JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

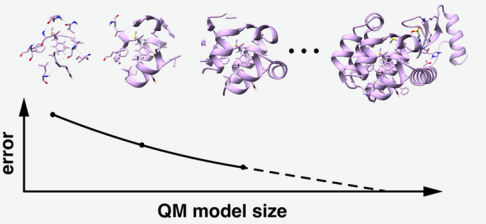
Quick-and-Easy Validation of Protein−Ligand Binding Models Using Fragment-Based Semiempirical Quantum Chemistry

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

Cite This: J. Chem. Inf. Model. 2025, 65, 937−949 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Electronic structure calculations in enzymes converge very slowly with respect to the size of the model region that is described using quantum mechanics (QM), requiring hundreds of atoms to obtain converged results and exhibiting substantial sensitivity (at least in smaller models) to which amino acids are included in the QM region. As such, there is considerable interest in developing automated procedures to construct a QM model region based on well-defined criteria. However, testing such procedures is burdensome due to the cost of large-scale electronic structure calculations. Here, we show that semiempirical methods can be used as alternatives to density functional theory (DFT) to assess convergence in sequences of models generated by various automated protocols. The cost of these convergence tests is reduced even further by means of a many-body expansion. We use this approach to examine convergence (with respect to model size) of protein−ligand binding energies. Fragment-based semiempirical calculations afford well-converged interaction energies in a tiny fraction of the cost required for DFT calculations. Two-body interactions between the ligand and single-residue amino acid fragments afford a low-cost way to construct a "QM-informed" enzyme model of reduced size, furnishing an automatable active-site model-building procedure. This provides a streamlined, user-friendly approach for constructing ligand binding-site models that requires neither *a priori* information nor manual adjustments. Extension to model-building for thermochemical calculations should be straightforward.



1. INTRODUCTION

Convergence of electronic structure calculations on systematically larger enzyme models is slow,[1−15] requiring 300−600 atoms or more before the result no longer changes with respect to the inclusion of additional amino acids in the quantum mechanical (QM) model region. This is true whether the quantity of interest is a barrier height or a reaction energy,[1−13] or whether it is the interaction energy for noncovalent binding of a ligand to a protein.[15] In view of this, the current state-of-the-art for modeling enzymatic active sites or ligand binding sites using quantum chemistry relies on bespoke or "artisanal" QM models, constructed to purpose by hand, without well-defined criteria to guide the process. Slowly this is beginning to change, as tools for automated QM model selection are developed.[11−13,15−19]

In the present work, we evaluate the use of such procedures for obtaining energetically converged molecular models of ligand binding sites in enzymes. Our strategies combines a semiempirical quantum chemistry model (namely, HF-3c)[20] with a fragment-based procedure for computing the interaction energy (ΔE_{int}) between a ligand and an enzyme model.[15] The latter is constructed in an automated way, and this facilitates high-throughput investigation of a large number of enzyme models at low cost. Given an appropriate model, one can then apply convergent, fragment-based protocols to evaluate ΔE_{int} at higher levels of theory.[15] That might be density functional theory (DFT), although the fragments are small enough that the use of correlated wave function models is also feasible.

The fragment-based approach leverages the power of distributed computing to reduce a single, monolithic (and potentially intractable) calculation into a large but manageable number of subsystem calculations.[21−27] This enables large-scale quantum chemistry calculations using only workstation-level resources (i.e., single-node parallelism),[15,28−30] as the storage footprint of a given calculation is reduced to that of the largest subsystem. This is an important consideration for investigators at under-resourced institutions. The present calculations bring protein−ligand binding calculations, at QM levels of theory, into the realm of what can be accomplished readily on workstation resources.

2. METHODS

2.1. Fragmentation. We use the many-body expansion (MBE) for calculations on proteins. This is a telescoping expansion for the total ground-state energy E, starting from

Received: October 26, 2024
Revised: December 2, 2024
Accepted: December 16, 2024
Published: January 3, 2025

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

energies $\{E_I\}$ for a collection of independent fragments ($I = 1, \cdots, N_{\text{frag}}$):

$$E = \sum_{I=1}^{N_{\text{frag}}} E_I + \sum_{I=1}^{N_{\text{frag}}} \sum_{J<I} \Delta E_{IJ} + \sum_{I=1}^{N_{\text{frag}}} \sum_{J<I} \sum_{K<J} \Delta E_{IJK} + \cdots \quad (1)$$

Here, the gross energy $\sum_I E_I$ is corrected via two-body terms

$$\Delta E_{IJ} = E_{IJ} - E_I - E_J \quad (2)$$

three-body terms

$$\Delta E_{IJK} = E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} - E_I - E_J - E_K \quad (3)$$

and so forth.[25,31] If eq 1 is truncated at n-body terms, then we refer to the resulting method as MBE(n).

As in previous work on proteins,[15,32] we use single-residue fragments obtained by cutting the C−C bond at C_α−C(=O), avoiding the more polar peptide (C−N) bond. The severed valence is capped with a hydrogen atom positioned at

$$\mathbf{r}_{\text{cap}} = \mathbf{r}_1 + \left(\frac{R_1 + R_H}{R_1 + R_2}\right)(\mathbf{r}_2 - \mathbf{r}_1) \quad (4)$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the positions of the atoms in the original C−C bond.[33] The quantities $R_1 = R_2 = 0.76$ Å and $R_H = 0.31$ Å are covalent radii for carbon and for hydrogen, respectively.[34] (This procedure affords C−H bond lengths of about 1.07 Å for the capping atoms.) Although more sophisticated capping strategies have been suggested,[35−38] we have not found them to be necessary.

In some of the calculations presented below, distance-based screening is used to reduce the number of subsystem calculations required for MBE(n). In that case, subsystems are omitted if the minimum interatomic distance between any two fragments exceeds a specified threshold, R_{cut}. In previous work on protein fragmentation,[15,32] we showed that $R_{\text{cut}} = 8$ Å affords results that are converged (with sub-kcal/mol fidelity) with respect to the equivalent MBE(n) calculation performed using all possible subsystems. As an example of the cost savings that is engendered, consider the T4-lysozyme complex with the protein data bank (PDB) code 181L, which is one of the systems considered below. In that case, $N_{\text{frag}} = 164$ for the entire protein system but the use of $R_{\text{cut}} = 8$ Å reduces the number of subsystems for a MBE(3) calculation from 708,561 to 16,016, a 98% reduction.

Both the capping in eq 4 and the distance-based screening are performed automatically using our open-source FragmeNT code,[29,39] which drives all of the calculations reported here. FragmeNT implements both distance- and energy-based screening protocols[28−30] and is interfaced with a variety of quantum chemistry packages. All calculations reported in this work use Q-Chem v. 6.0 as the quantum chemistry engine.[40] Calculations were performed on 28-core nodes (Dell Intel Xeon E5-2680 v4) by packing four subsystem calculations onto each node with seven threads assigned to each Q-Chem process.

Single-pose protein−ligand interaction energies ΔE_{int} are computed according to

$$\Delta E_{\text{int}} = E_{\text{P:L}} - E_{\text{P}} - E_{\text{L}} \quad (5)$$

with consistent application of MBE(n) to compute both the energy of the isolated protein (E_{P}) and that of the protein−ligand complex ($E_{\text{P:L}}$). The ligand energy (E_{L}) is computed without fragmentation. Many of the n-body terms will cancel in eq 5 and need not be computed. The present version of FragmeNT identifies these terms *a priori* and removes them, using the algorithm described in ref 30, which leads to considerable cost savings for ΔE_{int} calculations. However, the present calculations were performed contemporaneously with that development and this savings is not exploited here. As such, timing data reported herein reflect the cost of all n-body terms in eq 5, subject only to a distance cutoff (R_{cut}).

Use of eq 5 is subject to basis-set superposition error (BSSE) because we use atom-centered Gaussian basis sets. This effect can be quite significant in protein−ligand models containing hundreds of atoms, especially when the ligand is large. In protein−ligand models with ∼300 atoms, for example, BSSE effects up to ∼50 kcal/mol have been documented when double-ζ basis sets are used,[41] as quantified by the difference between counterpoise-corrected and uncorrected values of ΔE_{int}. Versions of counterpoise correction designed for use with MBE(n) have been reported[42−47] but are not yet implemented in FragmeNT, although that work is underway. In lieu of counterpoise correction, we will consider the use of larger basis sets in order to evaluate the importance of BSSE.

2.2. Systems and Structure Preparation. Systems considered here were previously examined in the course of establishing MBE(n) protocols for protein−ligand interaction energies.[15] As baseline cases, we selected two structures (181L and 1LI2) from a set of T4-lysozyme complexes,[48−50] whose ligands are benzene and phenol. Both complexes contain two Cl^- ions, which we combined into a single fragment along with all residues within 2.5 Å of the ion. We also consider four complexes (1O48, 1BOZ, 1MMQ, and 1ZP5) containing large ligands that are more representative of typical noncovalent inhibitors. Ligand structures, in the protonation states that are examined here, are provided in Figure 1. Each ligand is charge-neutral.

Crystal structures were obtained from the PDB and protonated using the H++ web server,[51] specifying pH = 7.0,

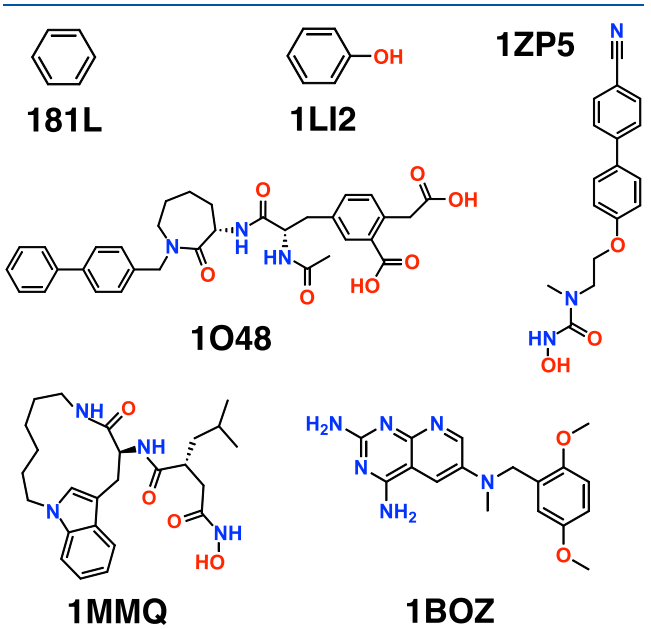


Figure 1. Ligands examined in this work, labeled with the PDB code of the corresponding protein−ligand complex.

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

salinity of 0.15 M, $\epsilon_{in} = 10$, and $\epsilon_{out} = 80$.[52] The large ligands were protonated separately using the PyMol program.[53] As in previous work,[15] geometries were then relaxed using the GFN2-xTB semiempirical method,[54] in conjunction with a generalized Born/solvent-accessible surface area (GBSA/SASA) implicit solvation model for water.[55] Most crystallographic water molecules were removed after relaxation, although the ones coordinated to the Cl^- ions were retained, as were any crystallographic water molecules within 2.5 Å of the ligand.

2.3. Model Construction. We insist on automated methods that provide a reproducible, black-box approach to QM model construction, which does not rely on any system-specific information beyond what is contained in a crystal structure. Structural models for the complexes described in Section 2.2, containing anywhere from 120 to 1726 atoms, were generated by one of several different approaches that are described below.

The simplest approach uses a distance criterion (d) to select amino acid residues proximate to the ligand. In this case, residue selection was performed using PyMol with a cutoff ranging from $d = 2.5$ Å to $d = 10.0$ Å. Residues are included in the QM model if they have at least one atom (including hydrogen) within a distance d from any ligand atom. This method is simple and systematic but its weakness lies in the fact that many enzymatic active sites are aspherical. In such cases, one might expect a structure-aware algorithm to converge more quickly than a brute-force approach based on distance. To examine this, we consider models generated using a Probe method,[56,57] or alternatively via the Arpeggio library.[58,59] Both methods take atomistic and residue-specific information into account. These methods are used as implemented in the rinrus toolkit,[60] with construction and capping of the QM models completed within the Fragme∩t code. Note that rinrus contains additional functionality for building QM-cluster models that is not exploited here; see ref 18 for a recent overview.

These various approaches are illustrated in Figure 2. The Probe method rolls a sphere over the van der Waals surface of a seed moiety (for which we use the ligand), in order to identify close-contact residues.[56,57] These are classified into different categories depending on the contact distance, with hydrogen bonds as a separate category that also depends on atomic identity. rinrus uses the Probe classifications to assemble a list of residues that come into contact with the seed, at which point users can select the number of residues to include in the model. In the present work, the maximum number of residues suggested was used to construct the Probe-based models.

The Arpeggio method operates similarly but uses atom types, interatomic distances, and angles to classify inter-residue interactions into 15 different categories,[58] which are used by rinrus to construct a model. In some cases, the Probe and Arpeggio methods produce the same enzyme model and in either case, the result is a PDB-formatted file that can be read by Fragme∩t.

A final method for model construction uses two-body interaction energies ΔE_{IJ} to select residues, considering only those terms where either I or J represents the ligand. For definiteness, let J = ligand. A model is then created by retaining all residues I for which

$$|\Delta E_{I,\text{ligand}}| > \tau_{2B} \tag{6}$$

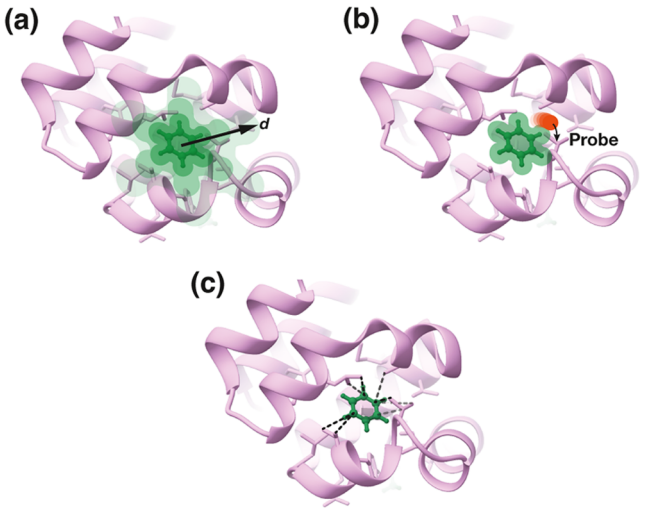


Figure 2. Illustration of methods for selecting amino acid residues around a benzene ligand (shown atomistically, in green): (a) distance-based selection using a cutoff d; (b) Probe-based selection, rolling a test sphere over the van der Waals surface of the ligand; and (c) Arpeggio selection based on atomic information.

where τ_{2B} is a user-specified threshold. MBE(3) calculations are built upon this two-body screening by including $\Delta E_{I,\text{ligand},K}$ for all residues K where $\Delta E_{I,\text{ligand}}$ satisfies eq 6. It would be relatively easy to develop an interface between rinrus and Fragme∩t, using the latter to create a ranked list of residues (according to eq 6), then reporting that information to rinrus for automated construction of QM models, but we have not done so here.

2.4. Quantum Chemistry Calculations. The primary electronic structure method used in this work is HF-3c,[20] as implemented in Q-Chem.[61] HF-3c uses a minimal-basis Hartree-Fock (HF) calculation, to which three empirical corrections are added: for dispersion, for BSSE, and for short-range basis-set incompleteness.[20] The dispersion correction is based on Grimme's D3 scheme,[62] but omits the three-body Axilrod-Teller-Muto correction that is present in conventional DFT+D3. All three corrections are parametrized for use with a specific basis set ("MINIX").[20] We will not indicate the basis set for HF-3c calculations, as it is always MINIX.

While simple and expedient, HF-3c also performs surprisingly well for large supramolecular complexes. Errors average ~4 kcal/mol for large supramolecular benchmarks,[63] such as L7[64] and S30L,[65] which is comparable to the performance of the best DFT methods.[66−69] (For molecules with $\gtrsim$100 atoms, DFT is less accurate than its performance in small van der Waals complexes would suggest.[69])

Some conventional DFT calculations are reported as well, using the ωB97X-V functional[70] as an example that performs well for small van der Waals complexes.[69] For these calculations, we use minimally augmented ("ma") versions of the standard Karlsruhe basis sets,[71−73] known as def2-ma-SVP, def2-ma-TZVP, and def2-ma-QZVP.[73] Diffuse functions are important for noncovalent DFT calculations, even when triple-ζ basis sets are employed, but minimal augmentation appears to be sufficient.[41] Our preference for the simple MBE(n) fragmentation scheme, without any kind of charge embedding, is based on a desire to use diffuse basis functions and large basis sets. Fragment-based charge embedding tends to be unstable in the presence of diffuse basis functions.[25,74−77]

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

Ionic residues inevitably arise when native protonation states are considered but are sometimes neutralized in fragment-based quantum chemistry procedures,[78−80] in order to reduce many-body effects. In contrast, our calculations use native protonation states, determined as described in Section 2.2. For systems with ionic side chains, we have documented that the use of low-dielectric boundary conditions, implemented by means of a polarizable continuum model (PCM),[81] is critical for obtaining converged results.[32] This necessity was ultimately traced to interplay between fragmentation and delocalization error,[30] and is consistent with "charge sloshing" behavior that is observed in DFT calculations for large biomolecular models.[82,83] Other studies have also noted that a PCM can improve self-consistent field (SCF) convergence behavior in DFT calculations,[84] presumably by providing a stabilizing induced charge to counterbalance charge delocalization driven by self-interaction error. As such, it is best not to view PCM as a solvation model *per se*, but rather as a simple form of dielectric boundary conditions, which are superior to vacuum boundary conditions for macromolecular electronic structure calculations.

That said, the use of a dielectric constant in the range $\varepsilon = 2-4$ to model the hydrophobic interior of a protein has a long history in biomolecular electrostatics and pK_a calculations.[85−90] (Even larger values of ε have sometimes been suggested.[90−96]) In the present work, all DFT and HF-3c calculations employ the conductor-like PCM (C-PCM),[97] with $\varepsilon = 4$. Previous work on enzyme models points to significant differences between vacuum boundary conditions ($\varepsilon = 1$) and C-PCM with $\varepsilon = 2-4$ but the effect quickly saturates for larger values of ε, especially when the enzyme model is large.[32,98−102]

C-PCM is implemented here using the switching/Gaussian procedure.[97,103−105] A molecular cavity is constructed from modified Bondi atomic radii,[106] setting $R_{atomic} = 1.2R_{Bondi}$ (per standard convention),[81] then discretized using atom-centered Lebedev grids.[103] For the n-body DFT calculations, we use 110 Lebedev points for hydrogen and 194 points for other nuclei, whereas for the HF-3c subsystem calculations and the HF-3c full-protein calculations we employ 50 Lebedev points for hydrogen and 110 for other nuclei. A conjugate gradient PCM solver was used for the full-protein calculations.[105]

For all calculations, the SCF convergence threshold was set to 10^{-8} E_h and the integral and shell-pair drop tolerances were both set to 10^{-12} a.u. The latter setting is appropriate for calculations in medium-size molecules where diffuse functions are used, as looser thresholds may engender convergence problems.[107]

3. RESULTS AND DISCUSSION

The primary goal of this work is to demonstrate that fragment-based semiempirical calculations can be used as an efficient means to test convergence of automated procedures for QM model construction in enzyme calculations. To do so, we first demonstrate that two-body HF-3c calculations provide robust convergence in large-ligand complexes (Section 3.1). We then validate the use of HF-3c against conventional DFT (Section 3.2), before considering energy screening of the two-body HF-3c calculations as a means to improve efficiency (Section 3.3). The resulting method is used to evaluate the convergence of ΔE_{int} for several binding-site models (Section 3.4), and comparisons to DFT results are presented in Section 3.5.

3.1. Convergence Tests. To illustrate that two-body HF-3c calculations provide robust convergence even for large-ligand models, we briefly recapitulate some results from our previous work establishing convergent fragment-based protocols for ligand−protein interaction energies.[15] Figure 3 shows

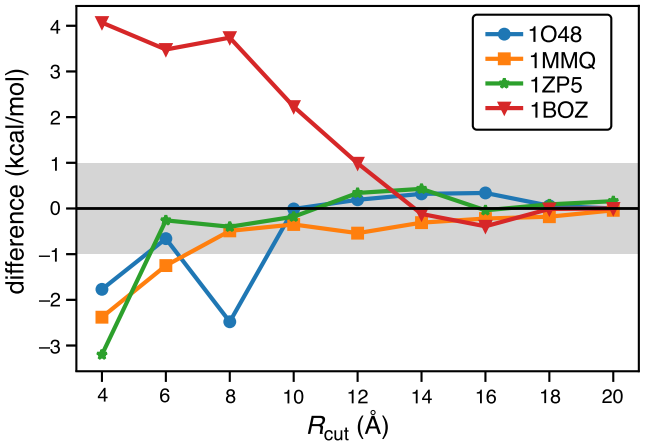


Figure 3. Difference in ΔE_{int} for large-inhibitor complexes as a function of a cutoff distance (R_{cut}), based on MBE(2) calculations at the HF-3c level. The baseline calculation is MBE(2) for the full-protein−ligand complex, and the shaded region indicates ±1 kcal/mol with respect to that baseline. Reprinted with permission from ref 15; copyright 2025 American Chemical Society.

how MBE(2) calculations at the HF-3c level converge as a function of a cutoff radius R_{cut}, for four different protein−ligand complexes. The four ligands in question are large but otherwise chemically distinct (see Figure 1), and the enzyme is different in each case. Nevertheless, the convergence behavior is remarkably similar. The largest complex is 1BOZ at 3124 atoms, and results in this case are converged by R_{cut} = 14 Å. The other three complexes range in size from 1781 atoms (1O48) to 2637 atoms (1MMQ), and each converges by R_{cut} = 10−12 Å.

3.2. Comparing HF-3c to DFT. We next consider two- and three-body contributions to ΔE_{int}, meaning $\Delta E_{I,ligand}$ and $\Delta E_{IJ,ligand}$, comparing values computed at the HF-3c and DFT levels. The DFT calculations are performed in basis sets up to quadruple-ζ quality. Correlations between the two methods are illustrated in Figure 4 for one particular protein−ligand complex (181L), and analogous plots for other complexes can be found in Figures S1−S3.

Correlation between HF-3c and ωB97X-V is quite good for the two-body terms (Figure 4a), and there is a clear separation between energetically important terms ($|\Delta E_{IJ}| > 10^{-3}$ E_h) and those that are very nearly zero. Linear fits to the data in Figure 4a afford slopes of 1.12, 0.85, and 0.81 for the def2-ma-SVP, def2-ma-TZVP, and def2-ma-QZVP basis sets, respectively, with $R^2 \geq 0.9$ in each case. (Results are similar for the other systems and best-fit parameters can be found in Table S1.) A slope greater than unity implies that ΔE_{IJ} is more attractive at the ωB97X-V level as compared to HF-3c. In three of four examples, this happens only for the def2-ma-SVP basis set while other slopes are less than unity. In the remaining case (1O48), the slope is closest to unity for def2-ma-SVP and smaller in the more complete basis sets (Table S1). All of this behavior is indicative of significant BSSE in the double-ζ calculations. Close agreement between triple- and quadruple-ζ values for the two-body corrections suggests that the BSSE is largely eliminated using def2-ma-TZVP, which is typical for small fragments.[41]

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

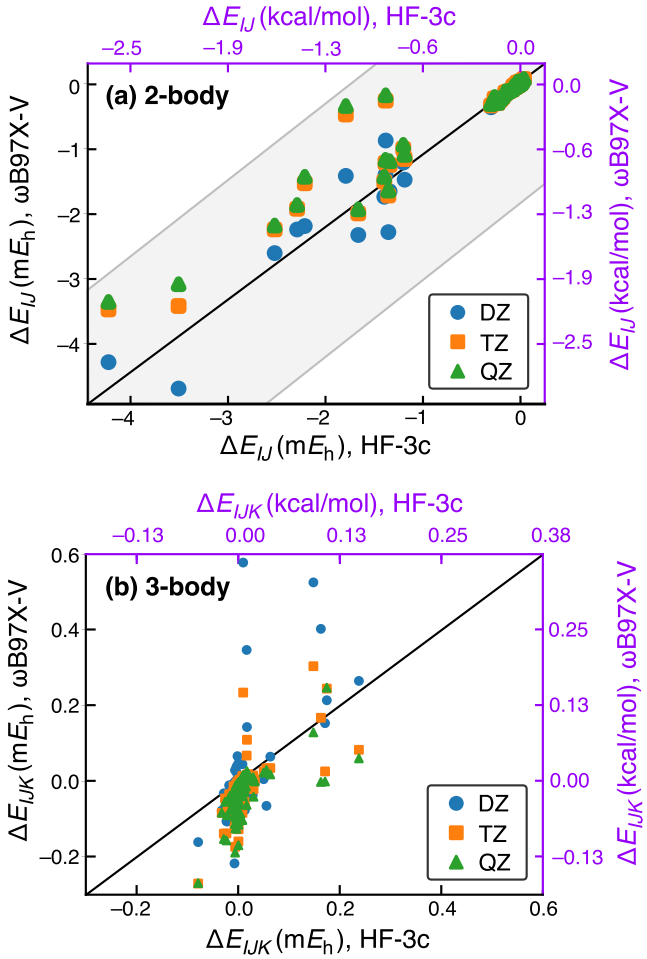


Figure 4. Correlations between (a) ΔE_{IJ} and (b) ΔE_{IJK} for the complex 181L, comparing HF-3c and ωB97X-V results, with the latter evaluated in several different basis sets. A millihartree scale (in black) is consistent with the units used for the τ_{2B} criterion in eq 6, but a kcal/mol scale (in purple) is also shown. Diagonal lines indicate where the value is the same for both methods, and the gray area in (a) represents ±1 kcal/mol difference. Only terms that involve the ligand (benzene) are plotted, using R_{cut} = 8 Å for the three-body terms. For the ωB97X-V calculations, the basis sets are def2-ma-SVP (labeled "DZ"), def2-ma-TZVP ("TZ"), and def2-ma-QZVP ("QZ").

Correlations between HF-3c and ωB97X-V are much less pronounced for the three-body terms (Figure 4b), with $R^2 \approx 0.4$. The lone exception to this trend is that HF-3c and ωB97X-V values of ΔE_{IJK} correlate very well for 1O48, with $R^2 = 0.93$ for HF-3c versus ωB97X-V/def2-ma-QZVP, for example. We regard this as a coincidence as it is not borne out in the other three systems considered. Three-body terms may not be reliably captured using HF-3c due to the minimal-basis set, since polarization is the most important three-body contribution,[25] although it is also possible that the three-body interactions are exaggerated by ωB97X-V calculations in a double-ζ basis set.[30] Setting aside the ωB97X-V/def2-ma-SVP results in Figure 4b, which are significantly impacted by BSSE, it does appear that HF-3c can at least identify the small number of three-body terms whose magnitude is significant.

The remaining analysis focuses on two-body interactions because MBE(2) can be used for rapid screening and to evaluate convergence of binding-site models. Figure 5 provides a closer look at two-body terms computed at the HF-3c level, organized into histograms for each of four protein−ligand complexes. These histograms include only those terms $|\Delta E_{I,\text{ligand}}|$, meaning that one of the fragments is the ligand. Each distribution in Figure 5 is asymmetric about zero. Additionally, there does not seem to be a single energy threshold that would be viable across all four of these systems, as the energy scale for $|\Delta E_{I,\text{ligand}}|$ is rather different in each of the four examples.

3.3. Selecting τ_{2B}. We next consider the construction of enzyme models via the two-body energy criterion in eq 6. Table 1 compares errors for MBE(2) and MBE(3) approximations for models generated in this way. All calculations were performed at the HF-3c level and the error is defined with respect to a full-system calculation performed at the same level. Timings for the full-system calculations can be found in Table 2.

With the exception of 1O48, the MBE(2) method affords sub-kcal/mol fidelity with respect to a full-protein calculation, if the model is constructed using a sufficiently small value of the two-body energy threshold τ_{2B} in eq 6. Even for 1O48, sub-kcal/mol fidelity is achievable but in that case it requires MBE(3), and comes with a significant increase in cost. For 1O48, MBE(3) is consistently and significantly more accurate than MBE(2) but for the other three systems, MBE(2) and MBE(3) results are typically within ∼1 kcal/mol of one another.

For high-fidelity calculations, the best choice appears to be $\tau_{2B} = 2.5 \times 10^{-4}$ E_h, for both MBE(2) and MBE(3). The tighter value $\tau_{2B} = 1.25 \times 10^{-4}$ E_h produces larger models, for which the results are actually marginally worse in a few cases, as judged by comparison to ΔE_{int} computed using the full protein. This indicates that convergence of ΔE_{int} need not be monotonic (to the full supramolecular result) with increasing model size, and that there is some interplay between the model size and the order of the n-body expansion. Larger models may introduce noise, stemming from finite-precision issues,[15,31,108,109] while including less relevant residues that do not contribute meaningfully to the accuracy. Conversely, a smaller but well-chosen model can focus on the most energetically significant interactions, leading to more accurate predictions for ΔE_{int} at lower cost. In MBE(n) calculations, one should not assume that larger models are always more faithful to the full-system result, except possibly in very small models.

As we refine these models, it is also crucial to consider how we evaluate their performance, particularly in terms of error reporting. Standard practice in fragment-based quantum chemistry calculations is to report errors on a per-monomer basis. For applications of MBE(n) to water clusters, a target accuracy of 0.1 kcal/mol/monomer has been suggested,[46] representing 10% of k_BT at T = 298 K. The idea is that fragmentation errors of this magnitude are indistinguishable from thermal noise. Models with $\tau_{2B} = 2.5 \times 10^{-4}$ E_h do achieve this level of accuracy, although the 0.1 × k_BT standard is probably unnecessarily stringent for macromolecular ΔE_{int} calculations. Even with the best conventional density functionals such as ωB97X-V, the disparity between single-pose ΔE_{int} calculations (or even ensemble-averaged values, $\langle \Delta E_{int} \rangle$) and experimental binding affinities ΔG°_{bind} is many times larger than $0.1k_BT$. (For a lengthy discussion of this point, see our recent work on fragmentation protocols for protein−ligand interaction energies.[15]) In addition, it is important to recognize the intrinsic limitations of semiempirical quantum chemistry,

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

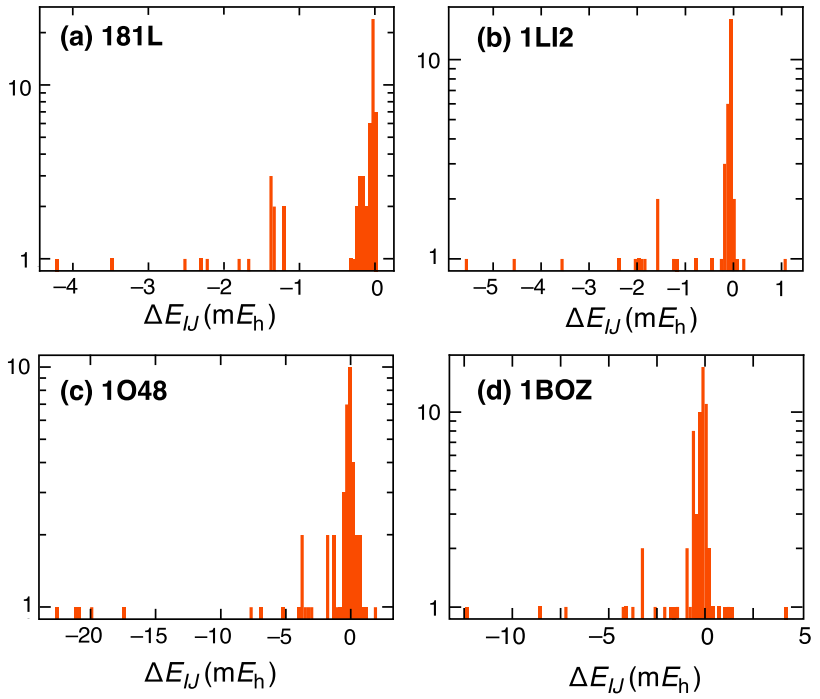


Figure 5. Histograms of the two-body terms ΔE_{IJ} where either I or J represents the ligand, for protein−ligand complexes (a) 181L, (b) 1LI2, (c) 1O48, and (d) 1BOZ. All calculations were performed using HF-3c.

Table 1. Errors in ΔE_{int} for HF-3c Calculations on Enzyme Models Constructed Based on Two-Body Energies

			MBE(2)			MBE(3)		
			error (kcal/mol)[a]			error (kcal/mol)[a]		
system	$\tau_{2B}/10^{-4}\ E_h$	no. atoms	absolute	per-monomer	time[b] (h)	absolute	per-monomer	time[b] (h)
181L	10.0	266	3.10	0.22	1	3.32	0.24	10
	5.0	284	2.23	0.15	1	2.43	0.16	10
	2.5	360	1.58	0.08	2	1.91	0.10	21
	1.25	451	0.92	0.04	3	1.20	0.05	47
1LI2	10.0	263	1.10	0.08	1	1.22	0.09	10
	5.0	285	0.59	0.04	1	0.70	0.04	10
	2.5	333	0.36	0.02	1	0.61	0.03	18
	1.25	572	0.08	0.00	4	0.22	0.01	91
1O48	10.0	494	4.20	0.17	6	0.84	0.03	98
	5.0	644	3.92	0.12	10	0.41	0.01	246
	2.5	797	4.32	0.10	15	0.04	0.00	457
	1.25	991	4.47	0.08	22	0.00	0.00	891
1BOZ	10.0	439	3.05	0.15	5	4.20	0.21	77
	5.0	737	0.59	0.02	10	1.10	0.03	246
	2.5	1145	0.57	0.01	22	0.49	0.01	892
	1.25	1637	1.85	0.02	45	2.85	0.03	3512

[a]Error is defined with respect to a full-system HF-3c calculation. [b]Total time (aggregated across processors) on hardware described in Section 2.1.

Table 2. Full-System HF-3c Interaction Energies and Timings, without Fragmentation

system	no. atoms	ΔE_{int} (kcal/mol)	time[a] (h)
181L	2636	−19.4	4156
1LI2	2637	−18.8	5542
1O48	1781	−89.9	854
1BOZ	3124	−31.3	5018

[a]Supersystem calculations were performed using a single 48-core node (Intel Xeon Platinum 8268).

as there is no sense in pushing for higher fidelity than is warranted by the intrinsic accuracy of the electronic structure method. As noted in Section 2.4, errors in HF-3c interaction energies average 4 kcal/mol for standard test sets of large supramolecular complexes.[63]

One of the primary reasons to complete these calculations using fragmentation is the significant reduction in the cost per calculation. Given sufficient hardware, the wall-time cost of fragment-based calculations can be made very small because the subsystem calculations are inherently distributable. However, we are more interested in the extent to which the total (aggregate) computing time can be reduced via

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

Table 3. Errors in ΔE_{int} for MBE(n) Calculations at the HF-3c Level for Various Enzyme Models[a]

			MBE(2)			MBE(3)		
			error (kcal/mol)[b]			error (kcal/mol)[b]		
system	model	no. atoms	absolute	per-monomer	time[c] (h)	absolute	per-monomer	time[c] (h)
181L	d = 4 Å	243	4.15	0.32	1	4.11	0.32	5
	d = 6 Å	452	1.26	0.05	2	1.61	0.07	19
	Probe	221	4.90	0.41	1	4.87	0.41	4
	Arpeggio	243	4.15	0.32	1	4.11	0.32	5
1LI2	d = 4 Å	244	2.31	0.17	1	2.31	0.17	6
	d = 6 Å	475	0.32	0.01	2	0.01	0.00	22
	Probe	222	3.59	0.33	1	3.59	0.33	3
	Arpeggio	247	1.33	0.10	1	1.42	0.10	7
1O48	d = 4 Å	420	3.29	0.16	3	0.82	0.04	21
	d = 6 Å	623	2.42	0.08	5	1.15	0.04	44
	Probe	383	2.55	0.14	2	1.70	0.09	17
	Arpeggio	449	2.48	0.11	3	1.88	0.09	24
1BOZ	d = 4 Å	467	3.34	0.15	4	1.53	0.07	28
	d = 6 Å	947	3.95	0.08	10	2.53	0.05	109
	Probe	371	0.52	0.03	2	0.94	0.06	14
	Arpeggio	476	2.86	0.12	4	1.47	0.06	31

[a]MBE(n) calculations use R_{cut} = 8 Å. [b]Error with respect to a full-system HF-3c calculation. [c]Total time (aggregated across processors) on hardware described in Section 2.1.

fragmentation. Aggregate computing time is a better metric for evaluating the cost because it reflects the carbon footprint of a given calculation, whereas wall time is a selfish time-to-solution metric.[25,110] Table 1 provides the aggregate computing time for the HF-3c MBE(n) calculations and Table 2 provides the same data for the supersystem HF-3c calculations. The latter were performed on a single compute node so they do not suffer from the low parallel efficiencies that typically characterizes massively parallel electronic structure calculations.[110]

Even so, the cost reduction is significant for the MBE(2) calculations, whose cost is no more than 1−2% of the cost of the corresponding supersystem calculation. For the largest system considered here (1BOZ, with 3124 atoms), and for the model constructed using $\tau_{2B} = 2.5 \times 10^{-4}$ E_h, the MBE(2) calculation requires 22 h or 0.4% of the conventional HF-3c cost, while MBE(3) requires 892 h or 18% of the unfragmented cost. For 1O48 (with 1781 atoms), the cost of the $\tau_{2B} = 2.5 \times 10^{-4}$ E_h model is 2% of the supersystem cost for MBE(2) or 54% for MBE(3). Thus, fragmentation dramatically reduces the cost even for low-scaling methods like HF-3c that are already affordable in large systems. This presents a compelling advantage for high-throughput screening of different model-building algorithms, which is the topic of the next section.

3.4. Comparison of Enzyme Models. Having established that two-body energy screening is a viable means to construct binding-site models, we next consider the application of MBE(n) to models constructed in other ways, either using a simple distance criterion or else by means of the RINRUS code (as described in Section 2.3). Table 3 lists errors in MBE(2) and MBE(3) values of ΔE_{int} for various models, with all calculations performed at the HF-3c level.

We examined distance-based models ranging from d = 2.5 Å to d = 10.0 Å but only the 4 and 6 Å models are listed in Table 3, as these were judged to provide reasonable accuracy while also affording models that are comparable in size to those obtained in other ways. As we saw with the τ_{2B} models in Section 3.3, increasing d (to increase model size) improves the MBE(2) and MBE(3) accuracy only up to a point; errors eventually reach a plateau where larger models do not improve the results, as compared to a value of ΔE_{int} computed without fragmentation. For some systems, that plateau is reached at d = 5 Å while for other complexes the fragmentation errors continue to decrease until the model reaches d = 8 Å. Errors for models ranging from 2.5−10.0 Å can be found in Tables S2−S5.

The best-performing model in Table 3, according to both MBE(2) and MBE(3) calculations, is the d = 6 Å model. This construction also affords the largest model for each of the four protein−ligand complexes that we consider, and includes residues that were not picked up in the Probe or Arpeggio constructions or even by the τ_{2B} criterion. At the same time, a strictly distance-based construction almost certainly includes unimportant residues, leading to systematically larger models. For the T4-lysozymes 181L and 1LI2, the 6 Å model affords a smaller error at the MBE(2) level as compared to the τ_{2B}-derived model, but for those systems the ligand is much smaller and the binding site is more compact as compared to that in 1O48 or 1BOZ. Convergence to the supermolecular value of ΔE_{int} is slower for the latter two systems.

Models generated using the Probe and Arpeggio methods are generally similar to one another although the latter approach incorporates a slightly larger number of residues. (In each case considered here, residues selected by the Probe method are a subset of those selected using the Arpeggio construction.) The Arpeggio models are not significantly or consistently more accurate, however. Interestingly, for the large-ligand systems 1O48 and 1BOZ, the Probe and Arpeggio models typically have a lower absolute error as compared to the distance-based models, and they are competitive with the τ_{2B} models. In our view, this is a consequence of the ligand size. The ligand was the seed used to generate these models (see Section 2.3), and a larger ligand engenders more close contacts that are captured in a comprehensive way by the Probe and Arpeggio constructions.

Those methods do not outperform the distance-based models for the T4-lysozymes (181L and 1LI2), however. Although these are less realistic examples for noncovalent

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

inhibition, they are potential examples of fragment-based approaches to drug discovery,[111−119] which target interaction energies between small functional-group moieties and an enzyme binding site. There are several avenues that could be used to improve these model-building procedures, including the addition of a second interaction sphere or the use of a larger seed that contains some nearest-neighbor residues. The latter could be selected using a distance cutoff or based on *a priori* knowledge of relevant interactions. (For example, if investigating a mechanism then important nucleophiles or electrophiles could be included.) In the present work, we chose to examine only "naive" models that require no such *a priori* information. A recent development in RINRUS is an option to use a form of pairwise symmetry-adapted perturbation theory (SAPT),[120] as a means to decompose the interaction energy between a protein and individual residue main chains or side chains.[13] More cost-effective forms of SAPT, scaling as $O(N^3)$ rather than $O(N^5)$, could also be used for this purpose.[66−68,121] In any case, it is important to point out that the Probe, Arpeggio, and τ_{2B} models are generated in a reproducible manner, making them particularly well-suited for development of drug discovery workflows.

Models discussed in this section are generally smaller than the τ_{2B} models described in Section 3.3, which impacts both the computational load and the time required for processing. This is clearly reflected in the timing data in Table 3, where MBE(2) calculations using the 6 Å model require 10 h for the largest protein−ligand complex as compared to 22 h for our preferred τ_{2B}-derived model.

Figure 6 compares MBE(3) errors across the data set, using three different paradigms to construct the binding-site model:

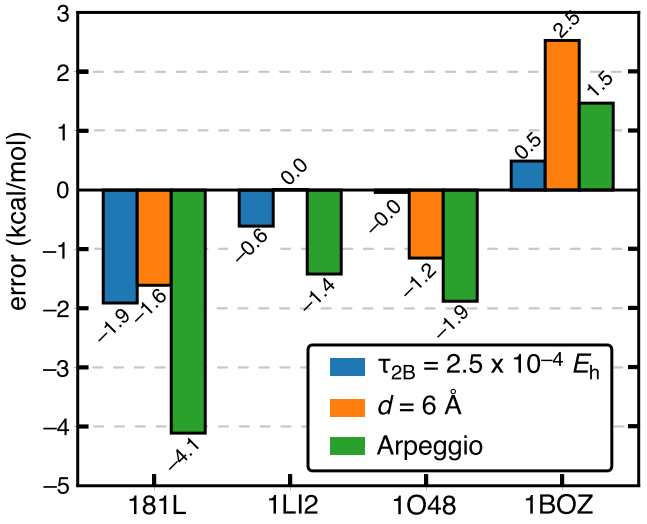


Figure 6. Errors in ΔE_{int} for MBE(3) calculations at the HF-3c level, comparing three different methods to construct a binding-site model for four different protein−ligand complexes.

eq 6 with $\tau_{2B} = 2.5 \times 10^{-4}$ E_h, a $d = 6$ Å model, and finally an Arpeggio model obtained using RINRUS. For three of the four protein−ligand complexes, all of these models overestimate the interaction strength whereas for 1BOZ they all underestimate it, suggesting there may be an enzyme size-related bias that is common to all three algorithms. None of these three procedures consistently outperforms the others but the τ_{2B} approach stands out as the most reliable overall, with a mean absolute fragmentation error of 0.8 kcal/mol for MBE(3) calculations using HF-3c. Furthermore, the $\tau_{2B} = 2.5 \times 10^{-4}$ E_h method also affords the smallest fragmentation error for MBE(2), which is 1.7 kcal/mol when averaged over the four complexes. However, the 6 Å model is only slightly less accurate on average, and more accurate in two out of four complexes. It is also considerably less expensive.

The 6 Å model contains several unique residues that do not appear in any of the τ_{2B} models: three such residues in 181L, eight in 1LI2, one in 1O48, and 11 in 1BOZ. For the distance-based models, the addition of residues need not reflect consequences for ΔE_{int}, so it is more interesting to compare residues that are unique to the energy-based models instead. The most significant differences between these two algorithms are found in 1O48 and 1BOZ, where the number and identity of residues varies greatly. For example, for 1O48 the 6 Å model contains only one unique residue (Leu45) but the model constructed using $\tau_{2B} = 2.5 \times 10^{-4}$ E_h includes 15 additional residues, ten of which are charged with a total combined charge of +5. This discrepancy manifests as a 2 kcal/mol difference in errors at the MBE(2) level, illustrating how the precise choice of residues can significantly impact the result, and furthermore demonstrating that larger models do not always lead to smaller errors.

3.5. DFT and Basis-Set Convergence. To ground the performance of MBE(2) for HF-3c in terms of more conventional quantum chemistry, we next examine the performance of various QM models when ΔE_{int} is computed using the ωB97X-V functional, in basis sets ranging from def2-ma-SVP to def2-ma-QZVP. Supersystem calculations at the ωB97X-V/def2-ma-QZVP level exceed our computational resources so instead we examine MBE(2) results that include all residues up to $R_{cut} = 8$ Å. The resulting interaction energies are provided in Table 4, comparing ωB97X-V (in various basis sets) to HF-3c. These data come from previous work,[15] where we established that ωB97X-V/def2-ma-SVP predicts stronger interaction energies than HF-3c whereas ωB97X-V with triple- and quadruple-ζ basis sets predicts weaker interactions. We also know that BSSE can be quite large for sizable protein−ligand models, especially in double-ζ basis sets,[41] and the basis-set convergence in Table 4 provides some measure of the BSSE. For the largest system considered here (1BOZ), the ωB97X-V/def2-ma-SVP and ωB97X-V/def2-ma-QZVP interaction energies differ by 23 kcal/mol.

Table 4. Interaction Energies Computed using MBE(2) with $R_{cut} = 8$ Å[a]

	ΔE_{int} (kcal/mol)			
		ωB97X-V		
system	HF-3c	DZ[b]	TZ[c]	QZ[d]
181L	−19.1	−21.0	−16.3	−15.4
1LI2	−19.8	−23.1	−18.0	−16.8
1O48	−93.7	−101.6	−82.5	−80.6
1BOZ	−36.8	−53.6	−34.1	−30.4

[a]From ref 15. [b]def2-ma-SVP. [c]def2-ma-TZVP. [d]def2-ma-QZVP.

We next use the MBE(2) interaction energies in Table 4 as benchmarks for MBE(2) applied to smaller QM models, using ωB97X-V in basis sets through def2-ma-QZVP. Errors in ΔE_{int}, relative to MBE(2) with $R_{cut} = 8$ Å, are listed in Table 5 for the best-performing model systems, as determined in Sections 3.3

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

Table 5. Errors in ΔE_{int} for MBE(2) Calculations Using ωB97X-V[a]

system	model	error (kcal/mol)[b] DZ[c]	TZ[d]	QZ[e]
181L	$\tau_{2B} = 2.5 \times 10^{-4}$ E_h	1.2	1.1	1.4
	d = 6 Å	0.9	0.9	0.8
	Arpeggio	4.4	4.0	3.8
1LI2	$\tau_{2B} = 2.5 \times 10^{-4}$ E_h	1.0	1.0	0.9
	d = 6 Å	0.6	0.5	0.4
	Arpeggio	2.6	2.3	2.2
1O48	$\tau_{2B} = 2.5 \times 10^{-4}$ E_h	0.2	0.2	0.2
	d = 6 Å	1.5	0.1	1.4
	Arpeggio	1.1	0.9	0.8
1BOZ	$\tau_{2B} = 2.5 \times 10^{-4}$ E_h	4.2	2.0	1.2
	d = 6 Å	2.4	3.7	1.5
	Arpeggio	4.4	3.5	2.8

[a]MBE(2) calculations use R_{cut} = 8 Å. [b]Error is measured with respect to the full-system (unfragmented) calculation at the same level of theory. [c]def2-ma-SVP. [d]def2-ma-TZVP. [e]def2-ma-QZVP.

and 3.4. For the small-ligand complexes 181L and 1LI2, the d = 6 Å model tends to exhibit the smallest errors, although errors for the $\tau_{2B} = 2.5 \times 10^{-4}$ E_h model are larger only by about 0.5 kcal/mol. For the large-ligand complexes 1O48 and 1BOZ, the τ_{2B} model is the most accurate one except in one case, namely, 1BOZ at the ωB97X-V/def2-ma-SVP level. Those results are probably significantly impacted by BSSE, since 1BOZ is the largest system considered here. In almost every case, fragmentation errors are smaller in the triple-ζ basis set as compared to the double-ζ one, with the 6 Å model of 1BOZ as the lone exception. MBE(2) errors at the ωB97X-V/def2-ma-QZVP level are all $\leq$1.5 kcal/mol for the $\tau_{2B} = 2.5 \times 10^{-4}$ E_h and the d = 6 Å models.

4. CONCLUSIONS

This work extends other recent work from our group,[15,32] whose goal is to develop automated methods for reliable and affordable QM calculations in enzymatic systems. Fragmentation offers significant advantages for calculating protein−ligand interaction energies in sizable binding-site models, and renders such calculations accessible to workstation-level computing resources. The open-source FRAGME∩T code[39] is a practical and immediate solution that makes accurate QM calculations available to a wide range of researchers who may not have access to supercomputer resources.

For protein−ligand systems, we have demonstrated that two-body interaction terms ΔE_{IJ}, computed using the semiempirical HF-3c method,[20] correlate very well with results from high-quality DFT calculations, exemplified by ωB97X-V/def2-ma-QZVP. The two-body terms vary significantly in both magnitude and sign, and provide a means to generate QM models in a well-defined way. A threshold $\tau_{2B} = 2.5 \times 10^{-4}$ E_h offers a good balance between accuracy and computational efficiency. Of the model-construction algorithms examined here, this is the most reliable one and does not require *a priori* biochemical information. Moreover, because HF-3c requires only a minimal-basis HF calculation plus analytic corrections along the lines of Grimme's D3 dispersion model,[20,62] it should be easy to implement in any modern quantum chemistry code that contains DFT + D3 with hybrid functionals.

Energy-based model construction typically results in larger models as compared to algorithms implemented in the RINRUS program;[11,13,60] nevertheless, the τ_{2B} models consistently deliver higher accuracy. Simple distance-based procedures with a 6 Å cutoff are also found to be effective. Models based on the Probe and Arpeggio functionality in RINRUS could be further improved by including a coordination sphere in the seed moiety, but this would require users to know which residues are relevant. Alternatively, two-body semiempirical calculations are affordable enough to be incorporated into model-building workflows and require no *a priori* information beyond a crystal structure.

For MBE(2)-DFT calculations with ωB97X-V, the best QM models constructed in this manner achieve a fidelity of 1−2 kcal/mol in triple- and quadruple-ζ basis sets, as compared to MBE(2)-DFT calculations on larger, converged models of the protein. The combination of fast semiempirical MBE(2) calculations, used to test convergence of ΔE_{int} with respect to model size, and convergent MBE(n) protocols for evaluating ΔE_{int},[15] represents a powerful tool chain for quantum-chemical studies of drug−protein interactions. The same semiempirical model-building and convergence tests should also be useful for studies of enzyme thermochemistry and kinetics, for which we have also reported convergent MBE(n) protocols.[32]

■ ASSOCIATED CONTENT

Data Availability Statement

All calculations were performed using the open-source FRAGME∩T code.[39] In the present work, FRAGME∩T is interfaced with Q-CHEM,[40] although other electronic structure engines can also be used. A trial license for Q-CHEM can be obtained from https://www.q-chem.com/try. Other software used to prepare protein models is available at the URLs indicated in the references, including H++,[51] PYMOL,[53] RINRUS,[60] and GFN2-xTB.[122] Coordinates for the protein−ligand complexes are provided in the Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c01987.

Additional data for fragmentation calculations (PDF)

List of residues included in the QM models (PDF)

Coordinates for the protein−ligand complexes (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

John M. Herbert − *Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210, United States; 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 − *Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210, United States; 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*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c01987

Author Contributions

P.E.B. and J.M.H. designed the research; D.R.B. and P.E.B. wrote the software; P.E.B. performed the calculations and analyzed the data; P.E.B. and J.M.H. wrote the manuscript. All authors read and approved the final manuscript.

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

Work by P.E.B. on protein−ligand interactions was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award No. 1R43GM148095-01A1. Development of the FRAGME∩T software (by D.R.B.) 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.[123]

■ REFERENCES

(1) Hu, L.; Eliasson, J.; Heimdal, J.; Ryde, U. Do quantum mechanical energies calculated for small models of protein-active sites converge? *J. Phys. Chem. A* **2009**, *113*, 11793−11800.
(2) Hu, L.; Söderhjelm, P.; Ryde, U. On the convergence of QM/MM energies. *J. Chem. Theory Comput.* **2011**, *7*, 761−777.
(3) Hu, L.; Söderhjelm, P.; Ryde, U. Accurate reaction energies in proteins obtained by combining QM/MM and large QM calculations. *J. Chem. Theory Comput.* **2013**, *9*, 640−649.
(4) Sumner, S.; Söderhjelm, P.; Ryde, U. Effect of geometry optimizations on QM-cluster and QM/MM studies of reaction energies in proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4205−4214.
(5) Liao, R.-Z.; Thiel, W. Comparison of QM-only and QM/MM models for the mechanism of tungsten-dependent acetylene hydratase. *J. Chem. Theory Comput.* **2012**, *8*, 3793−3803.
(6) Liao, R.-Z.; Thiel, W. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *J. Comput. Chem.* **2013**, *34*, 2389−2397.
(7) 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.
(8) Karelina, M.; Kulik, H. J. Systematic quantum mechanical region determination in QM/MM simulation. *J. Chem. Theory Comput.* **2017**, *13*, 563−576.
(9) 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.
(10) Yang, Z.; Mehmood, R.; Wang, M.; Qi, H. W.; Steeves, A. H.; Kulik, H. J. Revealing quantum mechanical effects in enzyme catalysis with large-scale electronic structure simulation. *React. Chem. Eng.* **2019**, *4*, 298−315.
(11) Summers, T. J.; Cheng, Q.; Palma, M. A.; Pham, D.-T.; Kelso III, D. K.; Webster, C. E.; DeYonker, N. J. Cheminformatic quantum mechanical enzyme model design: A catechol-O-methyltransferase case study. *Biophys. J.* **2021**, *120*, 3577−3587.
(12) Cheng, Q.; DeYonker, N. J. The glycine N-methyltransferase case study: Another challenge for QM-cluster models? *J. Phys. Chem. B* **2023**, *127*, 9282−9294.
(13) Agbaglo, D. A.; Summers, T. J.; Cheng, Q.; DeYonker, N. J. The influence of model building schemes and molecular dynamics on QM-cluster models: The chorismate mutase case study. *Phys. Chem. Chem. Phys.* **2024**, *26*, 12467−12482.
(14) Demapan, D.; Kussmann, J.; Ochsenfeld, C.; Cui, Q. Factors that determine the variation of equilibrium and kinetic properties of QM/MM enzyme simulations: QM region, conformation, and boundary condition. *J. Chem. Theory Comput.* **2022**, *18*, 2530−2542.
(15) 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.* **2024**, in press.
(16) Summers, T. J.; Daniel, B. P.; Cheng, Q.; DeYonker, N. J. Quantifying inter-residue contact through interaction energies. *J. Chem. Inf. Model.* **2019**, *59*, 5034−5044.
(17) Cheng, Q.; DeYonker, N. J. A case study of the glycoside hydrolase enzyme mechanism using an automated QM-cluster model building toolkit. *Front. Chem.* **2022**, *10*, 854318.
(18) 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.
(19) Csizi, K.-S.; Reiher, M. Universal QM/MM approaches for general nanoscale applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2023**, *13*, e1656.
(20) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(21) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.* **2012**, *112*, 632−672.
(22) Collins, M. A.; Bettens, R. P. Energy-based molecular fragmentation methods. *Chem. Rev.* **2015**, *115*, 5607−5642.
(23) Raghavachari, K.; Saha, A. Accurate composite and fragment-based quantum chemical methods for large molecules. *Chem. Rev.* **2015**, *115*, 5643−5677.
(24) *Fragmentation: Toward Accurate Calculations on Complex Molecular Systems*; Gordon, M. S., Ed.; John Wiley & Sons: Hoboken, 2017.
(25) Herbert, J. M. Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.* **2019**, *151*, 170901.
(26) Li, W.; Dong, H.; Ma, J.; Li, S. Structures and spectroscopic properties of large molecules and condensed-phase systems predicted by generalized energy-based fragmentation approach. *Acc. Chem. Res.* **2021**, *54*, 169−181.
(27) Liu, J.; He, X. Recent advances in quantum fragmentation approaches to complex molecular and condensed-phase systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2023**, *13*, e1650.
(28) 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.
(29) 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.
(30) Broderick, D. R.; Herbert, J. M. Delocalization error poisons the density-functional many-body expansion. *Chem. Sci.* **2024**, *15*, 19893−19906.
(31) 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.
(32) 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.
(33) Liu, J.; Herbert, J. M. Pair−pair approximation to the generalized many-body expansion: An efficient and accurate alternative to the four-body expansion, with applications to *ab initio* protein energetics. *J. Chem. Theory Comput.* **2016**, *12*, 572−584.
(34) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent radii revisited. *Dalton Trans.* **2008**, 2832−2838.
(35) Lin, H.; Truhlar, D. G. QM/MM: What have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2007**, *117*, 185−199.
(36) He, X.; Zhu, T.; Wang, X. W.; Liu, J. F.; Zhang, J. Z. H. Fragment quantum mechanical calculation of proteins and its applications. *Acc. Chem. Res.* **2014**, *47*, 2748−2757.

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

(37) Vornweg, J. R.; Wolter, M.; Jacob, C. R. A simple and consistent quantum-chemical fragmentation scheme for proteins that includes two-body contributions. *J. Comput. Chem.* **2023**, *44*, 1634−1644.
(38) Vornweg, J. R.; Jacob, C. Protein-ligand interaction energies from quantum-chemical fragmentation methods: Upgrading the MFCC-scheme with many-body contributions. *J. Phys. Chem. B* **2024**, *128*, 11597−11606.
(39) Broderick, D. R.; Bowling, P. E.; Shockey, J.; Higley, J.; Dickerson, H.; Ahmed, S.; Herbert, J. M. Fragme∩t, an open-source framework for fragment-based quantum chemistry calculations (https://gitlab.com/fragment-qc/fragment).
(40) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; 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.
(41) 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.
(42) 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.
(43) 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.
(44) Liu, K.-Y.; Herbert, J. M. Understanding the many-body expansion for large systems. III. Critical role of four-body terms, counterpoise corrections, and cutoffs. *J. Chem. Phys.* **2017**, *147*, 161729.
(45) Hirata, S.; Valiev, M.; Dupuis, M.; Xantheas, S. S.; Sugiki, S.; Sekino, H. Fast electron correlation methods for molecular clusters in the ground and excited states. *Mol. Phys.* **2005**, *103*, 2255−2265.
(46) Ouyang, J. F.; Bettens, R. P. A. Many-body basis set superposition effect. *J. Chem. Theory Comput.* **2015**, *11*, 5132−5143.
(47) Richard, R. M.; Bakr, B. W.; Sherrill, C. D. Understanding the many-body basis set superposition error: Beyond Boys and Bernardi. *J. Chem. Theory Comput.* **2018**, *14*, 2386−2400.
(48) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005**, *48*, 3714−3728.
(49) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. A model binding site for testing scoring functions in molecular docking. *J. Mol. Biol.* **2002**, *322*, 339−355.
(50) Mobley, D. L.; Gilson, M. K. Predicting binding free energies: Frontiers and benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531−558.
(51) H++ web server. http://newbiophysics.cs.vt.edu/H++ (accessed Dec 02, 2024).
(52) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulation. *Nucleic Acids Res.* **2012**, *40*, 537−541.
(53) The PyMOL molecular graphics system, v. 2.1 (Schrödinger, LLC). https://pymol.org (accessed Dec 02, 2024).
(54) 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.
(55) 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.
(56) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and quantifying molecular goodness-of-fit: Small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **1999**, *285*, 1711−1733.
(57) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747.
(58) Jubb, H. C.; Higueruelo, A. P.; Ochoa-Montaño, B.; Pitt, W. R.; Ascher, D. B.; Blundell, T. L. Arpeggio: A web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* **2017**, *429*, 365−371.
(59) Arpeggio, a webserver for calculating interatomic interactions in protein structures. https://biosig.lab.uq.edu.au/arpeggioweb (accessed Dec 02, 2024).
(60) Residue Network Interaction Residue Selector. https://github.com/natedey/RINRUS (accessed Dec 02, 2024).
(61) Liu, J.; Rana, B.; Liu, K.-Y.; Herbert, J. M. Variational formulation of the generalized many-body expansion with self-consistent embedding charges: Simple and correct analytic energy gradient for fragment-based ab initio molecular dynamics. *J. Phys. Chem. Lett.* **2019**, *10*, 3877−3886.
(62) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parameterization of density functional dispersion correction (DFT-D) for the 94 elements H−Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(63) Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S. Low-cost quantum chemical methods for noncovalent interactions. *J. Phys. Chem. Lett.* **2014**, *5*, 4275−4284.
(64) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. Accuracy of quantum chemical methods for large noncovalent complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364−3374.
(65) Sure, R.; Grimme, S. Comprehensive benchmark of association (free) energies of realistic host−guest complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3785−3801.
(66) Lao, K. U.; Herbert, J. M. Atomic orbital implementation of extended symmetry-adapted perturbation theory (XSAPT) and benchmark calculations for large supramolecular complexes. *J. Chem. Theory Comput.* **2018**, *14*, 2955−2978.
(67) Carter-Fenk, K.; Lao, K. U.; Liu, K.-Y.; Herbert, J. M. Accurate and efficient ab initio calculations for supramolecular complexes: Symmetry-adapted perturbation theory with many-body dispersion. *J. Phys. Chem. Lett.* **2019**, *10*, 2706−2714.
(68) Carter-Fenk, K.; Lao, K. U.; Herbert, J. M. Predicting and understanding non-covalent interactions using novel forms of symmetry-adapted perturbation theory. *Acc. Chem. Res.* **2021**, *54*, 3679−3690.
(69) Gray, M.; Herbert, J. M. Density functional theory for van der Waals complexes: Size matters. *Annu. Rep. Comput. Chem.* **2024**, *20*, 1−61.
(70) 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.
(71) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(72) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **2010**, *133*, 134105.
(73) 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.
(74) Fedorov, D. G.; Slipchenko, L. V.; Kitaura, K. Systematic study of the embedding potential description in the fragment molecular orbital method. *J. Phys. Chem. A* **2010**, *114*, 8742−8753.
(75) Jacobson, L. D.; Herbert, J. M. An efficient, fragment-based electronic structure method for molecular systems: Self-consistent polarization with perturbative two-body exchange and dispersion. *J. Chem. Phys.* **2011**, *134*, 094118.
(76) Holden, Z. C.; Richard, R. M.; Herbert, J. M. Periodic boundary conditions for QM/MM calculations: Ewald summation for extended Gaussian basis sets. *J. Chem. Phys.* **2013**, *139*, 244108.

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

(77) Fedorov, D. G.; Kitaura, K. Use of an auxiliary basis set to describe the polarization in the fragment molecular orbital method. *Chem. Phys. Lett.* **2014**, *597*, 99−105.
(78) Thapa, B.; Beckett, D.; Jose, K. V. J.; Raghavachari, K. Assessment of fragmentation strategies for large proteins using the multilayer molecules-in-molecules approach. *J. Chem. Theory Comput.* **2018**, *14*, 1383−1394.
(79) 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.
(80) 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.
(81) Herbert, J. M. Dielectric continuum methods for quantum chemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, e1519.
(82) Slattery, S. A.; Surjuse, K. A.; Peterson, C.; Penchoff, D. A.; Valeev, E. Economical quasi-Newton unitary optimization of electronic orbitals. *Phys. Chem. Chem. Phys.* **2024**, *26*, 6557−6573.
(83) Slattery, S. A.; Yon, J. C.; Valeev, E. F. Revisiting artifacts of Kohn−Sham density functionals for biosimulation. *J. Chem. Theory Comput.* **2024**, *20*, 6652−6660.
(84) 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.
(85) Gilson, M. K.; Honig, B. H. The dielectric constant of a folded protein. *Biopolymers* **1986**, *25*, 2097−2119.
(86) 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.
(87) Rodgers, K. K.; Silgar, S. G. Surface electrostatics, reduction potentials, and internal dielectric constant of proteins. *J. Am. Chem. Soc.* **1991**, *113*, 9419−9421.
(88) Nakamura, H. Roles of electrostatic interaction in proteins. *Quart. Rev. Biophys.* **1996**, *29*, 1−90.
(89) 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.
(90) Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A. M.; Huang, Y.; Milletti, F.; Nielsen, J. E.; 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.
(91) King, G.; Lee, F. S.; Warshel, A. Microscopic simulations of macroscopic dielectric constants of solvated proteins. *J. Chem. Phys.* **1991**, *95*, 4366−4377.
(92) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. Prediction of pH-dependent properties of proteins. *J. Mol. Biol.* **1994**, *238*, 415−436.
(93) Demchuk, E.; Wade, R. C. Improving the continuum dielectric approach to calculating pK_as of ionizable groups in proteins. *J. Phys. Chem. A* **1996**, *100*, 17373−17387.
(94) Grycuk, T. Revision of the model system concept for the prediction of pK_a's in proteins. *J. Phys. Chem. B* **2002**, *106*, 1434−1445.
(95) Truchon, J.-F.; Nicholls, A.; Roux, B.; Iftimie, R. I.; Bayly, C. I. Integrated continuum dielectric approaches to treat molecular polarizability and the condensed phase: Refractive index and implicit solvation. *J. Chem. Theory Comput.* **2009**, *5*, 1785−1802.
(96) Li, L.; Li, C.; Zhang, Z.; Alexov, E. On the dielectric "constant" of proteins: Smooth dielectric function for macromolecular modeling and its implementation in DelPhi. *J. Chem. Theory Comput.* **2013**, *9*, 2126−2136.
(97) 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.
(98) Himo, F. Quantum chemical modeling of enzyme active sites and reaction mechanisms. *Theor. Chem. Acc.* **2006**, *116*, 232−240.
(99) Sevastik, R.; Himo, F. Quantum chemical modeling on enzymatic reactions: The case of 4-oxalocrotonate tautomerase. *Bioorg. Chem.* **2007**, *35*, 444−457.
(100) Liao, R.-Z.; Yu, J.-G.; Himo, F. Mechanism of tungsten-dependent acetylene hydratase from quantum chemical calculations. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 22523−22527.
(101) Dasgupta, S.; Herbert, J. M. Using atomic confining potentials for geometry optimization and vibrational frequency calculations in quantum-chemical models of enzyme active sites. *J. Phys. Chem. B* **2020**, *124*, 1137−1147.
(102) Bowling, P. E.; Dasgupta, S.; Herbert, J. M. Eliminating imaginary frequencies in quantum-chemical cluster models of enzymatic active sites. *J. Chem. Inf. Model.* **2024**, *64*, 3912−3922.
(103) 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.
(104) 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.
(105) 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, pp 363−416.
(106) 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. A* **1996**, *100*, 7384−7391.
(107) 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.
(108) 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.
(109) 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.
(110) Gavini, V.; Baroni, S.; Blum, V.; et al. Roadmap on electronic structure codes in the exascale era. *Model. Simul. Mater. Sci. Eng.* **2023**, *31*, 063301.
(111) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51*, 3661−3680.
(112) Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009**, *1*, 187−192.
(113) Kumar, A.; Voet, A.; Zhang, K. Y. J. Fragment based drug design: From experimental to computational approaches. *Curr. Med. Chem.* **2012**, *19*, 5128−5147.
(114) Baker, M. Fragment-based lead discovery grows up. *Nat. Rev. Drug Discovery* **2013**, *12*, 5−7.
(115) Jhoti, H.; Williams, G.; Rees, D.; Murray, C. W. The 'rule of three' for fragment-based drug discovery: Where are we now? *Nat. Rev. Drug Discovery* **2013**, *12*, 644.
(116) Hopkins, A. L.; Keserü, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H. The role of ligand efficiency metrics in drug discovery. *Nat. Rev. Drug Discovery* **2014**, *13*, 105−121.
(117) Efremov, I. V.; Erlanson, D. A. Fragment-based lead generation. In *Lead Generation: Methods, Strategies, and Case Studies*, 1st ed.; Holenz, J., Ed.; Wiley-VCH: Weinheim, 2016; Chapter 6, pp 133−157.
(118) Doak, B. C.; Norton, R. S.; Scanlon, M. J. The ways and means of fragment-based drug design. *Pharmacol. Therapeut.* **2016**, *167*, 28−37.
(119) Kirsch, P.; Hartman, A. M.; Hirsch, A. K. H.; Empting, M. Concepts and core principles of fragment-based drug design. *Molecules* **2019**, *24*, 4309.

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949

(120) Parrish, R. M.; Parker, T. M.; Sherrill, C. D. Chemical assignment of symmetry-adapted perturbation theory interaction energy components: The functional-group SAPT partition. *J. Chem. Theory Comput.* **2014**, *10*, 4417−4431.
(121) Lao, K. U.; Herbert, J. M. Accurate and efficient quantum chemistry calculations of noncovalent interactions in many-body systems: The XSAPT family of methods. *J. Phys. Chem. A* **2015**, *119*, 235−253.
(122) Semiempirical Extended Tight-Binding Program Package. https://github.com/grimme-lab/xtb (accessed Dec 02, 2024).
(123) Ohio Supercomputer Center. https://osc.edu/ark:/19495/f5s1ph73 (accessed Dec 01, 2024).

https://doi.org/10.1021/acs.jcim.4c01987
J. Chem. Inf. Model. 2025, 65, 937−949