Fragment-Based Calculations of Enzymatic Thermochemistry Require Dielectric Boundary Conditions

Paige E. Bowling, Dustin R. Broderick, and John M. Herbert*

Cite This: J. Phys. Chem. Lett. 2023, 14, 3826–3834

ABSTRACT: Electronic structure calculations on enzymes require hundreds of atoms to obtain converged results, but fragment-based approximations offer a cost-effective solution. We present calculations on enzyme models containing 500–600 atoms using the many-body expansion, comparing to benchmarks in which the entire enzyme–substrate complex is described at the same level of density functional theory. When the amino acid fragments contain ionic side chains, the many-body expansion oscillates under vacuum boundary conditions but rapid convergence is restored using low-dielectric boundary conditions. This implies that full-system calculations in the gas phase are inappropriate benchmarks for assessing errors in fragment-based approximations. A three-body protocol retains sub-kilocalorie per mole fidelity with respect to a supersystem calculation, as does a two-body calculation combined with a full-system correction at a low-cost level of theory. These protocols pave the way for application of high-level quantum chemistry to large systems via rigorous, ab initio treatment of many-body polarization.

Fragment-based approximations are an attractive way to circumvent nonlinear scaling of computational quantum chemistry (QC), whose floating point cost normally grows like $O(N^p)$ as a function of system size ($N$), with exponents ($p$) ranging from 3 for density functional theory (DFT) to $\geq 7$ for levels of theory that can provide thermochemical benchmarks. Fragmentation into $N_f$ separate subsystems (fragments), each of size $n$, reduces that cost to $N_f \times O(n^p)$ in a manner that is amenable to distributed computing and does not require modification to electronic structure codes. Nonlinear growth in $N_f$ with respect to system size can be mitigated by means of distance- or energy-based thresholds.

The present work presents a protocol for using fragmentation to compute reaction energies and activation barriers for enzyme-catalyzed chemical reactions. Over the past decade, benchmark calculations have revealed that enzymatic thermochemistry does not converge until hundreds of atoms are included in the QC calculation, which is much larger than is typical in contemporary quantum mechanics/molecular mechanics (QM/MM) calculations. Fragmentation may therefore offer an efficient route for obtaining converged thermochemistry (for $N > 500$ atoms) at benchmark levels of theory, provided that errors associated with fragmentation can be controlled. We demonstrate that these errors can be reduced below the "thermochemical accuracy" threshold of 1 kcal/mol. At the same time, our calculations reveal that straightforward comparison of fragment-based approximations to full-system benchmarks (as a means to assess error) is illposed if the calculations are carried out under vacuum boundary conditions.

We will consider models of enzyme–substrate complexes containing approximately 500–600 atoms. Total energies are approximated by means of a many-body expansion (MBE),

$$E = \sum_i E_i + \sum_{i,j} \Delta E_{ij} + \sum_{i,j,k} \Delta E_{ijk} + \cdots$$

(1)

Individual terms are

$$\Delta E_{ij} = E_{ij} - E_i - E_j$$

(2)

for the two-body corrections, where $E_{ij}$ is the energy of the dimer formed from fragments $I$ and $J$, and

$$\Delta E_{ijk} = E_{ijk} - E_{ij} - E_{ik} - E_{jk} - E_i - E_j - E_k$$

(3)

for the three-body corrections. When eq 1 is truncated at $n$-body terms, we denote the resulting approximation as MBE($n$). Electrostatic embedding of the subsystem calculations, using classical point charges derived from the fragment wave functions, is often used to hasten convergence of the MBE. We avoid this, however, having found that charge embedding can lead to inconsistent convergence of the $n$-body expansion. Furthermore, the use of self-consistent point

Received: February 25, 2023
Accepted: April 6, 2023
charges significantly complicates the formulation of analytic energy gradients.\textsuperscript{32−36}

Fragments $I$, $J$, $K$, ... are taken to be individual amino acids of the enzyme (except where stipulated otherwise, for testing purposes), with the substrate as its own fragment. Although larger fragments have sometimes been used for proteins,\textsuperscript{37} we are able to achieve our target accuracy of 1 kcal/mol using single-residue fragments. (The treatment of the substrate is discussed below.) Alternatively, overlapping fragments have sometimes been used for polypeptides and proteins.\textsuperscript{34,35,38−45} This can be motivated in terms of a generalized (G)MBE,\textsuperscript{14,46,47} but most overlapping fragment applications to date have used a one-body approach that captures through-bond interactions but not through-space interactions.\textsuperscript{1} A two-body GMBE can capture both, but is relatively expensive in terms of the number of subsystems that are generated.\textsuperscript{33,39} As such, we stick to the simple MBE($n$) approach in this work.

As a first test, we consider $S_n$, 2 methyl transfer\textsuperscript{48} catalyzed by human catechol $O$-methyltransferase (COMT).\textsuperscript{19−51} This particular enzyme has become something of a benchmark,\textsuperscript{44,52−54} because it has a well-resolved crystal structure,\textsuperscript{51} kinetics data,\textsuperscript{49} and numerous known inhibitors.\textsuperscript{14−16} A Mg$^{2+}$ ion in the active site is essential to COMT's function\textsuperscript{55} but leads to charge-transfer effects in QC calculations that can significantly alter the barrier height, depending on the size of the model system.\textsuperscript{14,52,55} Kulik et al.\textsuperscript{14} considered a sequence of COMT models with QM regions with up to 940 atoms, and we selected “model 8” from ref 14, which contains 632 atoms and 35 fragments. The largest fragment consists of the octahedral coordination sphere around Mg$^{2+}$, including deprotonated catechol (2-hydroxyphenolate (C$_6$H$_4$O$_2^-$)), two aspartic acid residues, an asparagine residue, and a water molecule (58 atoms). Reactant, product, and transition state structures for the transfer of a methyl from SAM to catecholate were protonated and relaxed as described in Computational Details. All calculations were performed at DFT levels of theory, so that we may obtain energies for the full enzyme–substrate complex at the same level of theory and thereby examine convergence of the MBE toward a well-defined target. As such, the errors discussed below are defined with respect to a supersystem calculation at the same level of theory.

The overall charge on this QM model is $-1$, but the system contains nine fragments with non-zero charge. Small anions in the gas phase are sometimes inherently unstable (or metastable), as in the case of SO$_4^{2-}$,\textsuperscript{57,58} and delocalization errors in DFT can exacerbate this problem.\textsuperscript{58} To avoid artifacts, charged residues are often neutralized in fragment-based calculations on proteins.\textsuperscript{59−61} This is not always a viable or realistic option, however, as charged side chains may be directly involved in stabilizing the protein structure or binding to a ligand (as in the present example), or may be vital to a reaction mechanism. A general procedure for enzymatic thermochemistry must admit the possibility of fragments with non-zero charge.

When we naïvely apply MBE($n$) to a large COMT model with charged residues, however, we find that convergence is erratic. This is shown for the barrier height (activation energy $E_a$) in Figure 1a, where MBE(2) overestimates the barrier by 5.4 kcal/mol but MBE(3) underestimates it by 16.7 kcal/mol. To verify that charged residues are the problem, we prepared a second model of COMT in which nearest-neighbor fragments are combined to neutralize charge; e.g., a negatively charged aspartic acid residue is combined with a positively charged ligand, forming a single fragment. This increases the largest fragment size from 58 to 124 atoms but does not alter any protonation states. Using this “charge-coordinated” model of COMT, we observe rapid convergence of MBE($n$), such that two- and three-body calculations afford essentially identical values for both $E_a$ and $\Delta_{\text{rxn}}E$ (see Figure 1). Even one-body calculations perform reasonably well for the charge-coordinated model, due to the larger fragment size, but enlarging the fragments is not an attractive strategy for levels of theory beyond DFT.

Each of the calculations described above was performed using vacuum boundary conditions. As an alternative, we introduce low-dielectric boundary conditions using a polarizable continuum model (PCM).\textsuperscript{62} For protein electrostatics calculations based on the Poisson–Boltzmann equation, it is common to use a dielectric constant $\varepsilon$ in the range of 2−4 to represent the hydrophobic interior of the protein,\textsuperscript{63−68} although larger values have occasionally been suggested.\textsuperscript{68−74} The precise value of $\varepsilon$ may matter for $pK_a$ calculations, but $E_a$ and $\Delta_{\text{rxn}}E$ often converge quickly as a function of $\varepsilon$, such that results for $\varepsilon = 4$ are indistinguishable from much larger values.\textsuperscript{75,76} That is indeed the case for COMT, as shown in Figures S2 and S3. Gas-phase energetics ($\varepsilon = 1$) are an outlier, but $\varepsilon = 4$ is nearly indistinguishable from $\varepsilon = 32$. Subsequent PCM calculations in this work use $\varepsilon = 4$.

In 2023, we extended some of these results to $n = 4$ in order to check convergence. Using PCM boundaries, the difference between

![Figure 1. Errors in MBE($n$) calculations at the oB97X-D/def2-SVP level as compared to a supersystem calculation at the same level of theory: (a) $E_a$ for COMT, (b) $\Delta_{\text{rxn}}E$ for COMT, (c) $E_a$ for AspDC, and (d) $\Delta_{\text{rxn}}E$ for AspDC. For COMT, results are shown for vacuum boundary conditions ($\varepsilon = 1$) vs PCM boundary conditions ($\varepsilon = 4$) and for a charge-coordinated model that uses larger, charge-neutral fragments. The AspDC system does not contain charged fragments.](https://pubs.acs.org/doi/10.1021/acs.jpcl.3c00533)
the MBE(3) and MBE(4) results is $\lesssim$ 1 kcal/mol, while gas-phase calculations sometimes afford errors of $>150$ kcal/mol at the four-body level! MBE(3) calculations with low-dielectric boundary conditions consistently provide sub-kcal/mol fidelity for various functionals and basis sets, whereas MBE(3) with vacuum boundary conditions affords errors of 10–30 kcal/mol in many cases.

These results suggest that large errors for enzymatic thermochemistry obtained using MBE($n$) with vacuum boundary conditions originate not from the fragmentation approximation, nor do they arise from the simple hydrogen atom caps that we use to saturate the severed valencies. (Our approach to capping is less sophisticated than the use of "conjugated caps" that attempt to replicate amino acid moieties,\textsuperscript{15–28,30,34} yet our results demonstrate that sub-kcal/mol accuracy can be achieved with hydrogen atom caps.) Instead, large and oscillatory errors in MBE($n$) calculations under vacuum boundary conditions arise due to inconsistent charge delocalization in the $n$-body calculations. To obtain a polarization environment that is comparable to that of the supersystem, high-order $n$-body calculations are required, beyond $n = 4$. Alternatively, dielectric boundary conditions provide a simple and low-cost means of mimicking this polarization. In principle, one might consider the use of heterogeneous dielectric boundary conditions\textsuperscript{82–83} such that hydrophobic parts of the protein are treated differently from solvent-exposed portions. This has not been pursued in the present work, where we simply aim to demonstrate that convergence of the MBE in vacuo is not well-defined.

To confirm this explanation, we also examined a different enzymatic reaction that does not involve charged moieties near the active site. For this example, we chose the decarboxylation of l-aspartate by the enzyme l-aspartate α-decarboxylase (AspDC), which has also been studied using QC models of varying size.\textsuperscript{86} Here, we consider only the C–C cleavage step, using a model consisting of 30 monomers (511 atoms), corresponding to a 5 Å radial cutoff around the active site of the relaxed crystal structure. This system has zero net charge but two ionic amino acids, which were placed together in a single fragment to avoid having any charged fragments. Results for $E_a$ and $\Delta_{\text{react}}^\sigma E$ (Figure 1) demonstrate that $n$-body results converge similarly for both vacuum boundary conditions ($\epsilon = 1$) and PCM boundary conditions with $\epsilon = 4$, although the PCM-based error is smaller at the $n = 2$ level. Unlike the charge-coordinated results for COMT, where the fragments are large and thus many-body effects are small, here the $n = 1$ results are unacceptable but two-body results with low-dielectric boundary conditions are rather good.

Together, these results demonstrate that application of MBE($n$) with vacuum boundaries to large enzyme–substrate complexes need not converge to the supersmolecular result at low orders ($n \leq 4$). Oscillatory behavior results from inconsistent charge delocalization across the subsystems (monomer to dimer to trimer, etc.) when the fragments have net charge. Low-dielectric boundary conditions with $\epsilon \approx 1.5$ have been shown to reduce the density delocalization error in isolated peptide DFT calculations,\textsuperscript{77} and in the present context the use of $\epsilon \approx 4$ appears to prevent oscillatory changes to corrections $\Delta E_{\text{dip}}$ $\Delta E_{\text{ip}}$, etc.

Given a two- or three-body approximation for a large enzyme model with charged side chains, one might worry about neglect of long-range interactions. We address this by assessing a multilayer fragmentation scheme in which a low-level calculation on the entire system is used to correct for errors introduced by fragmentation, while the subsystems are described at a higher level of theory. This strategy has been suggested by others under various names,\textsuperscript{88–90} and is illustrated in Figure 2 by analogy to the “ONIOM” approach to QM/MM calculations.\textsuperscript{84} Both the subsystems and the supersystem are computed at the lower level of theory, and the difference between low-level supersystem and low-level MBE($n$) calculations provides a correction for the effects of fragmentation, including neglect of long-range polarization. Raghavachari and co-workers have made extensive use of this idea for calculations in biological macromolecules,\textsuperscript{59–61,92–95} and our two-layer procedure is equivalent to the “MIM2” strategy defined in ref 89.\textsuperscript{6}

We tested several low-level supersystem corrections in combination with various target levels of DFT, for the activation energy in COMT. Errors with respect to the target level (applied to the entire enzyme–substrate model) are illustrated in Figure 3, and numerical values for each supersystem correction are listed in Table S3. The low-level methods that we tested include the semiempirical threecorrected methods HF-3c\textsuperscript{96} and PBEh-3c,\textsuperscript{97} which use a minimal and a double-$\zeta$ basis set, respectively. We also tested conventional Hartree–Fock (HF) theory and the density functional LRC-ωPBEh\textsuperscript{98,99} each with the 6-31G basis set. (Note that 6-31G is much less expensive than other double-$\zeta$ basis sets if the electronic structure software can take advantage of compound shells.\textsuperscript{100}) For this particular 632-atom enzyme–substrate complex, all four of these supersystem corrections require a similar computational effort, which constitutes <20% overhead on top of a MBE(2) calculation.

Even without the supersystem correction, results in Figure 3a indicate that a two-body expansion can achieve $\sim$1 kcal/mol accuracy for $E_a$ using various density functionals. Low-cost supersystem corrections decrease this to $\sim$0.5 kcal/mol. MBE(3) is an order of magnitude more accurate than MBE(2), achieving $\sim$0.1 kcal/mol fidelity even without the supersystem correction. MBE(3) seems to represent something of an accuracy limit, as low-cost supersystem corrections no longer improve the results.

Importantly, the HF-3c supersystem correction performs just as well as HF/6-31G, despite using only a minimal basis set ("MINIX")\textsuperscript{96} For the 632-atom enzyme–substrate
complex used to model COMT, this means 1944 basis functions for HF-3c/MINIX versus 3510 functions for HF/6-31G. On a single 28-core node, these supersystem calculations can be completed in 0.6 h (HF-3c/MINIX) and 1.0 h (HF/6-31G), with 80–90% of that time spent in the PCM solver, which is less well-parallelized as compared to two-electron integrals. The PCM cost could be reduced by using a less dense surface discretization.

Having established that we can obtain converged results, we next turn to computational efficiency. The cost of fragmentation methods is not always discussed honestly and should be measured in aggregate computer time rather than wall time. Timing data for single-point energy calculations on COMT are provided in Figure 4, with the corresponding numerical data in Table S5. In the absence of any supersystem correction, MBE(2) with PCM boundary conditions is ~60% of the cost of a supersystem calculation at the same level of theory (ωB97X-D/def2-SVP), whereas MBE(3) is ~14 times more expensive than the supersystem calculation, although this cost can be distributed across a large number of compute nodes. Despite using larger fragments, the charge-coordinated MBE(3) calculation is actually ~10% cheaper than MBE(3) with single-residue fragments, because the former calculation decreases the number of unique subsystems from 7175 to 3581. This balance would likely shift in favor of the single-residue calculation if a method more expensive than DFT were used.

The number of subsystems required for MBE(\(n\)) increases as \(N^n\) for a protein with \(N\) residues, which imposes a severe computational bottleneck even for \(n = 3\). In what follows, we screen the dimers and trimers on the basis of distance, removing them from the calculation if the minimum interatomic distance between any two fragments exceeds a specified threshold, \(R_{\text{cut}}\). We then recompute \(E_a\) and \(\Delta_{\text{rel}}E\) for COMT, with the caveat that we are careful to ensure that the same residues are included in the reactant, product, and transition state models. Tests of a distance-screened MBE(3) approximation (Figure S5) demonstrate that the predicted value of \(E_a\) changes by <0.1 kcal/mol as \(R_{\text{cut}}\) is decreased from 25 to 8 Å. Setting \(R_{\text{cut}} = 8\) Å decreases the number of subsystems from 7175 to 1499 (as shown in Figure S6), with a negligible effect on accuracy. Using \(R_{\text{cut}} = 8\) Å, the computational effort for ωB97X-D/def2-SVP is reduced from 2025 h (the value shown in Figure 4) to 657 h. The latter
In summary, we find that low-dielectric boundary conditions lead to rapid convergence of the MBE, which otherwise suffers from oscillatory behavior in the presence of charged fragments. Larger, charge-neutral fragments can be used as an alternative strategy to avoid these oscillations, but this will significantly increase the cost if a correlated wave function method is used to describe the subsystems. Because ionic residues must be anticipated in general, this makes dielectric boundary conditions effectively mandatory for QC calculations of enzymatic thermochemistry. These observations furthermore suggest that the use of gas-phase supersystem calculations to benchmark fragmentation approximations distorts the performance of those approximations. Where charged fragments are involved, comparison to a gas-phase calculation may exaggerate the role of higher-order n-body terms.

When dielectric boundaries are employed, MBE(3) provides converged results with sub-kcal/mol fidelity, using single-residue fragments without the need for electrostatic embedding, conjugated caps, or an ONIOM-style supersystem correction. This relatively simple three-body approach represents a reliable fall-back procedure for systems that are too large for conventional DFT. A practical alternative is MBE(2) with distance screening in a double-ζ basis set, plus an ONIOM-style supersystem correction at the HF/6-31G level. This composite approach is converged below 1 kcal/mol with respect to a triple-ζ benchmark and is considerably less expensive than the full-system DFT calculation that it aims to approximate, with a cost that can be readily distributed across hardware.

In the end, we find that enzymatic thermochemistry can be reproduced with sub-kcal/mol fidelity using practical protocols based on fragmentation. The stage is set to push the accuracy of these calculations beyond the DFT level, by means of a hybrid approach that deploys high-level methods for the two-body interactions combined with three-body DFT to capture polarization by the protein environment. We are also exploring the use of fragment-based vibrational frequency calculations, as pioneered by others, to include zero-point corrections and finite-temperature thermal corrections. (The use of smooth cutoffs in gradient calculations has already been demonstrated.) Network analysis can be used to build sensible (if sizable) models of the enzyme–substrate complex and then the protocols developed here can provide converged results for any given model. Together, these developments promise to make QC modeling of enzymatic reactions more robust and systematic.
The crystal structure of COMT (Protein Data Bank (PDB) entry 3BWM) was protonated using the H++ web server at pH 7.0, a salinity of 0.15 M, $e_{in} = 10$, and $e_{out} = 80$. Ligand atoms were protonated separately using PyMOL and then validated against reactant and product structures taken from ref 14. As in that work, the inhibitor 3,5-dinitrocatechol in the crystal structure was replaced with catecholate ($\text{C}_6\text{H}_2\text{O}_{4}^{-}$). Reactant and product structures were relaxed using the GFN2-xTB semiempirical model, with a generalized Born/surface area (GBSA) implicit solvent model for water. Recent benchmarking demonstrates that GFN2-xTB with the implicit solvent affords protein structures that compare favorably to experiment. To obtain the transition state, we scanned the length of the bond between the sulfur atom on SAM and the transferred methyl group. The system was then trimmed to obtain the 632-atom “model 8” from ref 14, which contains residues within a 5 Å radius of the active site along with three important residues identified experimentally. This model affords converged energetics with respect to larger models. For AspDC (PDB entry 1UHE), a single monomer unit can be directly downloaded from the PDB, although the complete structure is an octamer. Starting from the latter, a large radial cutoff of 12 Å was used for structure relaxation with GFN2-xTB in implicit solvent. From that relaxed structure, a smaller 5 Å region was created for a scan along the bond-breaking coordinate. A transition state and a product structure were determined from that scan. For fragmentation calculations, the negatively charged ligand and the cationic arginine residue coordinated to it were included in a single fragment, such that all fragments are charged.

In creating fragments, we avoid cutting the polar $\text{C}–\text{N}$ peptide bond (following previous recommendations), and instead create fragments by cutting the $\text{C}_\text{α}–\text{C_{carbonyl}}$ bond as indicated in Figure S1. The severed valence is capped with a hydrogen atom that is positioned according to eq S1, as in previous work.

All QM calculations were run using a home-built fragmentation code (PyFragment@t), interfaced with Q-Chem. For all calculations, the self-consistent field convergence threshold is set to $\tau_{\text{SCF}} = 10^{-8}$ Ha and the integral and shell-pair drop tolerances are set to $\tau_{\text{int}} = 10^{-12}$ a.u. We use the conductor-like PCM (C-PCM) implemented with the switching/Gaussian discretization scheme. The continuum interface is defined by a van der Waals cavity, constructed using modified Bondi atomic radii that are scaled by a factor of 1.2. That surface is discretized using 110 Lebedev points for hydrogen and 194 points for other atoms. A conjugate gradient algorithm was used to solve the C-PCM equations for the full protein model, whereas matrix inversion was used for the subsystem C-PCM calculations. Calculations with ωB97X-D and M06-2X +D3 use the SG-2 quadrature grid, whereas SG-1 is used for other functionals.

**References**


**Associated Content**

- Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcl.3c00533.

Convergence tests with various functionals and basis sets (PDF)

Coordinates for the enzyme-substrate complexes (ZIP)
49. Lotta, T.; Vidgren, J.; Tilgmann, C.; Ulmanen, I.; Melen, K.; Julkunen, I.; Taskinen, J. Kinetics of human soluble and membrane-bound catechol O-methyltransferase: A revised mechanism and


(64) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. Proteins 1998, 4, 7–18.


(99) Lange, A. W.; Herbert, J. M. Both intra- and interstrand charge-transfer excited states in B-DNA are present at energies comparable to, or just above, the ππ* excitonic bright states. J. Am. Chem. Soc. 2009, 131, 3913−3922.


(102) Lin, H.; Truhlar, D. G. QM/MM: What have we learned, where are we, and where do we go from here? Theor. Chem. Acc. 2007, 117, 185−199.


