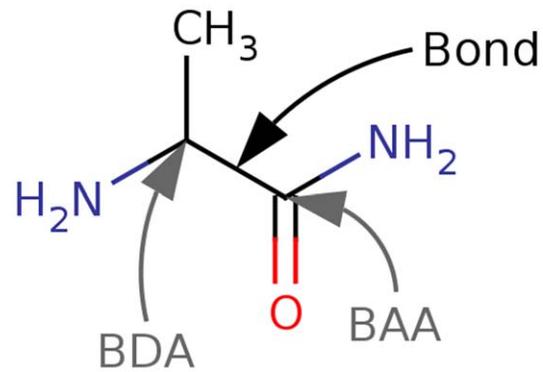
Figure 2. The current model system used in this study for fragmentation across peptide bonds. The model is constructed automatically for use with AFO. The central atoms are the bond attached atom (BAA) and the bond detached atom (BDA). The atoms which are connected directly to either the BAA or the BDA are included, capped with hydrogens as necessary.
doi:10.1371/journal.pone.0041117.g002
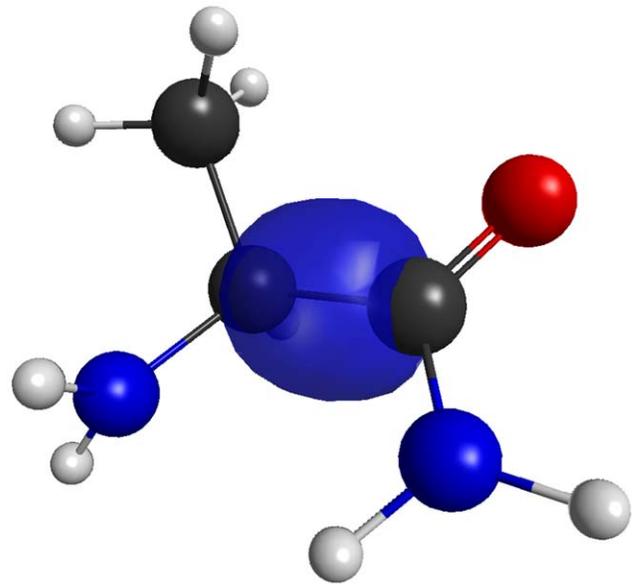
Figure 3. Special bond orbital for bond 13 in The Trp-cage protein. The orbital is obtained using RHF/6-31G(d) on a model system (Figure 2).
doi:10.1371/journal.pone.0041117.g003

For polypeptides, which is the main focus of this study, there is one SBO per pair of BAA and BDA. This SBO is associated with the fragment that contains the BAA. After the computation of all model systems, monomer calculations are done, followed by a Foster-Boys localization, where the SBO is kept frozen, i.e. not allowed to mix with the rest of the orbitals. This leads to a polarizable point in the centroid of the SBO (Figure 3), obtained from the model system across the bond (Figure 2). We have thus successfully eliminated the need to manually parametrize the bonds between pairs of fragments.

In the original formulation of EFMO, the electric field arising from a static multipole or induced dipole in fragment I is screened by a Tang-Toennis type expression.

$$k(\vec{R}, \alpha, \beta) = 1 - \exp\left(-\sqrt{\alpha\beta}|\vec{R}|^2\right)\left(1 + \sqrt{\alpha\beta}|\vec{R}|^2\right) \quad (3)$$

Here, α and β are the screening parameters associated with fragments I and J, respectively. The distance parameter $\vec{R}$ is the vector between an induced dipole in fragment I and any of the electric moments in fragment J. The above expression is also the default in EFP [21,22] with the parameters $\alpha = \beta = 0.6$. We emphasize that the screening parameters are associated with fragments and not individual polarizable points.

Correlation

The introduction of correlation energy in the EFMO method follows previous work in FMO [32–34]. The total correlated energy of a system of N fragments is given as.

$$E = E^{\mathrm{EFMO}} + E^{\mathrm{COR}}. \quad (4)$$

Here E^{COR} is given as the sum of monomer correlation energies E_I^{COR} and pairwise corrections, i.e.

$$E^{\mathrm{COR}} = \sum_I^N E_I^{\mathrm{COR}} + \sum_{IJ}^{R_{I,J} < R_{\mathrm{corsd}}} \left(E_{IJ}^{\mathrm{COR}} - E_I^{\mathrm{COR}} - E_J^{\mathrm{COR}}\right), \quad (5)$$

where E_{IJ}^{COR} is the correlation energy of dimer IJ. The distance parameter R_{corsd} determines whether or not correlation is included for a specific dimer. The value of the parameter is discussed in the computational methodology section below. Note that for the correlation energy any size-extensive post-HF scheme can be used.

Computational Methodology

All *ab initio* and fragment calculations were carried out in a locally modified version of GAMESS [35]. EFMO was parallelized with the generalized distributed data interface [36]. In all calculations, the 6-31G(d) [37–39] basis set was employed throughout unless specified otherwise. In all the geometry optimizations, a convergence criterion of $5.0 \cdot 10^{-4}$ Hartree/Bohr was used.

The *ab initio* MP2 calculations had their integral accuracy increased to 10^{-12} (ICUT = 12 in $CONTRL), SCF convergence criterion was raised from 10^{-5} to 10^{-7} (CONV = 1E-7 in $SCF) and the MP2 code by K. Ishimura *et. al* [40] with AO integral transformation threshold increased from 10^{-9} to 10^{-12} (CODE = IMS and CUTOFF = 1E-12 in $MP2) to match what is used in FMO.

For FMO (and EFMO), the AFO scheme was used throughout with the default settings for bond definitions (LOCAL = -RUEDNBRG in $CONTROL and RAFO(1) = 1,1,1 in $FMO). The parameters for the electrostatic treatment of dimers R_{resdim} and the threshold for the inclusion of correlation effects R_{corsd} were both set to 2.0 (RESDIM = 2.0 RCORSD = 2.0 in $FMO) unless otherwise specified. The distances are relative to the van-der-Waals radii of atoms (see ref [14] for details). The screening parameter for all fragments are set to 0.1 for fragments with and without the SBO (SCREEN(1) = 0.1,0.1 in $FMO), respectively unless specified otherwise.

The following structures used in this study were taken from previous work by Fedorov et. al. [32,34,41] This includes α-helices

$(\alpha-(\mathrm{ALA})_n)$ and β-sheets $(\beta-(\mathrm{ALA})_n)$ of alanine, Chignolin (PDB code: 1UAO) and the Trp-cage (PDB code: 1L2Y). Correlation effects on molecular clusters is carried out by investigating the structures from our previous study [14]. The crystal structure of the 42 residue protein Crambine (PDB code: 1CRN) is also included and protonated using the PDB2PQR tool [42,43].

The three polypeptides used in this study were constructed by selecting six neutral (at pH = 7) amino acids AIVGLT (P1) and AVSNTL (P2) as well as four neutral and two non-neutral (at pH = 7) residues AVKNTD (P3) and padded with two glycine residues at each end for a total peptide length of 10 residues. The polypeptides were protonated (at pH = 7) using the PDB2PQR tool. P1 had neutral termini (arguments –neutralc –neutraln) while P2 and P3 both had charged termini. For each polypeptide, a conformational search was carried out to locate twenty different structures using the ObConformer tool of the Open Babel package [44,45]. They were finally minimized using PM6 [46] in MOPAC [47] with a bulk solvent (EPS = 80.1).

Only results for two residues per fragment are discussed in detail below, and the results for one residue per fragment are shown in the supporting information (Table S1). We note that because of the large charge transfer in some charged systems the one residue per fragment division leads to very considerable errors.

When interpreting the accuracy of the results, the following quantities of errors are defined for energies. The error in energy.

$$\Delta E^{M,X} = E^M - E^X, \quad (6)$$

the average deviation of conformers

$$\Delta E_{\mathrm{avg}}^{M,X} = \frac{1}{N}\sum_I^N \left\{E_I^M - E_I^X\right\} \quad (7)$$

and the mean average deviation (MAD) for conformers

$$\Delta E_{\mathrm{MAD}}^{M,X} = \frac{1}{N}\sum_I^N |E_I^M - E_I^X|. \quad (8)$$

Here, M is FMO2/HOP, FMO2/AFO or EFMO and X is RHF or MP2. I runs through N conformers of polypeptides. To evaluate the quality of the EFMO gradient, numerical gradients (∇E^{num}) were calculated on α-(ALA)$_{10}$ and compared to its analytical counterpart (∇E^{ana}) by the root mean square (rms) deviation of the individual elements

$$\nabla E^{\mathrm{rms}} = \sqrt{\frac{|\nabla E^{\mathrm{ana}} - \nabla E^{\mathrm{num}}|}{3N_A}} \quad (9)$$

and the maximum deviation

$$\nabla E^{\mathrm{max}} = \max(|\nabla_i E^{\mathrm{ana}} - \nabla_i E^{\mathrm{num}}|). \quad (10)$$

N_A in equation 9 is the number of atoms in the molecule of interest, i in equation 10 runs through $3N_A$ atomic coordinates.

To measure the compactness of a protein we use the radius of gyration R_g^2 given as

$$R_g^2 = \frac{1}{N_A} \sum_{k=1}^{N_A} (\vec{r}_k - \vec{r}_{\text{mean}})^2. \qquad (11)$$

Results and Discussion

Application to Polypeptides

The performance of EFMO has a critical dependence on the screening parameter (equation 3, Figures S1, S2 and S3, and Tables S2 and S3) because of the close position of a) induced dipoles located at the centroid of the SBO in one fragment and b) the nearby electrostatic moments and induced dipoles in another (especially, adjacent) fragment. In the following, the screening parameter for all fragments is $\alpha = 0.1$ unless otherwise specified.

Figure 4 shows the MAD results obtained for two residues per fragment for all three polypeptides (P1, P2 and P3) using FMO2/HOP, FMO2/AFO and EFMO for both RHF and MP2. For P1, RHF MAD values are 0.82 kcal/mol, 0.94 kcal/mol and 2.02 kcal/mol for FMO2/HOP, FMO2/AFO and EFMO, respectively. The MP2 results yield 1.01 kcal/mol, 1.45 kcal/mol and 2.33 kcal/mol for P1 respectively.

For the charged polypeptide P2, MAD (Figure 4) increases by roughly a factor of two. The factor is about 3 for P3 (from 2.02 kcal/mol to 5.94 kcal/mol for the RHF energy). The inclusion of charged residues results in larger induced dipoles, which has a negative impact on the accuracy of the energy in EFMO. The accuracy of charged systems may be ameliorated by solvent screening. [48–50].

If one considers the average deviation (equation 7 and Figure 5) instead, it is interesting to note that EFMO compares well with FMO2, and the agreement for P3 is perhaps fortuitous (the error is

less than 0.5 kcal/mol for EFMO-MP2). The maximum deviations for EFMO, however, are larger in all cases by roughly a factor of two.

For all three peptide ensembles, there is a good correlation between the compactness of the peptide conformation (measured by the radius of gyration, equation 11) and the error in the energy (see supporting information Figures S4, S5 and S6). More compact structures place the charged groups closer to the polarizable points at the fragment boundaries resulting in large induced dipoles and errors in the total energy.

Application to Proteins

The above benchmark of EFMO serves as an initial probe for how the energy behaves for polypeptides as the number of residues per fragment and screening parameters change. Based on those tests, we now apply EFMO to proteins or protein-like structures. The alanine polypeptides are particularly good for studying any systematic error, albeit they are not a representative benchmark for real proteins.

In Table 1, deviations in EFMO energy of the various protein structures compared to *ab initio* RHF (MP2) are presented for two residues per fragment with cutoffs R_{resdim} and R_{corsd} both equal to 2.0. For Chignolin (1UAO), the deviation in energy for EFMO (equation 6) in RHF (MP2) energy is 1.79 (1.48) kcal/mol, and for FMO2/AFO it is 0.37 (1.38) kcal/mol. For the larger Trp-cage (1L2Y), the EFMO errors are −2.87 (−4.21) kcal/mol and for FMO2/AFO the values are 1.74 (6.35) kcal/mol. The Crambine protein (1CRN) had errors of 15.66 (26.23) kcal/mol for EFMO, which is comparable to the FMO2/AFO results of 3.45 (25.59) kcal/mol. EFMO shows the largest errors of a similar magnitude to FMO2/AFO. Using a 6-31+G(d) basis set on Chignolin, EFMO has the errors of 21.70 (−21.87) kcal/mol. FMO2 did not converge using the default settings.
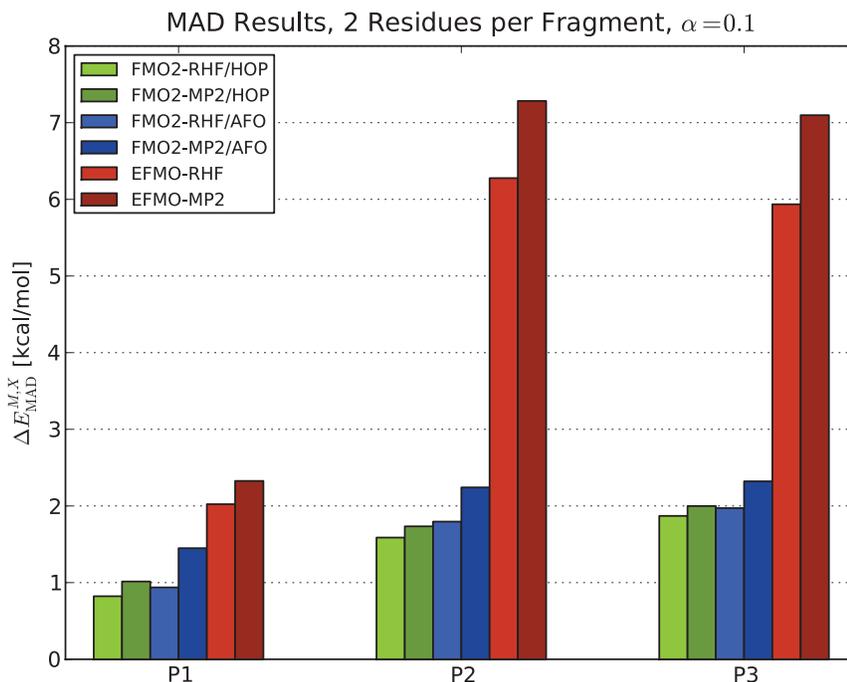
Figure 4. Mean average deviations of FMO2 and EFMO calculations. Results are compared to ab initio for conformers of the three polypeptides P1, P2 and P3 using two residues per fragment and the 6-31G(d) basis set. The screening parameter was set to $\alpha = 0.1$ for all calculations. Energies in kcal/mol.
doi:10.1371/journal.pone.0041117.g004
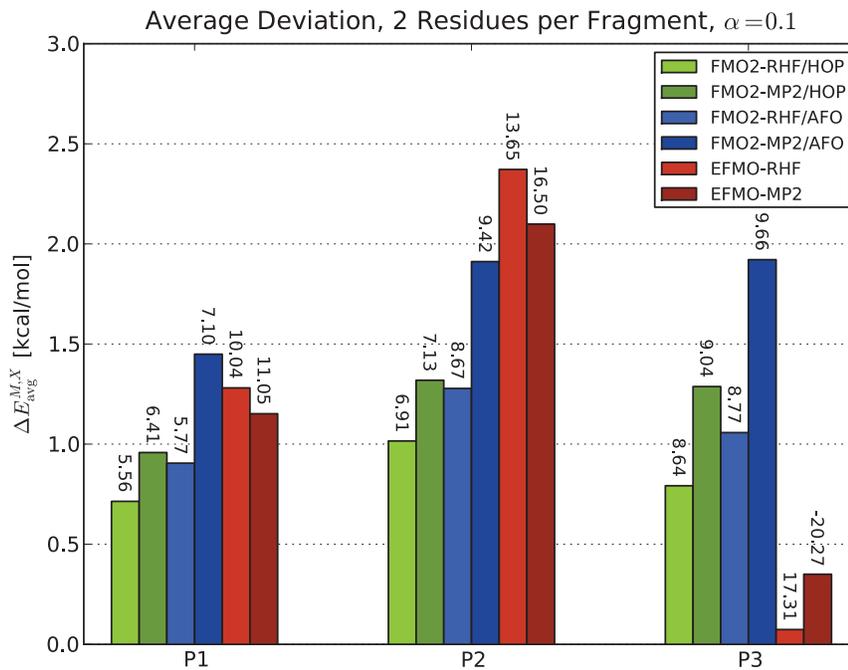
Figure 5. Average deviations of energy of FMO2 and EFMO calculations compared to RHF and MP2. All the three polypeptides P1, P2 and P3 using two residues per fragment are shown. Labels on the figure represent the maximum observed deviation. The screening parameter was set to $\alpha = 0.1$ for all calculations. Energies are in kcal/mol.
doi:10.1371/journal.pone.0041117.g005

The results from the α-helices and β-sheets are somewhat more detrimental. With the exception of the RHF EFMO results, the errors are roughly additive for the poly-alanine peptides, so the errors are discussed on a per residue basis. For α-helices, the error in energy increase with system size from -2.94 (0.32) kcal/mol for $\alpha-(\text{ALA})_{10}$ to 0.18 (-18.94) kcal/mol for the large $\alpha-(\text{ALA})_{40}$ helix, which corresponds to an average error per residue of 0.29 (0.03) kcal/mol for $\alpha-(\text{ALA})_{10}$ and less than 0.01 (-0.47) kcal/ mol for $\alpha-(\text{ALA})_{40}$. The α-helices tend to illustrate the case of

Table 1. Energy Error compared to *ab initio* calculations on proteins and protein-like structures using two residues per fragment.

	EFMO		FMO2/AFO	
	$R_{resdim} = 2.0$		$R_{resdim} = 2.0$	
	RHF	**MP2**	**RHF**	**MP2**
α-(ALA)$_{10}$	-2.94	0.32	-0.77	-0.08
β-(ALA)$_{10}$	0.60	0.89	0.08	0.25
α-(ALA)$_{20}$	-2.75	-9.66	-2.30	-0.53
β-(ALA)$_{20}$	1.74	2.78	0.22	0.71
α-(ALA)$_{40}$	0.18	-18.94	-5.47	-1.62
β-(ALA)$_{40}$	4.05	6.46	0.51	1.62
Chignolin	1.79	1.48	0.37	1.38
Trp-cage	-2.87	-4.27	1.74	6.35
Crambine[a]	15.66	26.23	3.45	25.59

[a]based on an FMO3-MP2/6-31G(d) calculation.
We used the 6-31G(d) basis set and $R_{resdim} = R_{corsd} = 2.0$. In all calculations, the screening parameter α was kept fixed at a value of $\alpha = 0.1$. All units in kcal/mol.
doi:10.1371/journal.pone.0041117.t001

over-polarization. For $\alpha-(\text{ALA})_{10}$, the total polarization energy is small (-12.89 kcal/mol) but as the system system size increase, so does the total polarization energy (-73.81 kcal/mol) in a non-linear fashion. We note that the MP2 energy for $\alpha-(\text{ALA})_{20}$ and $\alpha-(\text{ALA})_{40}$ increases linearly with system size but the RHF energy does not. The over polarization is also observed for FMO2/AFO, although the MP2 energies are much better (below 2 kcal/mol) which can only be attributed a better wave function of the individual fragments and their pairs. The β-sheets have errors which are lower than in the α-alanines the errors are from 0.60 (0.89) kcal/mol to 4.05 (6.46) kcal/mol for $\beta-(\text{ALA})_{10}$ and $\beta-(\text{ALA})_{40}$, respectively. Overall, the average error per residue becomes 0.06 (-0.50) kcal/mol and 0.10 (0.16) kcal/mol for $\beta-(\text{ALA})_{10}$ and $\beta-(\text{ALA})_{40}$, respectively. The β-sheets are planar and not prone to the same over-polarization (the $\beta-(\text{ALA})_{40}$ has a polarization energy of around 50 kcal/mol).

As noted above, the α-helices and β-sheets illustrate two very different polypeptides. The inaccuracy of EFMO for them is somewhat alleviated by the fact that the errors in energy for Chignolin and the Trp-cage proteins are smaller than the α-helices and β-sheets. The Trp-cage has 20 residues and its error in energy of -2.87 (-4.21) kcal/mol lie around the corresponding α-helices and β-sheets of the same size -2.75 (-9.66) kcal/mol to 1.74 (2.78) kcal/mol, respectively. The same is true for Chignolin.

Gradients and Geometry Optimizations

A key strength of EFMO over other similar methods [7–11] is the availability of the gradient. The gradient of FMO2/AFO has been investigated previously for zeolites [29] where errors in gradient were found to be ∇E^{rms}: $0.2 \cdot 10^{-3}$ Hartree/Bohr and ∇E^{max}: $1.4 \cdot 10^{-3}$ Hartree/Bohr when compared to numerical derivatives (equations 9 and 10) although with a smaller basis set than in this study. It was found, that even though these deviations

were present, geometry optimizations did result in satisfactory structures.

In this study, we present an investigation of the EFMO gradient comparing numerical and analytical values for proteins (Table 2). It has roughly the same accuracy-related issues found for zeolites, specifically around the bond regions where rms and maximum errors for FMO2-RHF/AFO with and without the electrostatic potential is $0.51 \cdot 10^{-3}$ Hartree/Bohr, $3.43 \cdot 10^{-3}$ Hartree/Bohr and $0.76 \cdot 10^{-3}$ Hartree/Bohr and $4.71 \cdot 10^{-3}$ Hartree/Bohr, respectively which is on par with what was found for zeolites. The latter result is particularly interesting as it is the FMO2/AFO result on top of which we add the EFP terms to obtain EFMO (equation 2).

Several different approaches to tackle the gradient were attempted. The first is the original approach taken for molecular clusters which is to transfer the gradient terms of the induced dipoles $\vec{\mu}^{\text{ind}}$ to the nearest atom only, in this study named EFMO$_{\text{org}}$. This is a clear improvement over the FMO2/AFO (without the ESP) result (∇E^{rms}: $0.73 \cdot 10^{-3}$ Hartree/Bohr, ∇E^{max}: $3.50 \cdot 10^{-3}$ Hartree/Bohr), but some deviations in gradient get worse using EFMO and will be discussed further below. Removing all torque contributions (EFMO$_{\text{nt}}$) reveals further improvements (∇E^{rms}: $0.68 \cdot 10^{-3}$ Hartree/Bohr, ∇E^{max}: $3.30 \cdot 10^{-3}$ Hartree/Bohr). Another approach, specifically for the induced dipole (EFMO$_{\text{nt}+\text{pct}}$) is to do a percentage based distribution of the induced dipoles based on the distance between two atoms (supporting information Text S1 and Figure S7). This only applies if the induced dipole is between two atoms and the gradient is distributed based on a percentage of the entire bond length. This further improves the results, but the improvement (∇E^{rms}: $0.66 \cdot 10^{-3}$ Hartree/Bohr, ∇E^{max}: $3.27 \cdot 10^{-3}$ Hartree/Bohr) reveals that the main source of the error is not due to EFMO (Figure 6), but pertains to approximations in the FMO2/AFO gradient. To make sure that the induced dipoles do not cause major problems, an approach was tried to not evaluate the electric field from the static multipole moments and the induced dipoles, both in the energy and the gradient, of adjacent fragments, that is fragment I covalently bound to fragment J does not induce dipoles in I and vice versa. Results with (EFMO$_{\text{nt}+\text{pct}+\text{adj}}$) and without (EFMO$_{\text{nt}+\text{adj}}$) percentage based distribution of induced dipoles are (∇E^{rms}: $0.66 \cdot 10^{-3}$ Hartree/Bohr, ∇E^{max}: $3.73 \cdot 10^{-3}$ Hartree/Bohr) and (∇E^{rms}: $0.66 \cdot 10^{-3}$ Hartree/Bohr, ∇E^{max} $3.74 \cdot 10^{-3}$ Hartree/Bohr) offer no clear advantage over EFMO$_{\text{nt}+\text{pct}}$ on the RHF level of theory, and consequently MP2 data are not presented.

From Figure 6, it is clear that EFMO fixes some of the issues that FMO2/AFO has, but evidently creates a few new ones at atom indices 111 (backbone nitrogen), 155 (backbone carbonyl), 231 (backbone nitrogen) and 236 (backbone C$_\alpha$). Common to all is that it is around the bonding region. Evidently, small perturbations in the geometry, specifically around the bonding region, has large implications for the generated EFP parameters. For FMO2-MP2/AFO and EFMO-MP2 (Figure 7 and Table 2), the errors in the gradient decrease for the EFMO$_{\text{nt}+\text{pct}}$ methodology (∇E^{rms}: $0.61 \cdot 10^{-3}$ Hartree/Bohr, ∇E^{max}: $2.89 \cdot 10^{-3}$ Hartree/Bohr) while FMO2-MP2/AFO errors are very similar to the corresponding RHF values.

Finally, geometry optimizations were carried out for α-(ALA)$_{10}$ using the 6-31G(d) basis set and the EFMO$_{\text{nt}+\text{pct}}$ procedure. Figure 8 shows the improvement in energy as a function of the number of steps taken in a geometry optimization. The obtained optimized structures have the lowest energies when comparing to all the taken steps, even for one residue per fragment. Compared to RHF (MP2) optimized structures, the rms between the optimized structures are 0.40 (0.44) angstrom (EFMO with one residue per fragment did slightly worse). This can be compared to the 0.3 angstrom that was obtained for FMO2-RHF with HOP previously [41].

EFMO offers a gradient whose quality is similar to FMO2/AFO calculations but at a reduced cost. The quality of the FMO2/AFO gradient could be improved if fully analytic derivatives available such as what was done by Nagata *et. al.* for HOP [51–53]. Another improvement can be obtained with an addition of the derivatives of the EFP monopoles (and higher order multipoles) as outlined by Xie *et al.* [5] We recommend EFMO$_{\text{nt}+\text{pct}}$ for geometry optimizations of polypeptides.

Molecular Clusters

Inclusion of correlation in EFMO (equation 4) warrants a new benchmark of the water clusters that was used in the original EFMO paper. In Table 3, results for MP2 energies are shown for $R_{\text{resdim}} = R_{\text{corsd}} = 2.0$ for various basis sets. Since there are no covalent bonds, the screening parameter was given its original value of $\alpha = 0.6$. In the original EFMO paper, the errors in energy for water clusters were discussed per hydrogen bond (HB) due to EFMO only describing higher order many-body effects for polarization (see ref [14] for full details), thus, the error is a lack of many-body terms per HB. For EFMO-MP2, only monomer and *ab initio* dimers are considered correlated and the lack of treatment separated dimers gives rise to new errors but we expect

Table 2. Errors in gradient of EFMO and FMO2/AFO for the $\alpha-$(ALA)$_{10}$ polypeptide using RHF and MP2.

	FMO2	[a]FMO2	EFMO$_{\text{org}}$	EFMO$_{\text{nt}}$	EFMO$_{\text{nt}+\text{pct}}$	EFMO$_{\text{nt}+\text{adj}}$	EFMO$_{\text{nat}+\text{pct}+\text{adj}}$
RHF							
∇E^{rms}	0.51	0.76	0.73	0.68	0.66	0.66	0.66
∇E^{max}	3.43	4.71	3.50	3.30	3.27	3.73	3.74
MP2							
∇E^{rms}	0.70	0.75	0.69	1.20	0.61		
∇E^{max}	3.57	4.53	2.84	2.91	2.89		

[a]No ESP.
Both RHF/6-31G(d) and MP2/6-31G(d) levels of theory are evaluated. All units in 10^{-3} Hartree/Bohr. The subscripts are: nt for not including torque contributions, pct is a percentage based distribution of the gradient arising from gradient contributins not located on atoms and adj ignores induced dipoles due to neighboring fragments. See text for details.
doi:10.1371/journal.pone.0041117.t002
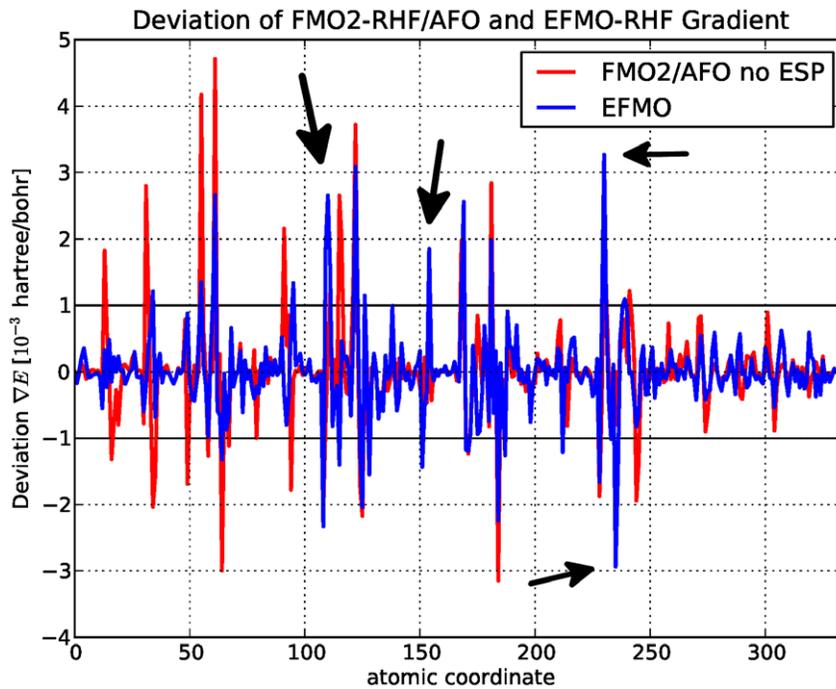
Figure 6. Deviations of analytic gradient from the numeric gradient for RHF on α-(ALA)$_{10}$. Shown in units of 10^{-3} Hartree/Bohr for FMO2-RHF/AFO and EFMO-RHF versus atomic coordinate for the 6-31G(d) basis set.
doi:10.1371/journal.pone.0041117.g006

these to be small. EFP does include dispersion terms [54], but these are not included in this work.

The EFMO-MP2/6-31G(d) results deviate by a maximum of 0.78 kcal/mol per HB, which is worse than FMO2-MP2/6-31G(d) which deviates by a maximum of -0.43 kcal/mol per HB.

Increasing the basis set shows that the EFMO errors are 0.02 and -0.05 kcal/mol per HB for 6-31+G(d) and 6-31++G(d), respectively. For FMO2, the respective errors are -0.76 and -0.48 kcal/mol. The errors we observe for the larger clusters
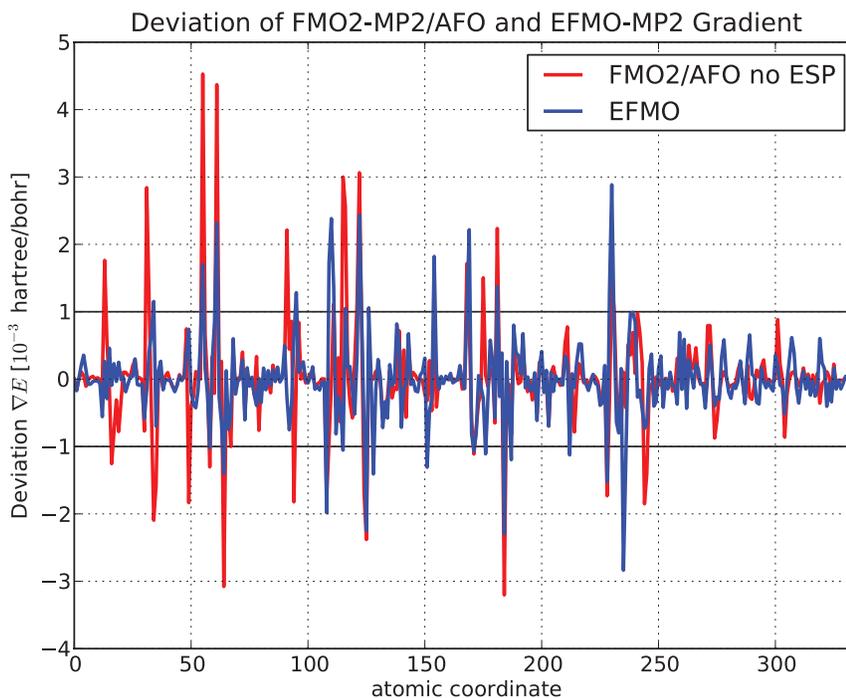
Figure 7. Deviations of analytic gradient from the numeric gradient for MP2 on α-(ALA)$_{10}$. Shown in units of 10^{-3} Hartree/Bohr for FMO2-MP2/AFO and EFMO-MP2 versus atomic coordinate for the 6-31G(d) basis set.
doi:10.1371/journal.pone.0041117.g007

Geometry Optimization of α-(ALA)$_{10}$
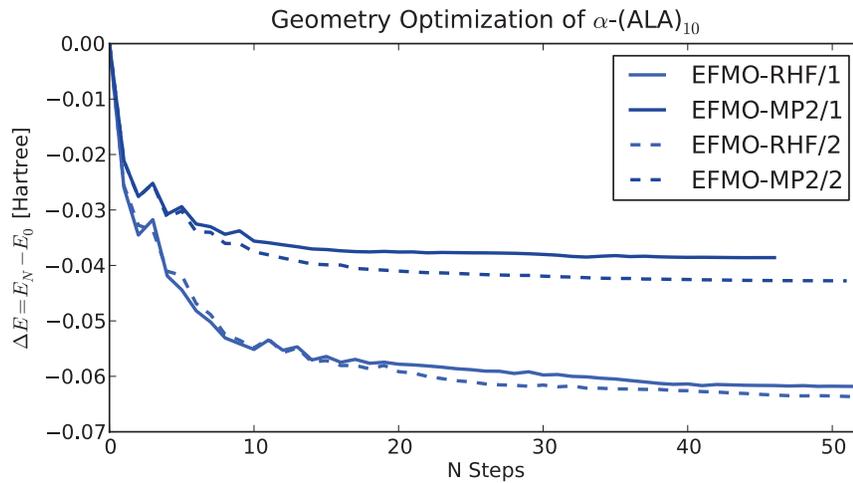
Figure 8. Convergence of energy as a function of number of geometry steps taken. Results are from an optimization of α-(ALA)$_{10}$ EFMO-RHF and EFMO-MP2 with both one and two residues per fragment calculated using the 6-31G(d) basis set. In all cases, the optimized geometries were optimized to a gradient threshold of $5.0 \cdot 10^{-4}$ Hartree/Bohr and all final structures had the lowest energies of all steps taken.
doi:10.1371/journal.pone.0041117.g008

containing 30, 40 and 50 water molecules are consistent with the smaller 20 water molecule cluster.

Timings

In our previous study [14], EFMO-RHF for molecular clusters were two (five) times faster than the corresponding FMO2 energy (gradient) calculation. In Table 4, results for Chignolin and the Trp-cage are presented for 5 nodes using 2 cores per node. All timings were carried out on Intel Xeon X5550 CPUs. Here, using EFMO-MP2 instead of EFMO-RHF increases the computation time by roughly a factor of two (from 14.0 minutes to 29.5 minutes for Chignolin using $R_{resdim} = 2.0$). For FMO2, the same calculation takes 38.5 minutes and 58.6 minutes, respectively. An EFMO-RHF gradient evaluation for Chignolin takes only three minutes longer than the energy, but becomes a five-fold increase when running EFMO-MP2 gradients. The same trends are observed for the Trp-cage. We note a significant speedup when lowering the cutoff distances R_{resdim} and R_{corsd}, especially for the larger Trp-cage. When the cut-off distances go down, the number of *ab initio* dimers decrease. Especially MP2 gradients require much CPU time due to the number of integrals that needs to be transformed [40].

We note that lowering of the cutoff distances R_{resdim} and R_{corsd} can have significant impact on the accuracy [18,32] like we observed for molecular clusters [14], however for a modest lowering of the thresholds to $R_{resdim} = R_{corsd} = 1.5$, the energy deviations from *ab initio* are not affected greatly (Table S3).

Table 3. Water cluster energy error for EFMO and FMO2 relative to ab initio MP2 (in kcal/mol per hydrogen bond).

N_{H_2O}	N_{HB}	EFMO	FMO2
6-31G(d)			
	31	0.63	−0.43
20	32	0.66	−0.37
	29	0.78	−0.38
6-31+G(d)			
	31	0.02	−0.69
20	32	0.01	−0.67
	29	0.02	−0.76
6-31++G(d)			
	31	−0.05	−0.44
20	32	−0.04	−0.43
	29	−0.05	−0.48
6-31G(d)			
30	51	0.59	−0.43
40	63	0.79	−0.41
50	86	0.74	−0.45

Energies are calculated using the 6-31G(d), 6-31+G(d) and 6-31++G(d) basis sets.
In all calculations $R_{resdim} = R_{corsd} = 2.0$ and $\alpha = 0.6$.
doi:10.1371/journal.pone.0041117.t003

Table 4. Timings for FMO2 and EFMO energy and gradient calculations on the Trp-cage protein.

	R_{resdim}, R_{corsd}	T(RHF)	$T(\nabla$RHF)	T(MP2)	$T(\nabla$MP2)
Chignolin					
EFMO	1.0	9.6	11.1	22.8	102.8
	1.5	13.2	13.1	28.7	106.4
	2.0	14.0	17.0	29.5	119.0
FMO2	2.0	38.5	59.7	58.6	149.9[c]
Trp-cage					
EFMO	1.0	24.2[b]	23.5	42.7	161.0
	1.5	33.7	38.3	70.7	261.6
	2.0	37.6	43.0	78.9	314.0
FMO2	2.0	100.4	187.0	142.5	408.6[d]

[a] tested for both RHF/6-31G(d) and MP2/6-31G(d). All units in minutes.
CPU utilization was [b]96%, [c]85% and [d]91%. All other were 99%.
All timings were carried out on 5 nodes containing Intel Xeon X5550 CPUs (10 CPU cores total).
doi:10.1371/journal.pone.0041117.t004

Summary

The effective fragment molecular orbital (EFMO) method is a merger of the effective fragment potential (EFP) method and the fragment molecular orbital (FMO) method and combines the general applicability of the FMO method (for example, to flexible biomolecules) with the speed of the EFP method. In this work, we have introduced new methodology needed to make EFMO work for systems with covalent bonds such as proteins. This, together with the analytical gradient provides an agile tool to treat proteins at a reasonable level of theory. We also showed how to incorporate electron correlation via Mø ller-Plesset perturbation theory.

We made an extensive study on small polypeptides to assess the need for screening when dealing with covalent bonds and found that an additional screening is needed compared to regular EFP. We showed that the deviations in energy on proteins are on par with FMO2 to within a few kcal/mol when using two residues per fragment. For example, Chignolin is reproduced to within 0.1 kcal/mol compared to FMO2. Timings were consistent with our previous work. We obtained two to five times speedup when using EFMO over FMO2 for RHF. The speedup was somewhat lower when employing MP2 gradients, resulting in speedups between 1.6 and 2.3.

There are many ways in which the EFMO method can be improved and extended, for example, interfacing EFMO with the polarized continuum model (PCM) or the classical dispersion interaction in EFP [54] which would enable us to lower R_{corsd} compared to R_{resdim}, thus speeding up the evaluation of the gradient greatly. Another direction is to follow the multilayer FMO method [55] and the recent frozen domain FMO (FMO/FD) method [56].

FMO has been applied [57–59] to a number of chemical problems, [60] and we expect that EFMO can be a useful method on its own, for example, in the structure optimization of protein-ligand complexes and other studies related to drug design.

Supporting Information

Figure S1 Deviations in energy from RHF and MP2 calculations of FMO2/HOP, FMO2/AFO and EFMO for the peptide P1 using two residues per fragment for different values of the screening parameter α.
(EPS)

Figure S2 Deviations in energy from RHF and MP2 calculations of FMO2/HOP, FMO2/AFO and EFMO for the peptide P2 using two residues per fragment for different values of the screening parameter α. Large positive values ($\gg 200$ kcal/mol) indicates that EFMO did not converge. See main text for full details.
(EPS)

Figure S3 Deviations in energy from RHF and MP2 calculations of FMO2/HOP, FMO2/AFO and EFMO for the peptide P3 using two residues per fragment for different values of the screening parameter α. Large positive values ($\gg 200$ kcal/mol) indicates that EFMO did not converge. See main text for full details.
(EPS)

Figure S4 Correlation between the deviation in energy of peptide P1 using two residues per fragment and the radius of gyration. Lower values of the radius of gyration is a more compact protein.
(EPS)

Figure S5 Correlation between the deviation in energy of peptide P2 using two residues per fragment and the radius of gyration. Lower values of the radius of gyration is a more compact protein.
(EPS)

Figure S6 Correlation between the deviation in energy of peptide P3 using two residues per fragment and the radius of gyration. Lower values of the radius of gyration is a more compact protein.
(EPS)

Figure S7 Two Carbon atoms (C_1 and C_2) and an the location of an induced dipole μ^{ind} above the bond midpoint (Drawn cartoonishly to emphasize the methodology).
(EPS)

Table S1 Energy Error of EFMO and FMO2/AFO compared to *ab initio* calculations on proteins and protein-like structures for different values of $R_{resdim} = R_{cor}$ using one residue per fragment. In all calculations, the screening parameter α was kept fixed at a value of $\alpha = 0.1$.
(TEX)

Table S2 Calculated mean average deviation ΔE_{MAD} and average deviation ΔE_{avg} for conformers of the peptides P1, P2 and P3 using two residues per fragment, the 6-31G(d) basis set and different values of the screening parameter α. For reference, FMO2/HOP and FMO2/AFO was included. All units in kcal/mol.
(TEX)

Table S3 Energy Error of EFMO compared to *ab initio* calculations on proteins and protein-like structures for different values of $R_{resdim} = R_{cor}$ using two residue per fragment. In all calculations, the screening parameter α was kept fixed at a value of $\alpha = 0.1$.
(TEX)

Text S1 Detailed description of the percentage based distribution of the gradient between two nearby atoms.
(TEX)

Author Contributions

Conceived and designed the experiments: CS. Performed the experiments: CS. Analyzed the data: CS DF JJ. Contributed reagents/materials/analysis tools: CS. Wrote the paper: CS DF JJ.

References

1. Zhang, DW; Zhang, JZH (2003) Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy. J Chem Phys 119: 3599.

2. Zhang, DW; Chen, XH; Zhang, JZH (2003) Molecular caps for full quantum mechanical computation of peptide-water interaction energy. J Comput Chem 24: 1846–1852.

3. Jiang N; Ma, J; Jiang, Y (2006) Electrostatic field-adapted molecular fractionation with conjugated caps for energy calculations of charged biomolecules. J Chem Phys 124: 114112.

4. Dahlke EE, Truhlar DG (2007) Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. J Chem Theo Comput 3: 46–53.

5. Xie W, Song L, Truhlar D, Gao J (2008) The variational explicit polarization potential and analytical first derivative of energy: Towards a next generation force field. J Chem Phys 128: 234108.

6. Xie W, Gao J (2007) Design of a next generation force field: the x-pol potential. J Chem Theory Comput 3: 1890–1900.

7. Söderhjelm P, Ryde U (2009) How Accurate Can a Force Field Become? A Polarizable Multipole Model Combined with Fragment-wise Quantum-Mechanical Calculations. J Phys Chem A 113: 617–627.

8. Söderhjelm P Aquilante F Ryde U (2009) Calculation of Protein-Ligand Interaction Energies by a Fragmentation Approach Combining High-Level Quantum Chemistry with Classical Many-Body Effects. J Phys Chem B: 11085–11094.

9. Beran GJO (2009) Approximating quantum many-body intermolecular interactions in molecular clusters using classical polarizable force fields. J Chem Phys 130: 164115.

10. Sebetci A, Beran G (2009) Spatially homogeneous qm/mm for systems of interacting molecules with on-the-fly ab initio force-field parametrization. Journal of Chemical Theory and Computation 6: 155–167.

11. Jacobson L, Herbert J (2011) An efficient, fragment-based electronic structure method for molecular systems: Self-consistent polarization with perturbative two-body exchange and dispersion. J Chem Phys 134: 094118.

12. Gordon MS, Mullin JM, Pruitt SR, Roskop LB, Slipchenko LV, et al. (2009) Accurate Methods for Large Molecular Systems. J Phys Chem B 113: 9646–9663.

13. Gordon MS, Fedorov DG, Pruitt SR, Slipchenko LV (2012) Fragmentation Methods: A Route to Accurate Calculations on Large Systems. Chem Rev 112: 632–672.

14. Steinmann C, Fedorov D, Jensen J (2010) Effective fragment molecular orbital method: A merger of the effective fragment potential and fragment molecular orbital methods. J Phys Chem A 114: 8705–8712.

15. Kitaura K, Ikeo E, Asada T, Nakano T, Uebayasi M (1999) Fragment molecular orbital method: an approximate computational method for large molecules. Chem Phys Lett 313: 701–706.

16. Nakano T, Kaminuma T, Sato T, Akiyama Y, Uebayasi M, et al. (2000) Fragment molecular orbital method: application to polypeptides. Chem Phys Lett 318: 614–618.

17. Nakano T, Kaminuma T, Sato T, Fukuzawa K, Akiyama Y, et al. (2002) Fragment molecular orbital method: use of approximate electrostatic potential. Chem Phys Lett 351: 475–480.

18. Fedorov DG, Kitaura K (2004) The importance of three-body terms in the fragment molecular orbital method. J Chem Phys 120: 6832–6840.

19. Fedorov DG, Kitaura K (2007) Extending the power of quantum chemistry to large systems with the fragment molecular orbital method. J Phys Chem A 111: 6904–6914.

20. Nagata T, Fedorov DG, Kitaura K (2011) Mathematical formulation of the fragment molecular orbital method. In: Zalesny R, Papadopoulos MGG, Mezey PGG, Leszczynski J, editors, Linear- Scaling Techniques in Computational Chemistry and Physics, Springer Netherlands, volume 13 of Challenges and Advances in Computational Chemistry and Physics. 17–64.

21. Day PN, Jensen JH, Gordon MS, Webb SP, Stevens WJ, et al. (1996) An effective fragment method for modeling solvent effects in quantum mechanical calculations. J Chem Phys 105: 1968–1986.

22. Gordon MS, Freitag MA, Bandyopadhyay P, Jensen JH, Kairys V, et al. (2001) The effective fragment potential method: A QM-based MM approach to modeling environmental effects in chemistry. J Phys Chem A 105: 293–307.

23. Ghosh D, Kosenkov D, Vanovschi V, Williams C, Herbert J, et al. (2010) Noncovalent interactions in extended systems described by the effective fragment potential method: Theory and application to nucleobase oligomers. J Phys Chem A 114: 12739–12754.

24. Nagata T, Fedorov DG, Kitaura K, Gordon MS (2009) A combined effective fragment potential– fragment molecular orbital method. I. The energy expression; initial applications. J Chem Phys 131: 024101.

25. Nagata T, Fedorov DG, Sawada T, Kitaura K, Gordon MS (2011) A combined effective fragment potential - fragment molecular orbital method. II. Analytic gradient and application to the geometry optimization of solvated tetraglycine and chignolin. J Chem Phys 134: 034110.

26. Stone AJ (1981) Distributed multipole analysis, or how to describe a molecular charge distribution. Chem Phys Lett 83: 233–239.

27. Minikis RM, Kairys V, Jensen JH (2001) Accurate intraprotein electrostatics derived from first principles: an effective fragment potential method study of the proton affinities of lysine 55; tyrosine 20 in turkey ovomucoid third domain. J Phys Chem A 105: 3829–3837.

28. Fedorov D, Jensen J, Deka R, Kitaura K (2008) Covalent bond fragmentation suitable to describe solids in the fragment molecular orbital method. J Phys Chem A 112: 11808–11816.

29. Fedorov D, Avramov P, Jensen J, Kitaura K (2009) Analytic gradient for the adaptive frozen orbital bond detachment in the fragment molecular orbital method. Chem Phys Lett 477: 169–175.

30. Kairys V, Jensen J (2000) QM/MM boundaries across covalent bonds: a frozen localized molecular orbital-based approach for the effective fragment potential method. J Phys Chem A 104: 6656–6665.

31. Edmiston C, Ruedenberg K (1963) Localized atomic and molecular orbitals. Rev Mod Phys 35: 457.

32. Fedorov D, Kitaura K (2004) Second order Møller-Plesset perturbation theory based upon the fragment molecular orbital method. J Chem Phys 121: 2483.

33. Fedorov D, Kitaura K (2005) Coupled-cluster theory based upon the fragment molecular-orbital method. J Chem Phys 123: 134103.

34. Fedorov D, Ishimura K, Ishida T, Kitaura K, Pulay P, et al. (2007) Accuracy of the three-body fragment molecular orbital method applied to Møller–Plesset perturbation theory. J Comput Chem 28: 1476–1484.

35. Schmidt MW, Baldridge KK, Boatz JA, Elbert ST, Gordon MS, et al. (1993) General atomic; molecular electronic structure system. J Comput Chem 14: 1347–1363.

36. Fedorov D, Olson R, Kitaura K, Gordon M, Koseki S (2004) A new hierarchical parallelization scheme: Generalized distributed data interface (gddi), and an application to the fragment molecular orbital method (fmo). J Comput Chem 25: 872–880.

37. Hariharan PC, Pople JA (1973) The influence of polarization functions on molecular orbital hydrogenation energies. Theor Chem Acc 28: 213–222.

38. Francl MM, Pietro WJ, Hehre WJ, Binkley JS, Gordon MS, et al. (1982) Self-consistent molecular orbital methods. xxiii. a polarization-type basis set for second-row elements. J Chem Phys 77: 3654–3665.

39. Gordon MS, Binkley JS, Pople JA, Pietro WJ, Hehre WJ (1982) Self-consistent molecularorbital methods. 22. Small split-valence basis sets for second-row elements. J Am Chem Soc 104: 2797–2803.

40. Ishimura K, Pulay P, Nagase S (2006) A new parallel algorithm of mp2 energy calculations. J Comput Chem 27: 407–413.

41. Fedorov D, Ishida T, Ubayasi M, Kitaura K (2007) The fragment molecular orbital method for geometry optimizations of polypeptides and proteins. J Phys Chem A 111: 2722–2732.

42. Dolinsky T, Nielsen J, McCammon J, Baker N (2004) Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. Nucleic Acids Res 32: W665.

43. Dolinsky T, Czodrowski P, Li H, Nielsen J, Jensen J, et al. (2007) Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res 35: W522.

44. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, et al. (2011) Open babel: An open chemical toolbox. J Cheminf 3: 33.

45. (2011). The open babel package, version 2.3.0. http://openbabel.org.

46. Stewart JJP (2007) Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. J Mol Model 13: 1173–1213.

47. Stewart JJP (2008) Mopac2009, stewart computational chemistry, colorado springs, co, usa. Web: "http://OpenMOPACnet".

48. Tomasi J, Mennucci B, Cammi R (2005) Quantum mechanical continuum solvation models. Chem Rev 105: 2999–3094.

49. Molina P, Li H, Jensen J (2003) Intraprotein electrostatics derived from first principles: Divideand-conquer approaches for qm/mm calculations. J Comput Chem 24: 1971–1979.

50. Jensen J, Li H, Robertson A, Molina P (2005) Prediction and rationalization of protein p k a values using qm and qm/mm methods. J Phys Chem A 109: 6634–6643.

51. Nagata T, Fedorov D, Kitaura K (2010) Importance of the hybrid orbital operator derivative term for the energy gradient in the fragment molecular orbital method. Chem Phys Lett 492: 302–308.

52. Nagata T, Brorsen K, Fedorov D, Kitaura K, Gordon M (2011) Fully analytic energy gradient in the fragment molecular orbital method. J Chem Phys 134: 124115.

53. Nagata T, Fedorov D, Ishimura K, Kitaura K (2011) Analytic energy gradient for second-order Møller-Plesset perturbation theory based on the fragment molecular orbital method. J Chem Phys 135: 044110.

54. Adamovic I, Gordon MS (2005) Dynamic polarizability, dispersion coefficient C 6; dispersion energy in the effective fragment potential method. Mol Phys 2: 379–387.

55. Fedorov DG, Ishida T, Kitaura K (2005) Multilayer formulation of the fragment molecular orbital method (FMO). J Phys Chem A 109: 2638–2646.

56. Fedorov D, Alexeev Y, Kitaura K (2011) Geometry optimization of the active site of a large system with the fragment molecular orbital method. J Phys Chem Lett 2: 282–288.

57. Sawada T, Fedorov DG, Kitaura K (2010) Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations. J Am Chem Soc 132: 16862–16872.

58. Ohno K, Mori K, Orita M, Takeuchi M (2011) Computational insights into binding of bisphosphates to farnesyl pyrophosphate synthase. Curr Med Chem 18: 220–233.

59. Mazanetz MP, Ichihara O, Law RJ, Whittaker M (2011) Prediction of cyclin-dependent kinase 2 inhibitor potency using the fragment molecular orbital method. J Cheminf 3: 2.

60. Fedorov DG, Nagata T, Kitaura K (2012) Exploring chemistry with the fragment molecular orbital method. Phys Chem Chem Phys 14: 7562–7577.