Fragment-Based Calculations of Enzymatic Thermochemistry Require Dielectric Boundary Conditions

Paige E. Bowling,1,2 Dustin R. Broderick,2 and John M. Herbert1,2
1Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210 USA
2Dept. of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210 USA

Abstract

Quantum-chemical calculations of enzymatic thermochemistry require hundreds of atoms to obtain converged results, severely limiting the levels of theory that can be used. Fragment-based approaches offer a means to circumvent this problem, and we present calculations on enzyme models containing 500–600 atoms using the many-body expansion with three- and four-body terms. Results are compared to benchmarks in which the supramolecular enzyme–substrate complex is described at the same level of theory. When the amino acid fragments contain ionic side chains, the many-body expansion oscillates under vacuum boundary conditions, exaggerating the role of many-body effects. 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 introduced by fragment-based approximations. For calculations with dielectric boundary conditions, a three-body protocol with distance cutoffs retains sub-kcal/mol fidelity with respect to a supersystem calculation at the same level of theory, as does a two-body protocol when combined with a full-system correction at a low-cost level of theory. Both calculations dramatically reduce the cost of large-scale enzymatic thermochemistry, paving the way for application of high-level ab initio methods to very large systems.

TOC Graphic

Fragment-based approximations1–6 represent an attractive way to circumvent the 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 ranging from \(p = 3\) for density functional theory (DFT) up to \(p = 7\) or higher for levels of theory that provide thermochemical benchmarks. Fragmentation into \(N_{\text{sub}}\) separate subsystems, each of size \(n\), reduces that cost to \(N_{\text{sub}} \times O(n^p)\) in a manner that is amenable to distributed computing and which does not require modification to electronic structure codes. Nonlinear growth in \(N_{\text{sub}}\) with system size can be mitigated by means of distance- or energy-based thresholds.6–8

The present work presents a protocol for using fragmentation to compute enthalpy changes and activation barriers for enzyme-catalyzed reactions. Over the past decade, benchmark QM calculations have revealed that enzymatic thermochemistry does not converge until hundreds of atoms are included,9–17 which is much larger than the QC region in typical quantum mechanics/molecular mechanics (QM/MM) calculations. Fragmentation may therefore offer an efficient route to obtain converged thermochemical calculations at benchmark levels of theory for \(N > 500\) atoms, provided that errors associated with the fragmentation approximation can be controlled. The present work demonstrates that these errors can be reduced below the “thermochemical accuracy” threshold of 1 kcal/mol, yet highlights the fact that straightforward comparison of fragment-based approximations to full-system benchmarks (as a means to assess errors) is ill-posed, if the calculations are carried out with vacuum boundary conditions.

We will consider sizable models of enzyme–substrate complexes containing \(N \sim 500–600\) atoms. Total energies are approximated by means of a many-body expans-
sion (MBE),

\[ E = \sum_{I=1}^{n} E_I + \sum_{I=1}^{n} \sum_{J>I} \Delta E_{IJ} + \sum_{I=1}^{n} \sum_{J>I} \sum_{K>J} \Delta E_{IK} + \cdots. \]  

(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

\[ \Delta E_{IK} = E_{IK} - E_{IJ} - E_{IK} - E_{JK} - E_I - E_J - E_K \]  

(3)

for the three-body corrections. Truncating eq. 1 at \( n \)-body terms, we will denote the resulting approximation as MBE\((n)\). Electrostochastic embedding of the subsystem calculations, using classical point charges derived from the fragment wave functions, is often used in an effort to hasten convergence of the MBE.\(^{18-28}\) We avoid this, however, because we have found that charge embedding can lead to inconsistent convergence of the \( n \)-body expansion.\(^{29-31}\) The use of self-consistent point charges also significantly complicates the formulation of analytic energy gradients.\(^{32-36}\)

Figments \( I, J, K \ldots \) 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,\(^{37}\) we are able to achieve our target accuracy of 1 kcal/mol using mostly single-residue fragments, except for the substrate whose treatment is discussed below. Alternatively, overlapping fragments have sometimes been used for polypeptides and proteins.\(^{24-28,37-42}\) which can be rationalized in terms of a generalized (G)MBE.\(^{1,6,43,44}\) To date, most overlapping-fragment applications use a one-body approach that captures through-bond interactions but not through-space interactions. A two-body GMBE can capture both, but is relatively expensive in terms of the number of subsys-
tems that are generated.\(^{31,37}\) As such, we stick to the simple MBE\((n)\) approach in this work.

As a first test, we consider the methyl transfer catalyzed by human catechol O-
methyltransferase (COMT).\(^{46-48}\) This particular enzyme has become something of a benchmark,\(^{14,49-52}\) because it has a well-resolved crystal structure,\(^{48}\) kinetics data,\(^{46}\) and numerous known inhibitors.\(^{14-16}\) A Mg\(^{2+}\) ion in the active site is essential to its function,\(^{53}\) but leading to charge-transfer effects in QC calculations that can significantly alter the barrier height, depending on the size of the model system.\(^{14,49,52}\) Kulik et al.\(^{14}\) considered a sequence of COMT models with QM regions 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 methyl transfer from S-adenosyl-L-methionine (SAM) to catecholate were protonated and relaxed as described in the 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 MBE\((n)\) towards a well-defined supersystem 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 9 fragments with non-zero charge. Small anions in the gas phase are sometimes inherently unstable (or metastable), as in the case of SO\(_4^{2-}\).\(^{54,55}\) and delocalization errors in DFT can exacerbate this problem.\(^{55}\) To avoid artifacts, charged residues are often neutralized in fragment-based calculations on proteins.\(^{56-58}\) 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 naively 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 Fig. 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 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 change 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 of both \( E_n \) (Fig. 1a) and \( \Delta_{\text{rxn}} E \) (Fig. 1b). 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).\(^{59}\) For protein electrostatics calculations based on the Poisson-Boltzmann equation, it is common to use a dielectric constant in the range \( \varepsilon = 2 - 4 \) to represent the hydrophobic interior of the protein,\(^{60-65}\) although larger values have occasionally been suggested.\(^{66-71}\) The precise value of \( \varepsilon \) may matter for \( pK_a \) calculations, but reaction barrier heights converge quickly as a function of \( \varepsilon \) and results for \( \varepsilon = 2 \) are often indistinguishable from much larger values,\(^{72,73}\) although different from gas-phase (\( \varepsilon = 1 \)) values.

When the fragment calculations required for MBE(n) are performed using PCM boundary conditions, and results compared to a supramolecular calculation with the same boundary conditions, we recover good convergence of MBE(n) even for single-residue fragments having net charge. Results for several other DFT functionals and basis sets are provided in Tables S1 and S2, and in Fig. S2 we extend some of these results to \( n = 4 \) in order to check convergence. Using PCM boundary conditions, the difference between the MBE(3) and MBE(4) results is \( \lesssim 1 \) kcal/mol, while gas-phase calculations sometimes afford errors > 150 kcal/mol at the four-body level! MBE(3) calculations with low-dielectric boundary conditions consistently provide sub-kcal/mol accuracy for various functionals and basis sets, whereas MBE(3) with vacuum boundary conditions affords errors of 10–30 kcal/mol in many cases. Notably, we obtain stable results even when the basis set contains diffuse functions. These can be problematic when self-consistent charge-embedding schemes are used.\(^{6,74-76}\)

These results suggest that large errors for enzymatic thermochemistry obtained using MBE(n) with vacuum boundary conditions originate not from the fragmentation approximation itself, or from the simple hydrogen atom caps that we use to saturate the severed valencies upon fragmentation. (This is less sophisticated as compared to “conjugated caps” that try to replicate amino acid moieties,\(^{25-28,77,78}\) but our results demonstrate that sub-kcal/mol accuracy is achievable even with hydrogen atom caps.) Instead, errors arise due to inconst-tent charge (de)localization in the \( n \)-body calculations for charged fragments. 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 to mimic this polarization. In principle, one might consider the use of heterogeneous dielectric boundaries,\(^{79-81}\) 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 \textit{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 \( \alpha \)-decarboxylase (AspDC), which has also been studied using QC models of varying size.\(^{82}\) 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 we placed together in a single fragment in order to avoid having any charged fragments. Results for \( E_a \) (Fig. 1c) and for \( \Delta_{\text{rxn}} E \) (Fig. 1d) demonstrate that \( n \)-body results with either vacuum (\( \varepsilon = 1 \)) or PCM (\( \varepsilon = 4 \)) boundary conditions converge similarly, 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 the application of MBE(n) with vacuum boundaries, to an enzyme–substrate model extracted from a crystal structure, need
Figure 3: Errors in $E_a$ for COMT, computed using two-layer fragmentation methods at the (a) MBE(2) or (b) MBE(3) level. Target levels of theory (used for the fragments) are indicated by the colored bars, and error is assessed with respect to a supersystem calculation at that level of theory. Several low-cost supersystem corrections are evaluated, as indicated along the horizontal axis. All calculations use PCM boundary conditions with $\varepsilon = 4$ without any distance-based cutoff applied to the MBE(n) calculations.

Given a two- or three-body approximation for a large enzyme model possessing 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 and is illustrated in Fig. 2 by analogy to the “ONIOM” approach for QM/MM calculations. 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 the possible neglect of long-range polarization. Raghavachari and co-workers have made extensive use of this idea for calculations in proteins and our two-layer procedure is equivalent to the “MIM2” strategy defined in Ref. 85.

We tested several low-level supersystem corrections in combination with four different target levels of DFT, for the activation energy in COMT. Errors with respect to the target level of DFT (applied to the entire enzyme–substrate model) are illustrated in Fig. 3 and numerical values for each supersystem correction can be found in Table S3. The low-level methods that we tested include the semi-empirical thrice-corrected methods HF-3c and PBEh-3c, which use a minimal and a double-$\zeta$ basis set, respectively. We also tested Hartree-Fock (HF) theory and the functional LRC-$\omega$PBEh, both 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 sp shells. For this particular 632-atom enzyme–substrate complex, all four of these supersystem corrections require similar computational time, which constitutes less than 20% overhead on top of a MBE(2) calculation.

Even without the supersystem correction, results in Fig. 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 reduce this to $\sim 0.5$ kcal/mol. MBE(3) is an order-of-magnitude more accurate than MBE(2) and achieves $\sim 0.1$ kcal/mol accu-
Figure 4: Total aggregate CPU time (on a logarithmic scale) for a single-point calculation on the COMT enzyme–substrate complex, at the ωB97X-D/def2-SVP level. The supersystem calculation contains 6,042 basis functions and was performed on a single 28-core node (Dell Intel Xeon E5-2680 v4). Fragment calculations were performed on the same hardware with 7 worker processes per node, each using 4 cores.

racy 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, HF-3c performs just as well as HF/6-31G as a supersystem correction despite using only a minimal basis set (“MINIX”). For the 632-atom COMT enzyme–substrate complex, this means 1,944 basis functions for HF-3c/MINIX versus 3,510 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 than the two-electron integrals. (The PCM cost could be reduced by using a less dense surface discretization.)

Having established that we can consistently 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.6,37,95 Timing data for single-point energy calculations on COMT are provided in Fig. 4, with the corresponding numerical data in Table S5. In the absence of any supersystem correction, MBE(2) with PCM boundary conditions costs about 60% as much a supersystem calculation at the same level of theory (ωB97X-D/def2-SVP), whereas MBE(3) is about 14× more expensive than the supersystem calculation. Despite using larger fragments, the charge-coordinated MBE(3) calculation is actually about 10% cheaper than MBE(3) with single-residue fragments, because the former calculation reduces the number of unique subsystems from 7,175 to 3,581. This balance would likely shift in favor of the single-residue calculation if a method more expensive than DFT were used, provided that good fidelity is maintained.

The number of subsystems required for MBE(n) grows as $N^n$ for a protein with $N$ residues, and this combinatorial growth imposes a severe computational bottleneck, even for $n = 3$. In what follows, we screen the dimers and trimers based on distance, removing them from the calculation if the minimum interatomic distance between any two fragments exceeds a specified threshold, $R_{cut}$. We then recompute $E_a$ and $\Delta_{rxn}E$ 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 (Fig. S3) demonstrate that the predicted value of $E_a$ for COMT changes by $< 0.1$ kcal/mol as $R_{cut}$ is reduced from 25 Å to 8 Å. Setting $R_{cut} = 8$ Å reduces the number of subsystems from 7,175 to 1,499 (as shown in Fig. S4), yet has negligible effect on accuracy. Setting $R_{cut} = 8$ Å, the computational effort is reduced from 2,025 h (which is the value shown in Fig. 4) to 657 h. This figure is still 5× greater than the cost of the corresponding supersystem calculation, however.

We include diffuse functions in our next set of tests (ωB97X-D/def2-SVDP), because a method that is intended for general application to enzymatic thermochemistry must be able to accommodate diffuse functions, in order to describe anionic side chains, yet these can be quite problematic for self-consistent charge schemes.6,74,75 Even if electrostatic embedding charges are fixed (say, from a force field), the use of diffuse functions can lead to overpolarization of the QM system by the MM charges.96 Errors in $E_a$ for COMT, computed using MBE(2) and MBE(3) approximations, are provided in Table 1. This includes results both with and without a HF/6-31G supersystem correction, and also with and without distance-based screening using $R_{cut} = 8$ Å. We have also tabulated errors with respect to a ωB97X-D/def2-TZVP calculation, which provides a measure of the basis-set incompleteness error when the smaller def2-TZVP basis set is used.

Both the MBE(2) and MBE(3) approximations at the ωB97X-D/def2-SVDP level afford sub-kcal/mol error with respect to supersystem results using the larger def2-TZVP basis set, suggesting that the basis-set incompleteness error is $< 1$ kcal/mol. MBE(3) achieves this feat without a supersystem correction, inclusion of which scarcely alters the results, whereas the supersystem correction does afford a small but noticeable improvement to MBE(2). It is worth noting that the supersystem ωB97X-D/def2-TZVP calculation on this 632-atom model consists of 11,767 basis functions and requires an aggregate computation time of 17,546 h running on a single 40-core node.

These results once again demonstrate that consistent, sub-kcal/mol accuracy is achievable in two ways: MBE(3) alone, or MBE(2) with a supersystem correction. Distance cutoffs with $R_{cut} = 8$ Å can safely be applied in either case. This consistency indicates that the supersystem correction (which is performed at the HF/6-31G level for the calculations reported in Table 1) primarily accounts for three-body polarization, and that four-
body terms make a negligible (sub-kcal/mol) contribution when PCM boundaries are applied. (See Fig. S2.) Of these two high-fidelity fragment-based procedures, MBE(2) with cutoffs and a supersystem correction is more affordable, by $11\times$ as compared to MBE(3) with cutoffs and no supersystem correction. Although the best measure of real-world cost is total (aggregate) time across all processors, if one wants to use throughput as the figure of merit then it is worth noting that the 379 h required for the supersystem-corrected MBE(2) calculation corresponds to 329 distinct subsystems that can be distributed across compute nodes.

In summary, we find that low-dielectric boundary conditions lead to rapid convergence of the many-body expansion, 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 for the two-body interactions. At the same time, ionic residues must be anticipated in general, and 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 benchmark 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, without the need for electrostatic embedding, conjugated caps, or an ONIOM-style supersystem correction, and using single-residue fragments for most of the protein. This relatively simple three-body approach represents a reliable fall-back procedure for systems that are too large even for conventional DFT. That said, even for a 632-atom enzyme–substrate model, a full-system DFT calculation is far less expensive when a high-performance electronic structure code is used. A practical alternative to MBE(3) is MBE(2) with distance screening, in a double-$\zeta$ 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-$\zeta$ benchmark and is generally less expensive than the full-system calculation. Moreover, that cost is readily distributable 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 now set to push the accuracy of these calculations beyond the DFT level, by means of hybrid methods that deploy high-level methods for the two-body interactions combined with three-body DFT to capture polarization by the environment. We are also exploring the use of fragment-based vibrational frequency calculations, as pioneered by others,97–99 to include zero-point corrections and finite-temperature thermal corrections. (The use of smooth cutoffs in gradient calculations has already been demonstrated.)7 Network analysis can be used to build sensible (if sizable) models of the enzyme–substrate complex,32,100,101 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.

### Computational Details

The crystal structure of COMT (PDB code: 3BMW) was protonated using the H++ web server,102 at pH = 7.0, salinity of 0.15 M, $\varepsilon_{\text{in}} = 10$, and $\varepsilon_{\text{out}} = 80$. Ligand atoms were protonated separately using PyMOL, 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 re-

### Table 1: Errors in $E_a$ for COMT, Computed at the $\omega$B97X-D/def2-SVPD Level.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error in $E_a$ (kcal/mol)</th>
<th>CPU Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBE(3)</td>
<td>-0.00</td>
<td>11.253</td>
</tr>
<tr>
<td>MBE(3) + HF/6-31G</td>
<td>0.04</td>
<td>11.656</td>
</tr>
<tr>
<td>MBE(3) + 8 Å</td>
<td>-0.05</td>
<td>4.170</td>
</tr>
<tr>
<td>MBE(3) + HF/6-31G + 8 Å</td>
<td>0.04</td>
<td>4.300</td>
</tr>
<tr>
<td>MBE(2)</td>
<td>0.81</td>
<td>460</td>
</tr>
<tr>
<td>MBE(2) + HF/6-31G</td>
<td>0.20</td>
<td>491</td>
</tr>
<tr>
<td>MBE(2) + 8 Å</td>
<td>0.83</td>
<td>354</td>
</tr>
<tr>
<td>MBE(2) + HF/6-31G + 8 Å</td>
<td>0.20</td>
<td>379</td>
</tr>
</tbody>
</table>

$^a$All calculations were performed using a PCM with $\varepsilon = 4$. $^b$Error with respect to a supersystem calculation at the same level of theory. $^c$Error with respect to a supersystem calculation at the $\omega$B97X-D/def2-TZVP level. $^d$Aggregate computer time for one single-point energy calculation, using a single 48-core node (Intel Xeon Platinum 8268). Fragment calculations employ 12 worker processes, each running on 4 cores. $^e$HF/6-31G as a supersystem correction. $fR_{\text{cut}} = 8$ Å screening threshold.
placed with catecholate (C₆H₅O₇⁻). Reactant and product structures were relaxed using the GFN2-xTB semi-empirical model with a generalized Born/surface area (GBSA) implicit solvent model for water. GFN2-xTB with implicit solvent has been recently benchmarked for protein structure, with results that compare well to larger models. To obtain the transition state, we scanned the bond length between the sulfur atom on SAM and the transferred methyl group. The system was then trimmed to the bond length 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 active site along with three important residues identified experimentally. This model affords converged energetics with respect to larger models. For AspDC (PDB code: 1UHE), a single monomer unit can be directly downloaded from the protein database although the complete structure is an octamer. Starting from the latter, a large radial cutoff of 12 Å was used for structure relaxation using GFN2-xTB in implicit solvent. From that relaxed structure, a smaller 5 Å region was created for a scan along the bond-breaking coordinate, and from that scan a transition state and a product structure were selected. 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 uncharged.

In creating fragments, we avoid cutting the polar C–N peptide bond (following previous recommendations) and instead create fragments by cutting the C–C bond at Cα, as indicated in Fig. S1. The severed valency is capped with a hydrogen atom that is positioned according to eq. S1, in as previous work.

All QM calculations were run using a home-built interface (PyFragment) to Q-Chem. For all calculations, the self-consistent field convergence threshold is set to $\tau_{SCF} = 10^{-8}$ Ha and the integral and shell-pair drop tolerances are set to $\tau_{ints} = 10^{-12}$ 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.

**Supporting Information**

Additional data including convergence tests with various functionals and basis sets.

**Notes**

The authors declare the following competing financial interest(s): J.M.H. serves on the board of directors of Q-Chem Inc.

**Acknowledgments**

We thank Atsu Agbaglo at the University of Memphis for assistance in creating the octomer of AspDC from the crystal structure of the monomer. This work was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences under Award No. DE-SC0008550. Calculations were performed at the Ohio Supercomputer Center.

**References**


115 “Ohio Supercomputer Center”, [http://osc.edu/ark:/19495/f5s1ph73](http://osc.edu/ark:/19495/f5s1ph73).