JCTC Journal of Chemical Theory and Computation

Open Access
This article is licensed under CC-BY 4.0

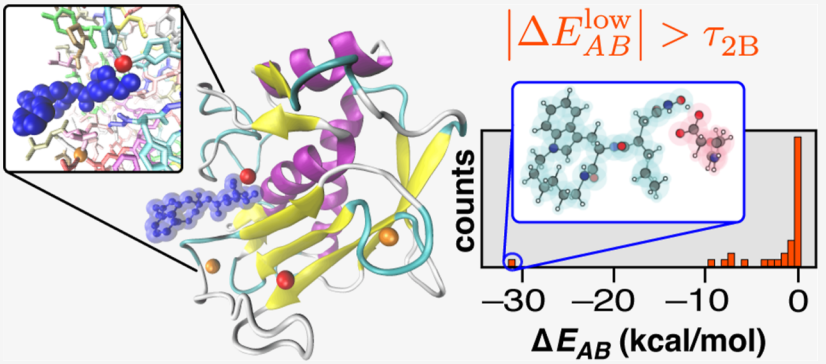
pubs.acs.org/JCTC Article

Energy-Screened Many-Body Expansion for Protein−Ligand Interactions: Examining Convergence for Metalloenzymes Through Seven−Body Interactions

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

Cite This: J. Chem. Theory Comput. 2026, 22, 3720−3731 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Fragment-based quantum chemistry is a powerful strategy for calculating protein−ligand interaction energies using quantum chemistry methods. Rigorous convergence often requires hundreds of atoms in the protein binding-site model, especially if that model is constructed using distance-based criteria to select amino acid residues, while three- and four-body calculations exhibit instability related to combinatorial proliferation in the number of subsystem calculations. Here, we report an energy-based screening protocol for the many-body expansion applied to protein−ligand interactions, implemented in the open-source FRAGME∩T code. Using a combination of aggressive screening based on semiempirical quantum chemistry, with an improved graph-theoretical algorithm to eliminate unimportant subsystems, we are able to perform n-body calculations up to $n = 7$ using density functional theory in triple-ζ basis sets. Distance cutoffs further reduce the cost without compromising accuracy. Rapid and stable convergence of the many-body expansion is obtained by $n = 4$, for a pair of metalloenzymes in which a divalent ion coordinates directly to the ligand. As compared to previous results that relied solely on distance cutoffs, oscillations in the n-body corrections are reduced or eliminated, although residual errors remain in one case. This work demonstrates that benchmark-quality protein−ligand interaction energies can be systematically converged using a method with excellent parallel efficiency and scalability.

Downloaded via OHIO STATE UNIV on April 14, 2026 at 11:13:18 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

1. INTRODUCTION

Enzyme inhibitors comprise up to 50% of existing drugs,[1−3] with noncovalent inhibition as the mechanism pursued by most drug-discovery teams.[4] Therefore, understanding the molecular physics of protein−ligand (P:L) binding and how it changes upon ligand redesign is a fundamental aspect of computer-aided drug discovery. Quantum chemistry (QC) has the potential to offer insight along these lines,[5] although P:L interaction energies are slow to converge with respect to the size of the binding-site model, often requiring 400 atoms or more.[6,7] The same problem affects active-site models of enzymatic reaction energies,[8−10] and results in large part from inter-residue and protein-to-solvent charge transfer.[9−13] These effects require a quantum-mechanical description but QC calculations at this scale are challenging, even with density functional theory (DFT), especially given that double-ζ basis sets do not afford converged results.[7,14] QC calculations at levels of theory beyond DFT are simply intractable without additional approximations.

Fragment-based QC offers a means to this end, and in the present work we examine the many-body expansion (MBE) as a simple paradigm.[15] It starts from one-body fragment energies $\{E_A\}_{A=1,\ldots,N}$ that are systematically corrected for two-body interactions (ΔE_{AB}), three-body interactions (ΔE_{ABC}), etc.:

Received: January 31, 2026
Revised: March 17, 2026
Accepted: March 18, 2026
Published: March 30, 2026

© 2026 The Authors. Published by American Chemical Society

$$E = \sum_{A=1}^{N} \left[E_A + \sum_{B>A} \left(\Delta E_{AB} + \sum_{C>B} \Delta E_{ABC} + \cdots \right) \right]. \quad (1)$$

Explicitly, the two- and three-body corrections are

$$\Delta E_{AB} = E_{AB} - E_A - E_B \quad (2)$$

and

$$\Delta E_{ABC} = E_{ABC} - \Delta E_{AB} - \Delta E_{AC} - \Delta E_{BC} - E_A - E_B - E_C. \quad (3)$$

Higher-order terms are explicated elsewhere.[16] If eq 1 is truncated at n-body terms, we refer to the resulting approximation as MBE(n).

Whereas MBE(n) calculations on water clusters and ion−water clusters unambiguously demonstrate the importance of three- and four-body corrections,[15−20] results for P:L interaction energies are more equivocal. Using single-residue fragments for the protein, we have identified cases where three- and four-body corrections were significant and oscillatory, in the sense that the MBE(4) result deviated from a supramolecular benchmark by a larger amount as compared to the MBE(3) approximation. The problematic cases were a pair of metalloenzymes in which a Zn^{2+} cation binds to the ligand. Divalent ions are know to exacerbate model-size effects,[21] so these are the test systems used in the present work.

Further investigation of these effects has been precluded by combinatorial proliferation of n-body subsystems in high-order MBE(n) calculations where the number of unique subsystems grows as $O(N^n)$. The cost can be mitigated somewhat using distance-based screening, but not to a degree that makes MBE(4) routinely feasible. Thus, MBE(n) calculations face a dilemma: either accept that fragmentation errors cannot be systematically reduced using higher-order terms, or face prohibitive cost (and potential loss of precision) by including terms with $n \geq 4$.[15−18]

For water and ion−water clusters, we have solved this problem with an energy-based screening procedure.[22,23] The idea is to cull the n-body subsystems using an inexpensive level of theory such as a classical force field or semiempirical QC method.[22−24] Only those n-body corrections that register above a user-defined threshold are evaluated at the target level of theory. For a target accuracy of ∼1 kcal/mol, we find this to be more efficient than distance-based screening,[22] in part because it incorporates cooperative effects that are omitted by aggressive distance-based thresholding.[24]

For screening based on semiempirical QC, this energy-based approach must be combined with graph-theoretical techniques to eliminate higher-order subsystems, else the cost to evaluate four-body interactions becomes prohibitive even for the low-level screening method. The resulting "bottom-up" screening algorithm[23] has enabled us to performed converged four-body expansions in $(H_2O)_{64}$ with $N = 64$ fragments. Of the 680,120 distinct subsystems in a complete MBE(4) calculation with $N = 64$, fewer than 1% were computed at the target level of theory. In this way, expansions with $n > 4$ were rendered feasible,[23] which exposed some artifacts in DFT calculations related to the interplay of delocalization error and the n-body expansion.[19,20]

In the present work, we report an implementation of energy-screened MBE(n) for P:L interactions. We then revisit the metalloproteases 1ZP5[25] and 1MMQ,[26] in which an inhibitor is bound to Zn^{2+} (Figure 1). These systems were considered in previous work,[6,7] where we found—unsurprisingly—that the divalent ion creates significant polarization interactions. These cannot be mitigated simply by neutralizing the fragments, as is sometimes done when fragment-based QC is applied to enzymatic systems,[27−29] because the ligand must be removed in order to compute the intermolecular interaction energy.

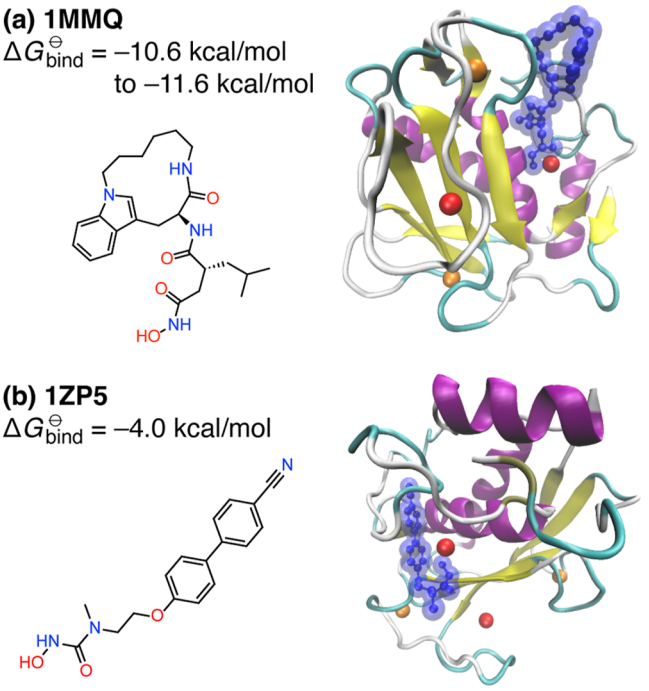


Figure 1. P:L systems considered in this work: (a) matrilysin complexed with a hydroxamate inhibitor (PDB: 1MMQ), and (b) matrix metalloproteinase-8 complexed with a *N*-hydroxyurea inhibitor (PDB: 1ZP5). Chemical structures for the ligands are shown along with crystal structures of the P:L complexes, In the latter, ball-and-stick space-filling ligand models are shown in blue (sans hydrogen atoms), while Zn^{2+} and Ca^{2+} ions appear as red and orange spheres, respectively.

Here, we demonstrate that converged interaction energies can be obtained using low-order n-body calculations based on single-residue fragments, if an appropriate screening strategy is employed. To verify convergence, we extend these calculations all the way to $n = 7$, using DFT in triple-ζ basis sets. This illustrates that the previously observed oscillatory behavior arises, in large part, due to accumulation of roundoff error associated with a rapidly growing number of subsystems. More generally, the protocols developed here provide a framework to extend high-accuracy QC to large metalloenzymes, with controllable convergence. While others have started to use fragment-based QC to obtain P:L interaction energies at levels of theory beyond DFT,[30] convergence with respect to the size of the enzyme model and the treatment of fragmentation have not been demonstrated. The present work provides a paradigm for doing so.

2. THEORY AND METHODS

The methodology described here is implemented in FRAGME∩T,[31] an open-source Python application for fragment-based QC that handles fragmentation, subsystem creation, database management, and parallelization for MBE-

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

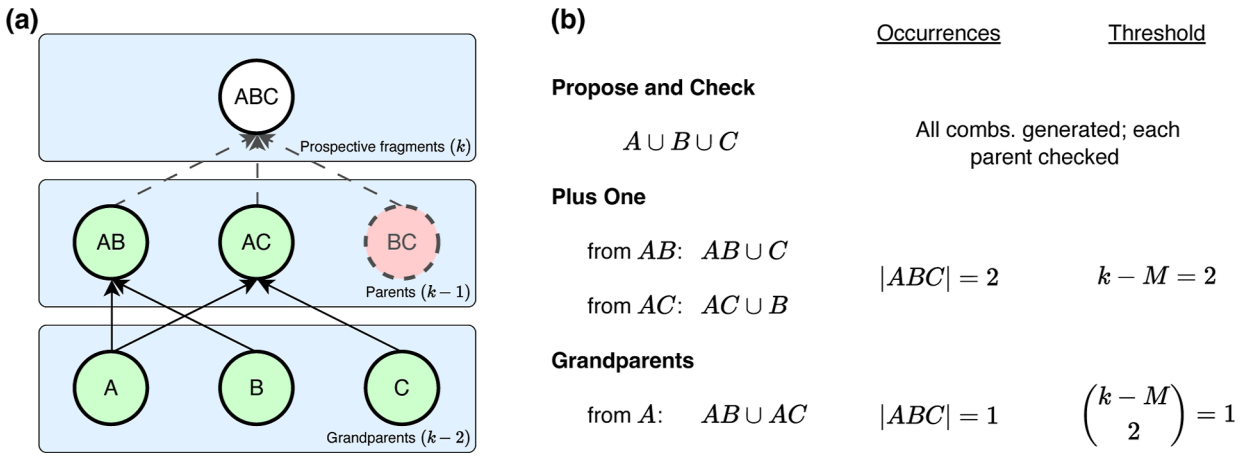


Figure 2. Illustration of bottom-up construction of the many-body interaction graph, applied to a hypothetical system consisting of monomers {*A*, *B*, *C*}. (a) Representation of the subsystems as a directed acyclic graph. Nodes *AB* and *AC* (in green) have survived the two-body screening process and are considered as parents for the trimer layer, whereas *BC* (highlighted in red) failed the screening test and was eliminated. (b) Three possible ways to implement a screening algorithm, as described in the text. All three algorithms afford the same result but their efficiencies differ.

(n) calculations. FRAGME∩T interfaces with numerous QC software packages to perform the electronic structure calculations and these interfaces are easy to add or modify.[31] QC calculations reported in this work were performed using Q-Chem.[32]

FRAGME∩T was developed around the idea of screening, which is described in Section 2.2 following a discussion of MBE(n) for the P:L interaction problem in Section 2.1. Setup of the biomolecular models is described in Section 2.3.

2.1. Interaction Energies

P:L interaction energies are computed using the supramolecular approach

$$\Delta E_{\text{int}} = E_{\text{P:L}} - E_{\text{P}} - E_{\text{L}} \tag{4}$$

by applying a consistent MBE(n) approximation to each of E_{P}, E_{L}, and $E_{\text{P:L}}$. Many terms cancel *a priori* in the difference $E_{\text{P:L}} - E_{\text{P}}$, because $\Delta E_{AB\cdots}$ does not contribute to ΔE_{int} if all fragments $AB\cdots$ reside within the protein.[19] This simplification is handled automatically within FRAGME∩T,[19,31] and it affords dramatic savings as compared to naive application of MBE(n) to eq 4. For the two P:L complexes in Figure 1, elimination of these unnecessary terms reduces the total compute cost by up to 97% for MBE(2) calculations at a semiempirical level of theory,[6] because the only terms ΔE_{AB} that are required are those where *B* is the ligand.

As in previous work,[6,7,33] we use fragments consisting of single amino acids for the protein. The ligand is not fragmented in these calculations, which increases the cost but serves to establish a baseline for convergence of ΔE_{int}. Ligand fragmentation can be considered at a later time, perhaps using tokenization strategies developed for fragment-based drug design.[34−37]

In previous applications of MBE(n) to enzymatic systems, we focused on active-site models with N = 30−35 residues, selected based on distance from the substrate,[33] and we have also used distance thresholds to construct binding-site models for ΔE_{int}(P:L) calculations.[7] However, the metalloenzymes examined here are significantly larger ($N \geq 160$ fragments), with sizable ligands (Figure 1). This necessitates a "bottom-up" graph-theoretical screening approach to make these calculations feasible.[23,31] This approach is described next and is reported here for the first time in protein systems.

2.2. Screening

FRAGME∩T represents a fragmentation scheme in the form of a directed acyclic graph that specifies parent/child relationships between subsystems.[23,31] Each child is the union of its parents; for example, the trimer *ABC* is the child of dimers *AB*, *AC*, and *BC*. The graph is constructed in layers starting with monomer terms followed by dimers, trimers, etc., up to a user-defined terminal order or until no new subsystems can be added due to lack of eligible parents.

As nodes are eliminated from the graph at each n-body order, due to energy and/or distance screening, a given child may be missing one or more of its parents (Figure 2). Let m denote the number of missing parents. Previous work on water clusters, using single-H_2O fragments, suggests that subsystems with $m > 1$ make negligible contribution and can be omitted *a priori*.[23] For those systems, however, the $m = 1$ terms contribute significantly and accuracy is degraded if a more aggressive $m > 0$ screening criterion is applied.

We define a screening parameter M indicating the maximum number of missing parents that is allowed, so that nodes on the graph with $m > M$ are deleted. For the two metalloenzymes considered here, we observe minimal increase in accuracy (∼0.1 kcal/mol) for $M = 0$ versus $M = 1$; see Figure S1. Testing an even more conservative $M = 2$ procedure is prohibitively expensive, adding almost 9000 subsystems for MBE(3) calculations on 1ZP5. However, the negligible difference between $M = 0$ and $M = 1$ results suggests that the latter is sufficient to achieve convergence.

In the present work, we use $M = 1$ as an initial screening criterion at each n-body order, prior to the application of any distance or energy cutoffs.[31] Several possible algorithms for performing this initial screening are discussed in Section 2.2.1. Energy and distance screening are then described in Section 2.2.2.

2.2.1. Screening the Parent/Child Relationships.

Eliminating nodes on the many-body interaction graph is crucial for obtaining the sparsity that makes higher-order MBE(n) calculations tractable.[23,31] Even with a fast semiempirical QC method, energy-based screening becomes prohibitively expansive for $n \geq 4$ unless the interaction graph is first culled based on the number of missing parents. In

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

previous work,[23] the requisite parental checks were performed at many-body order k by considering all

$$\binom{N_P}{k} = \frac{N_P!}{k!(N_P - k)!} \tag{5}$$

potential new subsystems, where N_P is the number of primary fragments. Each proposed subsystem was checked to determine whether $k - M$ parents were present in the graph. We refer to this as the *propose-and-check* algorithm. In the hypothetical example of Figure 2a, we propose adding ABC to the graph at order $k = 3$. Finding that this subsystem has one missing parent ($m = 1$), it is accepted and added as a node in the trimer layer of the graph.

The performance of the propose-and-check algorithm scales poorly with the number of fragments, as illustrated in Figure 3

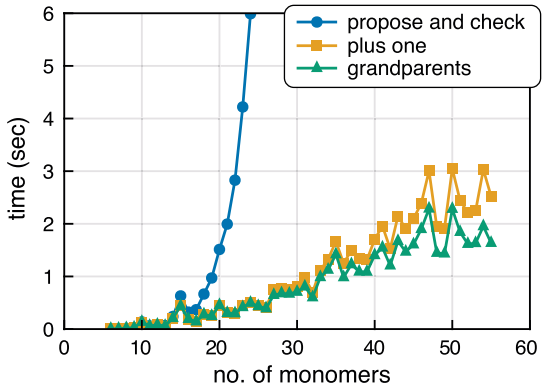


Figure 3. Performance of bottom-up fragmentation using the algorithms described in Figure 2. Benchmarks are reported for $(H_2O)_n$ clusters (N= 6−55) with distance cutoffs d_2 = 6 Å and d_3 = 5 Å. The MBE(n) expansions were allowed to progress up to n = 8.

for $(H_2O)_N$ clusters with N = 6−55 monomers. The cost scales as $O(N_P^k)$ albeit with a dramatically smaller prefactor as compared to performing a semiempirical energy evaluation on each k-mer of fragments. This makes it feasible to extend $(H_2O)_{64}$ calculations to MBE(8) even though there are formally ∼4×10^9 unique octamers. For an enzyme with N = 160 monomers, however, this is almost 9 × 10^{12} terms at n = 8, which is untenable.

Thus, for the present work the bottom-up algorithm was further optimized beyond the simple propose-and-check strategy. To see how, observe that each child differs from its parents by addition of a single monomer (Figure 2a). A *plus-one* algorithm attempts to add each monomer to each prospective parent. In the three-body example of Figure 2a, this means adding A, B, or C to either of AB or BC, since AC has been eliminated. More generally, to add k-body subsystems one must consider $S_P \times S_{k-1}$ where S_P is the set of primary monomers and S_{k-1} is the set of $(k - 1)$-body parents that have not been eliminated from the graph. The occurrences of each unique subsystem are counted and those occurring at least $k - M$ times are proposed for addition to the graph, subject to further energy or distance screening.

In the example of Figure 2a, prospective fragment ABC manifests twice during this process, which meets our threshold of required parents: $k - M = 2$ when $M = 1$ missing parent is allowed. As compared to the propose-and-check algorithm, this plus-one approach better leverages sparsity of the many-body interaction graph, avoiding unnecessary parental oversight. This is reflected in the timing data shown in Figure 3.

The plus-one algorithm can be further optimized by considering the grandparents of prospective fragments. This *grandparents* algorithm (Figure 2b) starts by considering the prospective grandparents, which are A, B, and C for the trimers in Figure 2a. When adding a new generation of subsystems at order k, the pairwise unions of the children of all grandparents (at order $k - 2$) are formed and the unique occurrences are counted. Fragments that occur at least $(k - M)(k - M - 1)/2$ times are proposed for addition to the graph. In the example of Figure 2a the trimer ABC occurs only once, via $AB \cup AC$, which meets the threshold with $k = 3$ and $M = 1$. In that example, grandparents B and C have only a single child each so they do not contribute to the count.

Timings for this algorithm are also shown in Figure 3 and it offers a modest improvement relative to the plus-one algorithm. In what follows, the grandparents algorithm is used wherever feasible. As its name implies, it requires there to be $(k - 2)$-order fragments already in the graph, so the grandparents algorithm is only suitable for three-body and higher-order interactions, and only when $k - M > 1$ where M is the number of missing parents that are allowed. If these conditions are not satisfied, then the plus-one algorithm is used instead.

2.2.2. Energy and Distance Screening. At a given n-body order, energy and/or distance screening is applied following the missing-parents screening. For example, at the n = 3 level we eliminate the trimer ABC if

$$|\Delta E^{\text{low}}_{ABC}| < \tau_{3B} \tag{6}$$

where $\Delta E^{\text{low}}_{ABC}$ is the three-body correction evaluated at a low level of theory and τ_{3B} is a user-specified threshold.

In addition, we might also examine

$$R_{ABC} = \max\{\mathbf{R}_{AB}, \mathbf{R}_{BC}, \mathbf{R}_{AC}\} \tag{7}$$

for some suitable definition of the interfragment distances, say

$$R_{AB} = \min_{a\in A, b\in B}\|\mathbf{r}_a - \mathbf{r}_b\| \tag{8}$$

Three-body distance screening would proceed by comparing R_{ABC} in eq 7 against a user-specified threshold d_3, eliminating ABC if

$$R_{ABC} < d_3 \tag{9}$$

2.3. Computational Methods

Structure preparation closely follows our previous work,[7] but for completeness the protocol is described below along with other details of the procedure.

2.3.1. Structure Preparation and Setup. Crystal structures from the protein data bank (PDB) were protonated using the H++ web server (pH = 7.0, salinity = 0.15 M, ε_{in} = 10, and ε_{out} = 80).[38] Ligands were protonated separately using PyMOL.[39] The resulting P:L structures were then relaxed using GFN-FF,[40] a polarizable force field designed for biological macromolecules, in conjunction with a generalized Born/solvent-accessible surface area (GB/SASA) implicit solvation model, representing water.[41] These relaxed structures are available in the Supporting Information. The 1MMQ structure is the same as that reported previously,[7] but the 1ZP5 structure includes some additional relaxation (using GFN2-xTB) to ensure correct protonation states of the histidine residues located nearest to each metal ion. Following structure

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

relaxation, most crystallographic water molecules were removed except for those that were directly coordinated to the ligand or to ionic moieties.

2.3.2. Fragmentation. As in previous work,[6,7,33] fragments consist of single amino acid residues, cutting the C_α−C(=O) carbon−carbon bond and capping with hydrogen atoms. Note that even the "molecular fractionation with conjugated caps" (MFCC) method for P:L interaction energies is nowadays used with hydrogen-atom caps (albeit with overlapping amino acid fragments),[42] suggesting that more sophisticated capping strategies are unnecessary. The ligand was considered as a single fragment. Ionic cofactors were combined into a monomer with their nearest residues (within 2.5 Å), in order to improve the stability of MBE(n) calculations and to reduce the number of fragments. However, both 1MMQ and 1ZP5 contain a divalent ion within 2.5 Å of the ligand that cannot be combined in this way, because it must be separated from the ligand in order to compute ΔE_{int}.

Energy-based screening was performed using the bottom-up approach described in Section 2.2, allow for $M = 1$ missing parent. This marks the first time that this procedure has been applied to biomolecular systems, so we tested a range of two- and three-body energy thresholds τ_{2B} and τ_{3B}, respectively, defined in the sense of eq 6 and the analogous two-body expression

$$|\Delta E_{AB}^{\text{low}}| < \tau_{2B} \tag{10}$$

We use the semiempirical HF-3c model[43] for the low-level screening. This marks a change from previous work on water clusters,[23] where we used the GFN2-xTB tight-binding model[44] for this purpose. For subsystems with net charge, we find that GFN2-xTB often exhibits convergence problems whereas HF-3c is more robust. Two-body energy screening was performed using thresholds τ_{2B} in the range 0.05−1.00 kcal/mol, while three-body body screening was performed with a consistent threshold $\tau_{3B} = 0.05$ kcal/mol.

A distance threshold was used to further reduce the number of systems in some cases, which will be indicated explicitly. The distance between two subsystems is defined as the minimum interatomic distance between any two atoms (R_{AB} in eq 8), and dimers are eliminated if $R_{AB} < d_2 = 8$ Å. This is the same distance cutoff used in previous work.[6,7,33]

2.3.3. QC Calculations. All calculations were performed using FRAGME∩T,[23] interfaced to Q-Chem[32] (v. 6.3) for the QC calculations. For timing data, calculations were run on 48-processor nodes (Dell PowerEdge C6620 two-socket server), using four worker processes per node, with each individual Q-Chem calculation employing four threads. Supersystem calculations, used to obtain benchmarks for MBE(n) approximations, were performed using a single 48-core node (Intel Xeon Platinum 8268). Timings are reported in terms of total central processing unit (CPU) time, aggregated across all processors, rather than "wall time", because aggregate CPU time is the more appropriate metric for evaluating the true cost of fragment-based QC calculations.[15,45] The self-consistent field convergence criterion was set to 10^{-8} E_h for all calculations. Integral screening and shell-pair drop tolerances were set to 10^{-12} a.u., consistent with recommendations for large-molecule calculations using diffuse basis sets.[46]

The HF-3c model is used for energy screening but we also report some full-protein ΔE_{int}(P:L) values at this level of theory. In addition, DFT calculations are performed using the ωB97X-V functional,[47] which performs well for noncovalent interactions.[47,48] Whereas HF-3c uses a specialized minimal basis set,[43] for the DFT calculations we employ minimally augmented versions[49] of the Karlsruhe augmented basis sets.[50] These are denoted def2-ma-SVP, def2-ma-TZVP, and def2-ma-QZVP.[49] Diffuse functions can be important for noncovalent interaction energies but minimal augmentation is typically sufficient for DFT.[14]

All QC calculations are performed using dielectric boundary conditions with a dielectric constant $\varepsilon = 4$, implemented via the conductor-like polarizable continuum model (C-PCM).[51,52] The value $\varepsilon = 4$ is appropriate for the hydrophobic interior of a protein,[53−58] but at the same time the use of low-dielectric boundaries mitigates spurious oscillations in the MBE(n) sequence of approximations.[19,20,33] Oscillations with respect to the n-body order are especially problematic in the presence of ionic side chains[33] and are driven by self-interaction error that is amplified by fragmentation.[19,20] Even low-dielectric boundaries can provide a driving force for charge localization, however.[20,59,60] (The same effect can be accomplished with embedding charges.[61]) The solute cavity for C-PCM calculations was constructed using Bondi's atomic radii,[62] scaled by 1.2, which is a standard "van der Waals" cavity construction.[52] This interface was discretized using the switching/Gaussian procedure,[63−65] with 50 Lebedev points for hydrogen and 110 points for other nuclei. For larger supersystem calculations of the entire protein, a conjugate gradient implementation of C-PCM was used.[65]

3. RESULTS AND DISCUSSION

As in previous work,[6,7,33] we begin by using HF-3c to compute ΔE_{int} for the unfragmented P:L complex. Using this full-protein baseline, we can then estimate the error introduced by the fragmentation approximation and test the convergence of the screening and n-body approximations (Section 3.1). Results obtained with DFT calculations are discussed subsequently (Section 3.2). Raw data for the plots that follow can be found in the Supporting Information.

3.1. HF-3c Calculations

For 1MMQ, the structure is unchanged from previous work.[7] Thus, the HF-3c result is the same: $\Delta E_{\text{int}} = -178.6$ kcal/mol, requiring 5619 CPU hours on 48 processors. In contrast, $|\Delta E_{\text{int}}|$ for 1ZP5 is smaller than previously estimated due to additional structural relaxation in the present work: $\Delta E_{\text{int}} = -76.7$ kcal/mol, requiring 3138 CPU hours on the same hardware.

3.1.1. Convergence Tests. We first examine convergence of MBE(n) calculations at the HF-3c level for these supramolecular P:L benchmarks, with results for 1MMQ shown in Figure 4 and those for 1ZP5 in Figure 5. Both plots have the same structure and Figure 4a, for example, plots absolute errors in the fragmentation approximation relative to a supramolecular calculation at the same level of theory, using MBE(n) approximations up to $n = 7$. Results are juxtaposed for several different values of the two-body screening threshold (τ_{2B}), using a fixed value $\tau_{3B} = 0.05$ kcal/mol that has worked well for water clusters.[23] (Results for other values of τ_{3B} can be found in Tables S1 and S2). All calculations set $M = 1$ as the maximum number of missing parents, as described in Section 2.2. In addition, a pairwise distance cutoff $d_2 = 8$ Å is used in some calculations. This particular cutoff value is a conservative choice that we have used in previous MBE(n) calculations targeting enzymatic thermochemistry.[6,7,33]

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

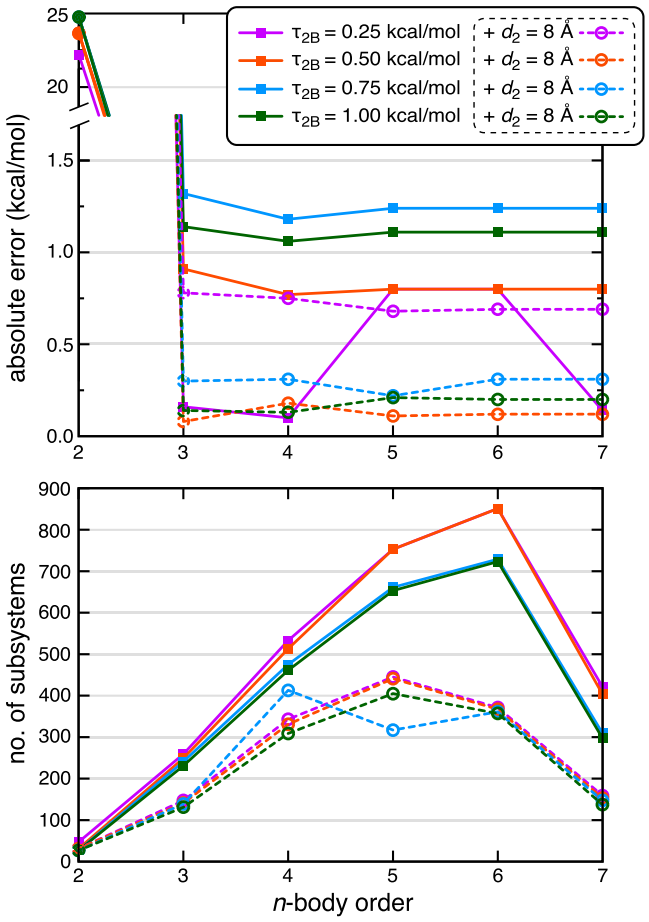


Figure 4. Convergence of MBE(n) calculations at the HF-3c level for ΔE_{int}(P:L) in 1MMQ: (a) absolute errors versus a full-protein benchmark at the same level of theory, and (b) number of distinct subsystem QC calculations required. Results are shown for color-coded values of τ_{2B}, using either energy screening alone (solid lines and square symbols) or else a combination of energy screening with a distance cutoff d_2 = 8 Å (dashed lines and round symbols). All calculations use τ_{3B} = 0.05 kcal/mol and M = 1 to screen the many-body interaction graph. The energy scale in (a) is broken to illustrate the very large errors (up to 25 kcal/mol) at the two-body level, which drop below 1.5 kcal/mol for $n \geq 3$.

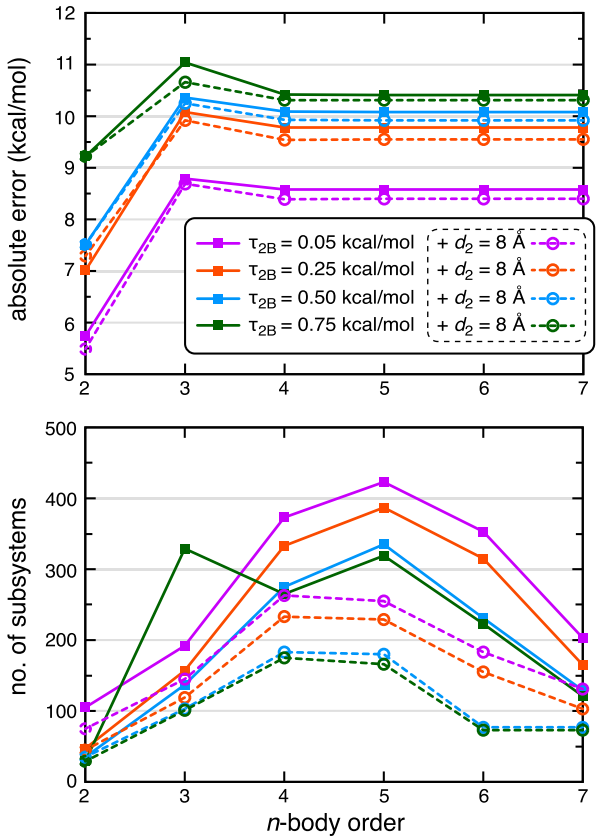


Figure 5. Convergence of MBE(n) calculations at the HF-3c level for ΔE_{int}(P:L) in 1ZP5: (a) absolute errors versus a full-protein benchmark at the same level of theory, and (b) number of distinct subsystem QC calculations required. Results are shown for color-coded values of τ_{2B}, using either energy screening alone (solid lines and square symbols) or else a combination of energy screening with a distance cutoff d_2 = 8 Å (dashed lines and round symbols). All calculations use τ_{3B} = 0.05 kcal/mol and M = 1 to screen the many-body interaction graph.

The number of unique subsystems, which is the number of distinct QC calculations that is required (following screening), is plotted for 1MMQ in Figure 4b at each n-body order, and for 1ZP5 in Figure 5b. We find that the combination of energy and distance screening can reduce the number of target-level QC calculations by ∼50% as compared to energy screening alone, without much effect on ΔE_{int} (typically <1 kcal/mol change). In other words, only residues within 8 Å of the ligand need to be considered. While that number might be expected to be somewhat system-dependent, the presence of several divalent ions in these metalloenzymes likely makes this a fairly conservative estimate for application to other P:L complexes.

Applying d_2 = 8 Å to the 1MMQ system engenders a small reduction in error as compared to the case where no distance threshold is applied (Figure 4a), except for the most conservative value of τ_{2B}. This may simply be error cancellation, as the errors are already small and the same effect is not consistently observed for the tight threshold τ_{2B} = 0.25 kcal/mol. On the other hand, distance screening substantially reduces the number of subsystems (Figure 4b) and may result in a more robust n-body expansion. Distance screening has a much smaller effect on ΔE_{int} for 1ZP5 (Figure 5a), for which the number of subsystems is $\lesssim$ 400 even without distance screening (Figure 5b). This is comparable to the number of subsystems in 1MMQ once d_2 = 8 Å is applied, supporting the idea of minor stability problems when the number of subsystems is larger that ∼400.

A curious effect is that use of higher-order n-body expansions can actually *reduce* the total number of subsystems.[23,31] This effect is explained carefully in ref 31. Briefly, if one imagines inserting formulas for ΔE_{AB} (eq 2) and ΔE_{ABC} (eq 3) into the MBE(3) formula (eq 1), the result would be a large number of individual terms with numerous redundancies. These can be eliminated by analytic resummation, affording only the unique subsystems multiplied by combinatorial coefficients.[16,66] Computationally, FRAGME∩T's database and cryptographic hash architecture accumulates these coefficients (in a manner that is more efficient than other competing software architectures),[31] ensuring that only unique QC calculations are performed. These are subsequently multiplied by an appropriate combinatorial factor. Thus, numerous lower-order terms are subsumed into higher-order corrections as the n-body order increases. The result is a decrease in the number of distinct QC calculations, although

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

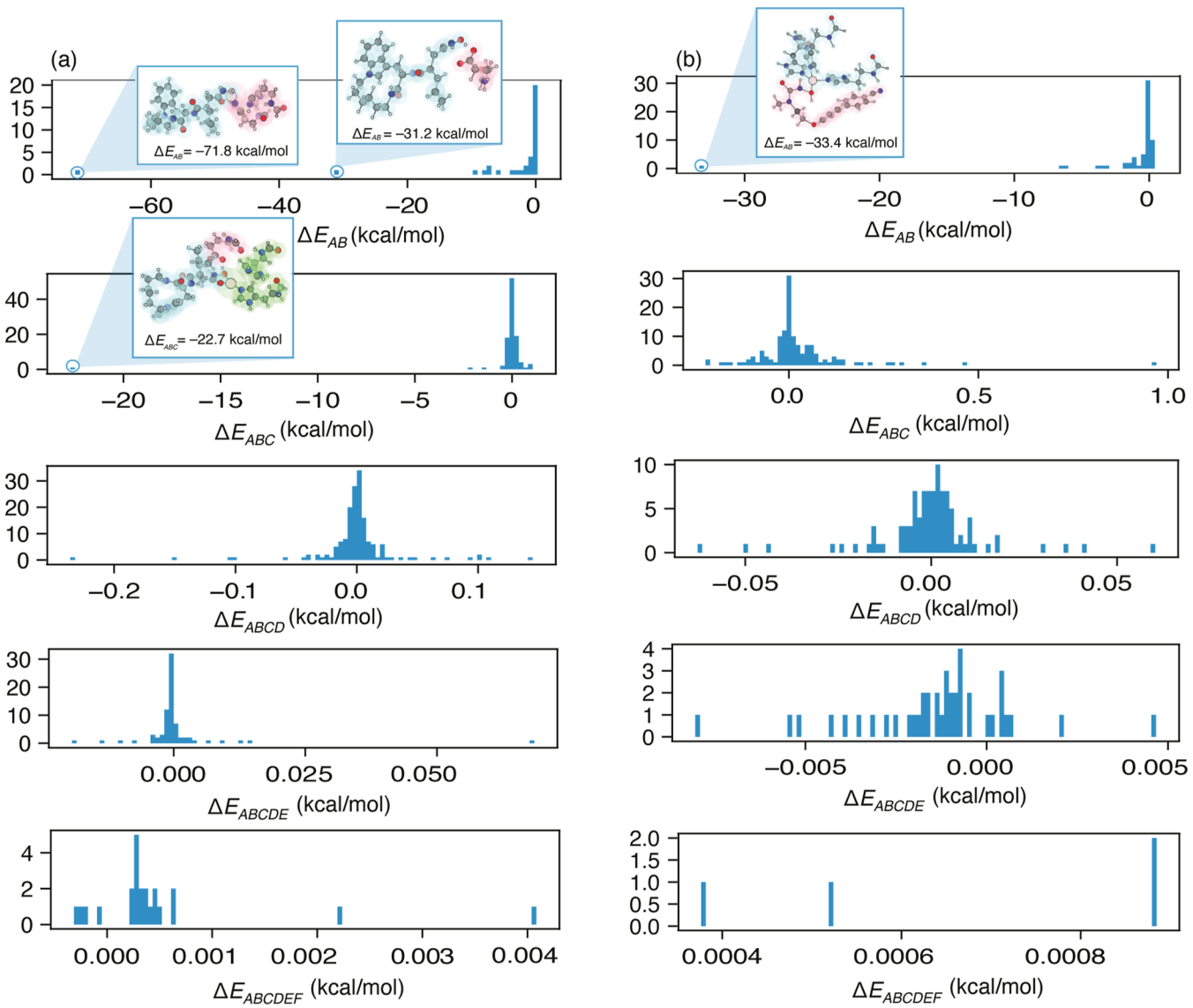


Figure 6. Distribution of MBE(n) terms contributing to ΔE_{int} for (a) 1MMQ and (b) 1ZP5 at the HF-3c level of theory. Various n-body contributions are arranged from the top (two-body corrections ΔE_{AB}) to the bottom (six-body corrections ΔE_{ABCDEF}). These histograms include all terms that are not screened out by the combination of parameters τ_{2B} = 0.25 kcal/mol, τ_{3B} = 0.05 kcal/mol, and M = 1 for 1MMQ and τ_{2B} = 0.05 kcal/mol, τ_{3B} = 0.05 kcal/mol, and M = 1 for 1ZP5. Two- and three-body corrections larger than 20 kcal/mol in magnitude are highlighted. See Table S3 for a tabulation of the net n-body contributions.

the subsystem calculations themselves become larger and more expensive.

A main takeaway from the error convergence plots in Figures 4a and 5a is that it is possible to converge these ΔE_{int}(P:L) calculations at relatively low n-body orders, arguably at n = 3 (for 1MMQ) and certainly by n = 4 (for 1ZP5). These MBE(n) calculations require only a few hundred subsystems. The new screening procedure allows us to demonstrate convergence unambiguously, by extending the calculations all the way to n = 7. For the 1MMQ complex the converged errors are <1.5 kcal/mol, which is <1% of ΔE_{int}.

The 1ZP5 complex, on the other hand, exhibits greater sensitivity with respect to τ_{2B} and also manifests a systematic error that does not converge to zero as n increases. A residual error of 8−10 kcal/mol remains, depending somewhat on the value of τ_{2B}, and does not disappear when τ_{3B} is reduced by a factor of 2 (Table S2). We do not have an explanation for this residual error, which we plan to investigate in future work including the use of larger, overlapping fragments, as in recent MFCC calculations of P:L interactions.[42] That said, the residual error is no larger than 0.07 kcal/mol/fragment. It has been argued that fragmentation errors smaller than 0.1 kcal/mol/fragment are ignorable,[7,24,31] because that value represents 10% of k_BT (per fragment) at room temperature. Electronic structure errors at that level cannot be distinguished from thermal fluctuations.

For 1ZP5, we observe a significant difference in the magnitude of ΔE_{int} as compared to earlier MBE(3) results at the HF-3c level. Partly this is due to better convergence of the present calculations but may also be partially attributable to structural differences between the 1ZP5 models. Here, positions of the protons coordinated to the Zn^{2+} ion nearest the ligand have been adjusted, which in turn affects the hydrogen-bonding network involving the ligand, a water molecule, and the E119 residue. This highlights the critical impact of both structure preparation and also fluctuations,

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

Table 1. ωB97X-V Results for 1MMQ Using Energy (+Distance) Screening

		no. subsystems		ΔE_{int} (kcal/mol)		CPU time (h)	
basis set[a]	MBE (n) order	E-only[b]	$\Delta(E+R)$[c]	E-only[b]	$\Delta(E+R)$[c]	E-only[b]	$\Delta(E+R)$[c]
DZ	2	31	(+0)	−129.5	(+0.00)	40	(+0)
DZ	3	249	(−108)	−147.7	(−1.2)	509	(−182)
DZ	4	513	(−182)	−148.0	(−1.3)	1876	(−635)
DZ	5	753	(−312)	−148.4	(−1.1)	4220	(−1723)
DZ	6	851	(−484)	−148.3	(−1.2)	6236	(−3430)
DZ	7	403	(−250)	−148.3	(−1.2)	4124	(−2900)
TZ	2	31	(+0)	−116.2	(+0.0)	182	(+0)
TZ	3	249	(−108)	−135.8	(−1.4)	2250	(−761)
TZ	4	513	(−182)	−135.5	(−1.4)	8528	(−2813)
TZ	5	753	(−312)	−135.8	(−1.2)	18,878	(−7515)
TZ	6	851	(−484)	−135.7	(−1.3)	27,142	(−14,592)
TZ	7	403	(−250)	−135.7	(−1.2)	15,935	(−10,655)
QZ	2	31	(+0)	−115.7	(+0.0)	1992	(+0)
QZ	3	249	(−108)	−135.9	(−1.4)	24,095	(−8191)

[a]DZ = def2-ma-SVP, TZ = def2-ma-TZVP, QZ = def2-ma-QZVP. [b]Result obtained using energy screening only, with parameters τ_{2B} = 0.5 kcal/mol, τ_{3B} = 0.05 kcal/mol, and M = 1. [c]Change in the energy-screening result when distance screening with d_2 = 8 Å is added.

Table 2. ωB97X-V Results for 1ZP5 Using Energy (+Distance) Screening[bc]

		no. subsystems		ΔE_{int} (kcal/mol)		CPU time (h)	
basis set[a]	MBE(n) order	E-only[a]	$\Delta(E+R)$[a]	E-only[a]	$\Delta(E+R)$[a]	E-only[a]	$\Delta(E+R)$[a]
DZ	2	35	(+0)	−74.5	(+0.0)	28	(+0)
DZ	3	137	(−34)	−75.1	(+0.1)	265	(−60)
DZ	4	275	(−92)	−76.5	(+0.1)	1093	(−402)
DZ	5	335	(−155)	−77.0	(+0.1)	1836	(−841)
DZ	6	231	(−154)	−77.0	(+0.1)	1340	(−907)
DZ	7	129	(−52)	−77.0	(+0.1)	822	(−389)
TZ	2	35	(+0)	−60.7	(+0.0)	135	(+0)
TZ	3	137	(−34)	−59.9	(+0.1)	1315	(−1159)
TZ	4	275	(−92)	−61.0	(+0.1)	5588	(−2047)
TZ	5	335	(−155)	−61.2	(+0.2)	9609	(−8811)
TZ	6	231	(−154)	−61.1	(+0.2)	7146	(−4804)
TZ	7	129	(−52)	−61.1	(+0.2)	4397	(−2055)
QZ	2	35	(+0)	−58.6	(+0.0)	1421	(+0)
QZ	3	137	(−34)	−58.1	(+0.1)	13,611	(−2971)

[a]DZ = def2-ma-SVP, TZ = def2-ma-TZVP, QZ = def2-ma-QZVP. [b]Result obtained using energy screening only, with parameters τ_{2B} = 0.5 kcal/mol, τ_{3B} = 0.05 kcal/mol, and M = 1. [c]Change in the energy-screening result when distance screening with d_2 = 8 Å is added.

which can have a large impact on thermochemical calculations.[67] For this reason, others have emphasized that the use of QC in drug design does not obviate the requirement to perform sampling.[68] QC calculations often converge in 50−100 snapshots along a molecular dynamics trajectory for a single binding pose,[69,70] which is certainly feasible with protocols developed here but lies beyond the scope of the present work.

3.1.2. Analysis of Many-Body Terms. Using energy-based screening, we achieve errors per monomer that are nearly equivalent to what we reported previously using only distance-based screening,[7] yet the present calculations incur only a fraction of the computational cost. Distributions of n-body corrections $\Delta E_{AB...}$ are plotted in Figure 6 for both complexes and the aggregate n-body contributions are provided in Table S3. The most significant contributions to ΔE_{int} arise from the two- and three-body terms with a modest contribution from the four-body terms, suggesting that the convergence observed above is not accidental.

For 1MMQ, the largest contributions to ΔE_{int} arise from interactions between the ligand and either the nearby Zn^{2+} ion or the hydrogen-bonded glutamate residue, the latter of which has a +1 charge. A 24 kcal/mol reduction in the error between n = 2 and n = 3 (Figure 4a) results from inclusion of a trimer consisting of all three moieties, with only marginal contributions from other trimers. These strongly interacting dimers and trimer are shown explicitly in Figure 6a.

In contrast, 1ZP5 has only a single large interaction term, between the ligand and a nearby Zn^{2+} ion (Figure 6b). Nevertheless, there is a significant difference in the accuracy of the MBE(2) and MBE(3) approximations (Figure 5a). This arises from the aggregate effect of numerous three-body terms, none of which is overwhelmingly larger than the others. Unlike 1MMQ, where the importance of the ligand−glumate−Zn^{2+} trimer was probably identifiable *a priori*, this example requires a systematic, automatable approach to identify the important interactions.

3.2. DFT Interaction Energies

HF-3c results demonstrate convergence behavior but may not afford reliable interaction energies, although we have argued that HF-3c values for ΔE_{int}(P:L) may be good enough for

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

some practical purposes if combined with sampling over geometries.[7]

In any case, we wish to examine the use of screened MBE(n) approximations to obtain DFT interaction energies in large binding-site models. Here, one must additionally consider convergence with respect to the choice of basis set. Tables 1 and 2 present DFT results for the two metalloenzymes using the ωB97X-V functional and basis sets ranging from double-ζ to quadruple-ζ. All calculations employ energy screening with HF-3c, and we also consider additional distance screening with d_2 = 8 Å.

A possible concern is that minimal-basis HF-3c calculations might converge at a lower n-body order as compared to DFT calculations that incorporate a more complete description of polarization, but this is not observed in practice. For both P:L complexes, the DFT-MBE(3) value of ΔE_{int} lies within 2 kcal/mol of the DFT-MBE(7) result in a given basis set, and the DFT-MBE(4) value lies within 0.5 kcal/mol of the n = 7 result. These observations are valid for both the double- and triple-ζ basis sets; the quadruple-ζ results extend only to n = 3, for reasons of cost. To some extent, that high cost is an artifact of our decision not to fragment the ligand and doing so might allow higher-order n-body calculations at the basis-set limit although it is not clear what else we would learn about convergence behavior. Many-body counterpoise corrections,[73,74] which are being added to the FRAGME∩T code, are likely to render quadruple-ζ calculations unnecessary.[14]

Although n-body convergence behavior is insensitive to the choice of basis set (including minimal basis sets when considering the HF-3c results), the numerical value of ΔE_{int} certainly depends on the basis set. Differences between converged MBE(n) interaction energies in double- versus triple-ζ basis sets are 12 kcal/mol for 1MMQ and 15 kcal/mol for 1ZP5. Quadruple-ζ results at the two- and three-body level suggest that the triple-ζ results may be within ∼1 kcal/mol of the basis-set limit, which is consistent with other large-scale DFT results for P:L interaction energies.[14]

While energy screening alone facilitates convergence of these DFT-MBE(n) results, both with respect to n and with regard to basis set, the addition of a conservative 8 Å distance cutoff significantly accelerates the calculations. Tables 1 and 2 quantify this in terms of how the cutoff reduces the number of distinct subsystem calculations, modifies ΔE_{int}(P:L), and reduces the aggregate CPU time. With regard to ΔE_{int}(P:L), the 8 Å cutoff changes the result by only about 1.2 kcal/mol for 1MMQ and by ∼0.1 kcal/mol for 1ZP5, while reducing the cost by hundreds or thousands of CPU hours. In some cases, this is a 50% reduction in CPU time as compared to the use of energy screening alone.

This is worth considering with an eye toward eventual sampling over geometries. First, note that P:L interaction energies $|\Delta E_{int}|$ computed using QC are generally at least an order of magnitude larger than thermodynamic binding free energies, $|\Delta G^{\ominus}_{bind}|$.[7,69,70] For the present examples, ΔE_{int}(1MMQ) ≈ − 136 kcal/mol whereas experimental estimates of the binding affinity $\Delta G^{\ominus}_{bind}$(1MMQ) range from −10.6 kcal/mol[26,71] to −11.6 kcal/mol.[71,72] Likewise, $\Delta G^{\ominus}_{bind}$(1ZP5) = −4.0 kcal/mol[71,72] is much smaller than the interaction energy obtained here, ΔE_{int}(1ZP5) ≈ − 61 kcal/mol. These disparities persists even when ΔE_{int} is corrected for the effects of thermal averaging, to obtain an ensemble-averaged value $\langle \Delta E_{int} \rangle$. Thermal fluctuations in ΔE_{int} may be ∼2 kcal/mol or larger under ambient conditions,[69,70] exceeding the convergence errors that are documented here, and entropic corrections are considerably larger still. Nevertheless, correlations between single-pose QC values of ΔE_{int} and experimental binding affinities are sometimes found to be reasonable, even without careful consideration of convergence.[27−29,75−77] In short, the ∼1 kcal/mol convergence targeted here is probably overly conservative. However, the present work does provide a testable means to evaluate convergence, should questions or discrepancies arise.

Along those lines and to contextualize the improvement offered by the combined energy- and distance-based screening approach, we compare the current results for 1MMQ with previous results based on MBE(n) with only distance screening.[7] For an identical geometry of the P:L complex, the distance-only approach affords ΔE_{int} values that differ from those reported here using the composite screening approach, by 3 kcal/mol for MBE(2) and by 7 kcal/mol for MBE(3). However, the older set of calculations were limited to lower-order n-body terms due to proliferation of subsystems in the absence of energy screening. Thus, the composite approach is more reliable for establishing the QC value of ΔE_{int} in the absence of a supramolecular (full-protein) benchmark. The cost is also substantially reduced. For 1MMQ, a DFT-MBE(3) calculation at the ωB97X-V/def2-ma-TZVP level required 62,962 CPU hours as compared to 1489 h for the analogous calculation with energy and distance screening, a 97% decrease in aggregate computing time.

4. CONCLUSIONS

Accurate modeling of P:L interactions using fragment-based QC presents a challenge insofar as long-range decay of electrostatic interactions may render distance-based fragmentation protocols inefficient or inaccurate. The problem is expected to be more severe in the presence of divalent metal cations,[21] as in the examples considered here. Nevertheless, a hybrid screening protocol combining energy-based selection with a conservative distance cutoff overcomes these limitations.

The pivotal modification driving this success is an energy-based filter for n-body subsystems. Distance-based screening, as in our previous approach to computing P:L interactions in these same metalloenzymes,[6,7] relies on cutoffs that indiscriminately incorporate a large number of negligible interactions yet miss others that are distant but electrostatically significant. By evaluating the various n-body corrections at a low-cost semiempirical level of theory, we retain only those terms that contribute significantly to the total energy, bypassing the expensive but physically irrelevant combinatorial proliferation of subsystems. This represents a fundamental paradigm shift toward tractability, compared to previous methods where truncation at MBE(3) or MBE(4) was mandatory and convergence behavior was unclear. By means of energy screening, we are able to extend the MBE(n) calculations to n = 7 and demonstrate unambiguously that results converge at the four-body level.

The addition of a conservative 8 Å distance cutoff on top of this procedure does not fundamentally alter the convergence properties but further reduces the cost, enabling DFT-MBE(n) calculations to be extended to triple-ζ basis sets with diffuse functions. Calculations of at least DFT/triple-ζ quality are crucial to establishing accurate benchmarks for noncovalent interaction energies in large systems.[14] This methodology paves the way for high-throughput, high-accuracy QC

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731

calculations as a means to generate training and/or assessment data for machine learning and other low-cost methods for P:L interactions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.6c00190.

Processed data tables (PDF)
Raw data (XLSX)
Coordinates for the relaxed structural models (PDB)
(PDB)

■ AUTHOR INFORMATION

Corresponding Author

John M. Herbert − *Biophysics Graduate Program and Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-1663-2278; Email: herbert@chemistry.ohio-state.edu

Authors

Paige E. Bowling − *Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States; Biophysics Graduate Program and Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*

Dustin R. Broderick − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States; Department of Chemistry, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-9085-4725

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.6c00190

Notes

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

■ ACKNOWLEDGMENTS

Methods-development 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. Some preliminary work on P:L interactions was supported by was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award no. 1R43GM148095-01A1. Calculations were performed at the Ohio Supercomputer Center.[78]

■ REFERENCES

(1) Imming, P.; Sinning, C.; Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discovery* **2006**, *5*, 821−834.

(2) Robertson, J. G. Enzymes as a special class of therapeutic targets: Clinical drugs and modes of action. *Curr. Opin. Struct. Biol.* **2007**, *17*, 674−679.

(3) Holdgate, G.; Meek, T.; Grimley, R. Mechanistic enzymology in drug discovery: A fresh perspective. *Nat. Rev. Drug Discovery* **2018**, *17*, 115−132.

(4) Singh, J.; Petter, R. C.; Baillie, T. A.; Whitty, A. The resurgence of covalent drugs. *Nat. Rev. Drug Discovery* **2011**, *10*, 307−3017.

(5) Heifetz, A., Ed.*Quantum Mechanics in Drug Discovery; Volume 2114 of Methods in Molecular Biology*; Springer Science+Business Media: New York, 2020.

(6) Bowling, P. E.; Broderick, D. R.; Herbert, J. M. Quick-and-easy validation of protein−ligand binding models using fragment-based semiempirical quantum chemistry. *J. Chem. Inf. Model.* **2025**, *65*, 937−949.

(7) Bowling, P. E.; Broderick, D. R.; Herbert, J. M. Convergent protocols for protein−ligand interaction energies using fragment-based quantum chemistry. *J. Chem. Theory Comput.* **2025**, *21*, 951−966.

(8) Wappett, D. A.; DeYonker, N. J. Accessible and predictable QM-cluster model building for enzymes with the Residue Interaction Network Residue Selector. *Annu. Rep. Comput. Chem.* **2024**, *20*, 131−155.

(9) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How large should the QM region be in QM/MM calculations? The case of catechol *O*-methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381−11394.

(10) Karelina, M.; Kulik, H. J. Systematic quantum mechanical region determination in QM/MM simulation. *J. Chem. Theory Comput.* **2017**, *13*, 563−576.

(11) Nadig, G.; Van Zant, L. C.; Dixon, S. L.; Merz, K. M. Charge-transfer interactions in macromolecular systems: A new view of the protein/water interface. *J. Am. Chem. Soc.* **1998**, *120*, 5593−5594.

(12) Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Charge transfer and polarization in solvated proteins from ab initio molecular dynamics. *J. Phys. Chem. Lett.* **2011**, *2*, 1789−1793.

(13) Kulik, H. J. Large-scale QM/MM free energy simulations of enzyme catalysis reveal the influence of charge transfer. *Phys. Chem. Chem. Phys.* **2018**, *20*, 20650−20660.

(14) Gray, M.; Bowling, P. E.; Herbert, J. M. Systematic examination of counterpoise correction in density functional theory. *J. Chem. Theory Comput.* **2022**, *18*, 6742−6756.

(15) Herbert, J. M. Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.* **2019**, *151*, 170901.

(16) Richard, R. M.; Lao, K. U.; Herbert, J. M. Understanding the many-body expansion for large systems. I. Precision considerations. *J. Chem. Phys.* **2014**, *141*, 014108.

(17) Richard, R. M.; Lao, K. U.; Herbert, J. M. Aiming for benchmark accuracy with the many-body expansion. *Acc. Chem. Res.* **2014**, *47*, 2828−2836.

(18) Lao, K. U.; Liu, K.-Y.; Richard, R. M.; Herbert, J. M. Understanding the many-body expansion for large systems. II. Accuracy considerations. *J. Chem. Phys.* **2016**, *144*, 164105.

(19) Broderick, D. R.; Herbert, J. M. Delocalization error poisons the density-functional many-body expansion. *Chem. Sci.* **2024**, *15*, 19893−19906.

(20) Broderick, D. R.; Herbert, J. M. Untangling sources of error in the density-functional many-body expansion. *J. Phys. Chem. Lett.* **2025**, *16*, 2793−2799.

(21) Mehmood, R.; Kulik, H. J. Both configuration and QM region size matter: Zinc stability in QM/MM models of DNA methyltransferase. *J. Chem. Theory Comput.* **2020**, *16*, 3121−3134.

(22) Liu, K.-Y.; Herbert, J. M. Energy-screened many-body expansion: A practical yet accurate fragmentation method for quantum chemistry. *J. Chem. Theory Comput.* **2020**, *16*, 475−487.

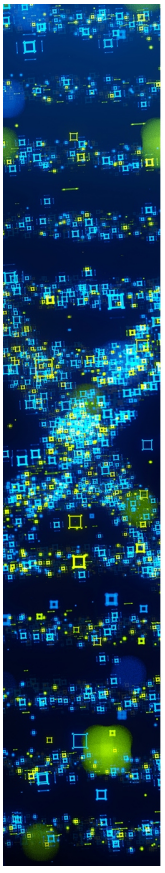
(23) Broderick, D. R.; Herbert, J. M. Scalable generalized screening for high-order terms in the many-body expansion: Algorithm, open-source implementation, and demonstration. *J. Chem. Phys.* **2023**, *159*, 174801.

(24) Ouyang, J. F.; Bettens, R. P. A. When are many-body effects significant? *J. Chem. Theory Comput.* **2016**, *12*, 5860−5867.

(25) Campestre, C.; Agamennone, M.; Tortorella, P.; Preziuso, S.; Biasone, A.; Gavuzzo, E.; Pochetti, G.; Mazza, F.; Hiller, O.; Tschesche, H.; Consalvi, V.; Gallina, G. *N*-hydroxyurea as zinc binding group in matrix metalloproteinase inhibition: Mode of

binding in a complex with MMP-8. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 20−24.
(26) Browner, M. F.; Smith, W. W.; Castelhano, A. L. Matrilysin-inhibitor complexes: Common themes among metalloproteases. *Biochemistry* **1995**, *34*, 6602−6610.
(27) Thapa, B.; Beckett, D.; Jovan Jose, K. V.; Raghavachari, K. Assessment of fragmentation strategies for large proteins using the multilayer molecules-in-molecules approach. *J. Chem. Theory Comput.* **2018**, *14*, 1383−1394.
(28) Thapa, B.; Beckett, D.; Erickson, J.; Raghavachari, K. Theoretical study of protein−ligand interactions using the molecules-in-molecules fragmentation-based method. *J. Chem. Theory Comput.* **2018**, *14*, 5143−5155.
(29) Thapa, B.; Raghavachari, K. Energy decomposition analysis of protein−ligand interactions using molecules-in-molecules fragmentation-based method. *J. Chem. Inf. Model.* **2019**, *59*, 3474−3484.
(30) Gupta, A.; Maier, S.; Thapa, B.; Raghavachari, K. Towards post-Hartree−Fock accuracy for protein−ligand affinities using the molecules-in-molecules fragmentation-based method. *J. Chem. Theory Comput.* **2024**, *20*, 2774−2785.
(31) Broderick, D. R.; Bowling, P. E.; Brandt, C.; Childress, S.; Shockey, J.; Higley, J.; Dickerson, H.; Ahmed, S. S.; Herbert, J. M. Fragme∩t: An open-source framework for multiscale quantum chemistry based on fragmentation. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2025**, *15*, No. e70058.
(32) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L.; et al. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *J. Chem. Phys.* **2021**, *155*, 084801.
(33) Bowling, P. E.; Broderick, D. R.; Herbert, J. M. Fragment-based calculations of enzymatic thermochemistry require dielectric boundary conditions. *J. Phys. Chem. Lett.* **2023**, *14*, 3826−3834.
(34) Firth, N. C.; Atrash, B.; Brown, N.; Blagg, J. MOARF, an integrated workflow for multiobjective optimization: Implementation, synthesis, and biological evaluation. *J. Chem. Inf. Model.* **2015**, *55*, 1169−1180.
(35) Bian, Y.; Xie, X.-Q. Computational fragment-based drug design: Current trends, strategies, and applications. *AAPS J.* **2018**, *20*, 59.
(36) Wilson, J.; Sokhansanj, B. A.; Chong, W. C.; Chandraghatgi, R.; Rosen, G. L.; Ji, H.-F. Fragment databases from screened ligands for drug discovery (FDSL-DD). *J. Mol. Graph. Model.* **2024**, *127*, 108669.
(37) He, G.; Liu, S.; Liu, Z.; Wang, C.; Zhang, K.; Li, H. Prototype-based contrastive substructure identification for molecular property prediction. *Brief. Bioinform.* **2024**, *25*, bbae565.
(38) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating p*K* prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulation. *Nucleic Acids Res.* **2012**, *40*, 537−541.
(39) PyMOL, a user-sponsored molecular visualization system on an open-source foundation, maintained and distributed by Schrödinger. https://pymol.org.
(40) Spicher, S.; Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem. Int. Ed. Engl.* **2020**, *59*, 15665−15673.
(41) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. Robust and efficient implicit solvation model for fast semiempirical methods. *J. Chem. Theory Comput.* **2021**, *17*, 4250−4261.
(42) Zhang, Y.; Xia, W.; Huang, K.; Xiao, J.; Zhang, J. Z. H. Accurate and efficient calculation of protein−ligand interaction energies using an electrostatically embedded fragmentation method. *J. Chem. Theory Comput.* **2026**, *22*, 1514−1523.
(43) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(44) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—an accurate and broadly parameterized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(45) Gavini, V.; et al. Roadmap on electronic structure codes in the exascale era. *Modell. Simul. Mater. Sci. Eng.* **2023**, *31*, 063301.
(46) Gray, M.; Bowling, P. E.; Herbert, J. M. Comment on "Benchmarking basis sets for density functional theory thermochemistry calculations: Why unpolarized basis sets and the polarized 6−311G family should be avoided". *J. Phys. Chem. A* **2024**, *128*, 7739−7745.
(47) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904−9924.
(48) Gray, M.; Herbert, J. M. Density functional theory for van der Waals complexes: Size matters. *Annu. Rep. Comput. Chem.* **2024**, *20*, 1−61.
(49) Gray, M.; Herbert, J. M. Comprehensive basis-set testing of extended symmetry-adapted perturbation theory and assessment of mixed-basis combinations to reduce cost. *J. Chem. Theory Comput.* **2022**, *18*, 2308−2330.
(50) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **2010**, *133*, 134105.
(51) Lange, A. W.; Herbert, J. M. Symmetric versus asymmetric discretization of the integral equations in polarizable continuum solvation models. *Chem. Phys. Lett.* **2011**, *509*, 77−87.
(52) Herbert, J. M. Dielectric continuum methods for quantum chemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1519.
(53) Gilson, M. K.; Honig, B. H. The dielectric constant of a folded protein. *Biopolymers* **1986**, *25*, 2097−2119.
(54) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins* **1988**, *4*, 7−18.
(55) Rodgers, K. K.; Silgar, S. G. Surface electrostatics, reduction potentials, and internal dielectric constant of proteins. *J. Am. Chem. Soc.* **1991**, *113*, 9419−9421.
(56) Nakamura, H. Roles of electrostatic interaction in proteins. *Q. Rev. Biophys.* **1996**, *29*, 1−90.
(57) Grochowski, P.; Trylska, J. Continuum molecular electrostatics, salt effects, and counterion binding—A review of the Poisson−Boltzmann theory and its modifications. *Biopolymers* **2008**, *89*, 93−113.
(58) Alexov, E.; Mehler, E. L.; Baker, N.; M Baptista, A.; Huang, Y.; Milletti, F.; Erik Nielsen, J.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. Progress in the prediction of pK_a values in proteins. *Proteins* **2011**, *79*, 3260−3275.
(59) Ren, F.; Liu, F. Impacts of polarizable continuum models on the SCF convergence and DFT delocalization error of large molecules. *J. Chem. Phys.* **2022**, *157*, 184106.
(60) Alam, B.; Jiang, H.; Zimmerman, P. M.; Herbert, J. M. State-specific solvation for restricted active space spin-flip (RAS-SF) wave functions based on the polarizable continuum formalism. *J. Chem. Phys.* **2022**, *156*, 194110.
(61) Jiang, Y.; Ho, J. Counterpoise correction and charge embedding as antidotes for delocalization error in density functional many-body expansion. *J. Phys. Chem. Lett.* **2025**, *16*, 13162−13169.
(62) Rowland, R. S.; Taylor, R. Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der Waals radii. *J. Phys. Chem.* **1996**, *100*, 7384−7391.
(63) Lange, A. W.; Herbert, J. M. Polarizable continuum reaction-field solvation models affording smooth potential energy surfaces. *J. Phys. Chem. Lett.* **2010**, *1*, 556−561.
(64) Lange, A. W.; Herbert, J. M. A smooth, nonsingular, and faithful discretization scheme for polarizable continuum models: The switching/Gaussian approach. *J. Chem. Phys.* **2010**, *133*, 244111.
(65) Herbert, J. M.; Lange, A. W. Polarizable continuum models for (bio)molecular electrostatics: Basic theory and recent developments

for macromolecules and simulations. In *Many-Body Effects and Electrostatics in Biomolecules*; Cui, Q.; Ren, P.; Meuwly, M., Eds.; CRC Press: Boca Raton, 2016; Chapter 11, pages 363−416.
(66) Kaplan, I. G.; Santamaria, R.; Novaro, O. Non-additive forces in atomic clusters. The case of Ag_n. *Mol. Phys.* **1995**, *84*, 105−114.
(67) Vysotskiy, V. P.; Ryde, U. Exploring the high sensitivity of DFT thermochemistry for protonation states of a ferredoxin model complex $[(CH_3S)_4Fe_2^{III}S_2H]^-$. *J. Chem. Phys.* **2025**, *162*, 165101.
(68) Merz, K. M., Jr. The role of quantum mechanics in structure-based drug design. In *Drug Design: Structure- and Ligand-Based Approaches*; Merz, K. M.; Ringe, D.; Reynolds, C. H., Eds.; Cambridge University Press: Cambridge, UK, 2010; Chapter 8, pages 120−136.
(69) Fox, S. J.; Dziedzic, J.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Density functional theory calculations on entire proteins for free energies of binding: Application to a model polar binding site. *Proteins* **2014**, *82*, 3335−3346.
(70) Gundelach, L.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Protein-ligand free energies of binding from full-protein DFT calculations: Convergence and choice of exchange-correlation functional. *Phys. Chem. Chem. Phys.* **2021**, *23*, 9381−9393.
(71) Hu, Y.; Furtmann, N.; Gütschow, M.; Bajorath, J. Systematic identification and classification of three-dimensional activity cliffs. *J. Chem. Inf. Model.* **2012**, *52*, 1490−1498.
(72) Thapa, B.; Erickson, J.; Raghavachari, K. Quantum mechanical investigation of three-dimensional activity cliffs using the molecules-in-molecules fragmentation-based method. *J. Chem. Inf. Model.* **2020**, *60*, 2924−2938.
(73) Richard, R. M.; Lao, K. U.; Herbert, J. M. Achieving the CCSD(T) basis-set limit in sizable molecular clusters: Counterpoise corrections for the many-body expansion. *J. Phys. Chem. Lett.* **2013**, *4*, 2674−2680.
(74) Richard, R. M.; Lao, K. U.; Herbert, J. M. Approaching the complete-basis limit with a truncated many-body expansion. *J. Chem. Phys.* **2013**, *139*, 224102.
(75) Maier, S.; Thapa, B.; Erickson, J.; Raghavachari, K. Comparative assessment of QM-based and MM-based models for prediction of protein−ligand binding affinity trends. *Phys. Chem. Chem. Phys.* **2022**, *24*, 14525−14537.
(76) Liu, J.; He, X. QM implementation in drug design: Does it really help?. In *Quantum Mechanics in Drug Discovery*, Vol. *2114*; Heifetz, A., Ed.; Springer Science+Business Media: New York, 2020; Chapter 2, pages 19−36.
(77) Cavasotto, C. N. Binding free energy calculation using quantum mechanics aimed for drug lead optimization. In *Quantum Mechanics in Drug Discovery*, Vol. *2114*; Heifetz, A., Ed.; Springer Science+Business Media: 2020; ding Chapter 16, pages 257−268.
(78) Ohio Supercomputer Center. http://osc.edu/ark:/19495/f5s1ph73 (accessed March 17, 2026).


CAS BIOFINDER DISCOVERY PLATFORM™
STOP DIGGING THROUGH DATA —START MAKING DISCOVERIES
CAS BioFinder helps you find the right biological insights in seconds
Start your search
CAS
A division of the American Chemical Society

https://doi.org/10.1021/acs.jctc.6c00190
J. Chem. Theory Comput. 2026, 22, 3720−3731