The effective fragment molecular orbital method for fragments connected by covalent bonds
Svendsen, Casper Steinmann; Fedorov, Dmitri G.; Jensen, Jan Halborg

Published in:
P L o S One

DOI:
10.1371/journal.pone.0041117

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
The Effective Fragment Molecular Orbital Method for Fragments Connected by Covalent Bonds

Casper Steinmann1,*, Dmitri G. Fedorov2, Jan H. Jensen1

1 Department of Chemistry, University of Copenhagen, Copenhagen, Denmark, 2NIRI, 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

\[ E_{\text{FMO2}} = \sum_1^N E_I^0 + \sum_{IJ} (E_{IJ}^0 - E_I^0 - E_J^0) \]  

(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

\[ E_{\text{EFMO}} = \sum_1^N E_I^0 + \sum_{IJ} (E_{IJ}^0 - E_I^0 - E_J^0 - E_{IJ}^{\text{POL}}) \]  

(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 \( R_{\text{resdim}} \).
$E_{IJ}^{\text{POL}}$ 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 $E_{ES}^{IJ}$ is the classical electrostatic interaction energy and applies to dimers separated by a distance greater than $R_{\text{resdim}}$. 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.

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 [20] 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.

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.
\[
    k(\mathbf{R}, \alpha, \beta) = 1 - \exp\left( -\alpha \beta |\mathbf{R}|^2 \right) \left( 1 + \sqrt{\alpha \beta |\mathbf{R}|^2} \right)
\]

Here, \( \alpha \) and \( \beta \) are the screening parameters associated with fragments \( I \) and \( J \), respectively. The distance parameter \( |\mathbf{R}| \) 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,

\[
    E = E_{\text{EFMO}} + E_{\text{COR}}.
\]

Here \( E_{\text{COR}} \) is given as the sum of monomer correlation energies \( E_{\text{COR}}^I \) and pairwise corrections,

\[
    E_{\text{COR}} = \sum_{I} E_{\text{COR}}^I + \sum_{IJ} (E_{\text{COR}}^{IJ} - E_{\text{COR}}^I - E_{\text{COR}}^J),
\]

where \( E_{\text{COR}}^{IJ} \) is the correlation energy of dimer \( IJ \). The distance parameter \( R_{\text{cord}} \) 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 \textit{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 parallelized with the generalized distributed data interface [36]. EFMO was a locally modified version of GAMESS [35]. EFMO was parallelized with the generalized distributed data interface [36].

The \textit{ab initio} MP2 calculations had their integral accuracy increased to \( 10^{-12} \) (ICUT = 12 in $\text{SCTRL}$), SCF convergence criterion was raised from \( 10^{-5} \) to \( 10^{-7} \) (CONV = 1E-7 in $\text{SCF}$) and the MP2 code by K. Ishimura et al [40] with AO integral transformation threshold increased from \( 10^{-9} \) to \( 10^{-12} \) (CO-DE = IMS and CUTOFF = 1E-12 in $\text{MP2}$) 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 = - RUEDNBFR in $\text{SCTRL}$ and RAFO(1) = 1,1,1 in $\text{FMO}$). The parameters for the electrostatic treatment of dimers \( R_{\text{resdim}} \) and the threshold for the inclusion of correlation effects \( R_{\text{cord}} \) were both set to 2.0 \( (\text{RESDIM} = 2.0, \text{RCORSD} = 2.0 \) in $\text{FMO}$) 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 is set to 0.1 for fragments with and without the SBO (SCREEN(1) = 0.1,0.1 in $\text{FMO}$), 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 \( \alpha \)-helices \( (\alpha-(\text{ALA})_n) \) and \( \beta \)-sheets \( (\beta-(\text{ALA})_n) \) 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 \( \text{pH} = 7 \)) amino acids AIVGLT (P1) and AVSNTL (P2) as well as four neutral and two non-neutral (at \( \text{pH} = 7 \)) 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 \( \text{pH} = 7 \)) using the PDB2PQR tool. P1 had neutral termini (arguments –neutral/–neutral) while P2 and P3 both had charged termini. For each polypeptide, a conformational search was carried out to locate twenty different structures using the OboConformer 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:

\[
    \Delta E_M^X = E_M - E_X, 
\]

the average deviation of conformers

\[
    \Delta E_{\text{avg}}^M = \frac{1}{N} \sum_{I} \{ E_M^I - E_X^I \},
\]

and the mean average deviation (MAD) for conformers

\[
    \Delta E_{\text{MAD}}^M = \frac{1}{N} \sum_{I} |E_M^I - E_X^I|.
\]

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 (\( \nabla E_{\text{num}} \)) were calculated on \( \alpha-(\text{ALA})_{10} \) and compared to its analytical counterpart (\( \nabla E_{\text{ana}} \)) by the root mean square (rms) deviation of the individual elements

\[
    \nabla E_{\text{rms}} = \sqrt{\frac{\sum_{i} (\nabla E_{\text{ana}} - \nabla E_{\text{num}})^2}{N_4}},
\]

and the maximum deviation

\[
    \nabla E_{\text{max}} = \max(|\nabla_i E_{\text{ana}} - \nabla_i E_{\text{num}}|).
\]

As discussed in Section 2, the computational methodology section below.
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 $R_{\text{resdim}}$ and $R_{\text{corsd}}$ 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 $2.87 (-2.21)$ 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.

\[
R^2_g = \frac{1}{N_g} \sum_{k=1}^{N_g} (g_k - \bar{g}_{\text{mean}})^2. \tag{11}
\]
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 $2.94 \pm 0.32$ kcal/mol for $\alpha-(\text{ALA})_{10}$ to $0.18 \pm 18.94$ kcal/mol for the large $\alpha-(\text{ALA})_{40}$ helix, which corresponds to an average error per residue of $0.29 \pm 0.03$ kcal/mol for $\alpha-(\text{ALA})_{10}$ and less than $0.01 \pm 0.47$ kcal/mol for $\alpha-(\text{ALA})_{40}$. The α-helices tend to illustrate the case of over-polarization. For $\alpha-(\text{ALA})_{10}$, the total polarization energy is small ($\sim 12.89$ kcal/mol) but as the system size increase, so does the total polarization energy ($\sim 73.81$ 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 $0.60 \pm 0.89$ kcal/mol to $4.05 \pm 6.46$ kcal/mol for $\beta-(\text{ALA})_{10}$ and $\beta-(\text{ALA})_{40}$, respectively. Overall, the average error per residue becomes $0.06 \pm 0.50$ kcal/mol and $0.10 \pm 0.16$ 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 ($\beta-(\text{ALA})_{40}$ has a polarization energy of around 50 kcal/mol).

### 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 $\Delta \epsilon_{\text{rms}}^\text{grad} = 0.2 \times 10^{-3}$ Hartree/Bohr and $\Delta \epsilon_{\text{max}}^\text{grad} = 1.4 \times 10^{-2}$ 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 · 10⁻³ Hartree/Bohr, 3.43 · 10⁻³ Hartree/Bohr and 0.76 · 10⁻³ Hartree/Bohr and 4.71 · 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 \( \vec{p}_{ind} \) to the nearest atom only, in this study named EFMO\(_{arg}\). This is a clear improvement over the gradient terms of the induced dipoles (without the ESP) result \( \langle V E_{\text{rms}} \rangle: 0.73 · 10^{-3} \) Hartree/Bohr, \( \langle V E_{\text{max}} \rangle: 3.50 · 10^{-3} \) Hartree/Bohr), but some deviations in gradient get worse using EFMO and will be discussed further below. Removing all torque contributions (EFMO\(_{nt}\)) reveals further improvements \( \langle V E_{\text{rms}} \rangle: 0.68 · 10^{-3} \) Hartree/Bohr, \( \langle V E_{\text{max}} \rangle: 3.30 · 10^{-3} \) Hartree/Bohr). Another approach, specifically for the induced dipole (EFMO\(_{nt+pct}\)) 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 \( \langle V E_{\text{rms}} \rangle: 0.66 · 10^{-3} \) Hartree/Bohr, \( \langle V E_{\text{max}} \rangle: 3.27 · 10^{-3} \) 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\(_{nt+pct+adj}\)) and without (EFMO\(_{nt+adj}\)) percentage based distribution of induced dipoles are \( \langle V E_{\text{rms}} \rangle: 0.66 · 10^{-3} \) Hartree/Bohr, \( \langle V E_{\text{max}} \rangle: 3.37 · 10^{-3} \) Hartree/Bohr and \( \langle V E_{\text{rms}} \rangle: 0.66 · 10^{-3} \) Hartree/Bohr, \( \langle V E_{\text{max}} \rangle: 3.74 · 10^{-3} \) Hartree/Bohr) offer no clear advantage over EFMO\(_{nt+pct}\) 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\(_2\)). 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\(_{nt+pct}\) methodology \( \langle V E_{\text{rms}} \rangle: 0.61 · 10^{-3} \) Hartree/Bohr, \( \langle V E_{\text{max}} \rangle: 2.89 · 10^{-3} \) Hartree/Bohr) while FMO2-MP2/AFO errors are very similar to the corresponding RHF values.

Finally, geometry optimizations were carried out for \( \alpha\)-[ALA]\(_{10}\) using the 6-31G(d) basis set and the EFMO\(_{nt+pct}\) 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 0.5 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\(_{nt+pct}\) 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 \( R_{\text{redim}} = R_{\text{redim}} = 2.0 \) 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. |
|---|---|---|---|---|---|---|
| **RHF** | **FMO2** | **EP** | **EFMO\(_{arg}\)** | **EFMO\(_{nt}\)** | **EFMO\(_{nt+pct}\)** | **EFMO\(_{nt+adj}\)** |
| \( \langle V E_{\text{rms}} \rangle \) | 0.51 | 0.76 | 0.73 | 0.68 | 0.66 | 0.66 |
| \( \langle V E_{\text{max}} \rangle \) | 3.43 | 4.71 | 3.50 | 3.30 | 3.27 | 3.73 |
| **MP2** | **FMO2** | **EP** | **EFMO\(_{arg}\)** | **EFMO\(_{nt}\)** | **EFMO\(_{nt+pct}\)** | **EFMO\(_{nt+adj}\)** |
| \( \langle V E_{\text{rms}} \rangle \) | 0.70 | 0.75 | 0.69 | 1.20 | 0.61 |
| \( \langle V E_{\text{max}} \rangle \) | 3.57 | 4.53 | 2.84 | 2.91 | 2.89 |

*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 contributions not located on atoms and adj ignores induced dipoles due to neighboring fragments. See text for details.

doi:10.1371/journal.pone.0041117.t002
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 −0.40 kcal/mol. The errors we observe for the larger clusters...
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 $R_{\text{resdim}} = 2.0$). 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 $R_{\text{resdim}}$ and $R_{\text{corsd}}$, 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 $R_{\text{resdim}}$ and $R_{\text{corsd}}$ 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 $R_{\text{resdim}} = R_{\text{corsd}} = 1.5$, 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).

<table>
<thead>
<tr>
<th>$N_{\text{H2O}}$</th>
<th>$N_{\text{HB}}$</th>
<th>EFMO</th>
<th>FMO2</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>0.63</td>
<td>−0.43</td>
<td></td>
</tr>
<tr>
<td>6-31G(d)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.66</td>
<td>−0.37</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.78</td>
<td>−0.38</td>
<td></td>
</tr>
<tr>
<td>6-31+G(d)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.02</td>
<td>−0.69</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.01</td>
<td>−0.67</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.02</td>
<td>−0.76</td>
<td></td>
</tr>
<tr>
<td>6-31++G(d)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>−0.05</td>
<td>−0.44</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>−0.04</td>
<td>−0.43</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>−0.05</td>
<td>−0.48</td>
<td></td>
</tr>
<tr>
<td>6-31G(d)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.59</td>
<td>−0.43</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.79</td>
<td>−0.41</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.74</td>
<td>−0.45</td>
<td></td>
</tr>
</tbody>
</table>

Energies are calculated using the 6-31G(d), 6-31+G(d) and 6-31++G(d) basis sets. In all calculations $R_{\text{resdim}} = R_{\text{corsd}} = 2.0$ and $\varepsilon = 0.6$.

doi:10.1371/journal.pone.0041117.t003

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

<table>
<thead>
<tr>
<th>$R_{\text{resdim}}$, $R_{\text{corsd}}$</th>
<th>$\tau$(RHF)</th>
<th>$\tau$(MP2)</th>
<th>$\tau$(MP2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chignolin</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFMO 1.0</td>
<td>9.6</td>
<td>11.1</td>
<td>22.8</td>
</tr>
<tr>
<td>1.5</td>
<td>13.2</td>
<td>13.1</td>
<td>28.7</td>
</tr>
<tr>
<td>2.0</td>
<td>14.0</td>
<td>17.0</td>
<td>29.5</td>
</tr>
<tr>
<td>FMO2 2.0</td>
<td>38.5</td>
<td>59.7</td>
<td>58.6</td>
</tr>
</tbody>
</table>

| Trp-cage                            |             |             |             |
| EFMO 1.0                            | 24.3$^a$    | 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 Möller-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 $R_{\text{cor}}$ compared to $R_{\text{resdim}}$, 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].

EFMO 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 $\alpha$.

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 $\alpha$. Large positive values ($>200$ kcal/mol) indicates that EFMO did not converge. See main text for full details.

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 $\alpha$. Large positive values ($>200$ kcal/mol) indicates that EFMO did not converge. See main text for full details.

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.

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.

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.

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

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

Table S2 Calculated mean average deviation $\Delta E_{\text{MAD}}$ and average deviation $\Delta E_{\text{avg}}$ 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 $\alpha$. For reference, FMO2/HOP and FMO2/AFO was included. All units in kcal/mol.

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

Acknowledgments
C.S. would like to thank Assistant Professor Hui Li of University of Lincoln-Nebraska and Assistant Professor Lyudmila Slipchenko of University of Purdue for interesting and fruitful discussions. D.G.F thanks Prof. Kazuo Kitaura for many fruitful discussions. All authors would like to thank Professor Mark S. Gordon for many fruitful discussions, and for suggesting the idea to use the EFP polarization in the FMO framework.

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