Quick-and-Easy Validation of Protein–Ligand Binding Models Using Fragment-Based Semi-Empirical Quantum Chemistry

Paige E. Bowling[1,2], Dustin R. Broderick[2], and John M. Herbert[1,2*]
[1] *Biophysics Graduate Program, The Ohio State University, Columbus, Ohio 43210 USA*
[2] *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210 USA*

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 semi-empirical 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 semi-empirical 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 provide affords 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,[1–14] 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 non-covalent binding of a ligand to a protein.[14] 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.[11–17]

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 semi-empirical quantum chemistry model[18] with a fragment-based procedure for computing the interaction energy (ΔE_{int}) between a ligand and an enzyme model.[14] 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.[14] 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.[19] This enables large-scale quantum chemistry calculations using only workstation-level resources (*i.e.*, single-node parallelism),[14,20–22] 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 ligand–protein binding calculations 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 energies $\{E_I\}$ for a collection of independent fragments ($I = 1, \ldots, N_{\text{frag}}$):

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

*herbert@chemistry.ohio-state.edu

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

$$\Delta E_{IJ} = E_{IJ} - E_I - E_J \ , \tag{2}$$

three-body terms

$$\begin{aligned}\Delta E_{IJK} = E_{IJK} - \Delta E_{IJ} - \Delta E_{IK} - \Delta E_{JK} \\ - E_I - E_J - E_K\end{aligned} \tag{3}$$

and so forth.[19,23] 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,[14,24] we use single-residue fragments obtained by cutting the C–C bond at C_α–C(=O). This avoids severing the more polar peptide (C–N) bond. The severed valencies thus created are capped with hydrogen atoms positioned at[25]

$$\mathbf{r}_{\rm cap} = \mathbf{r}_1 + \left(\frac{R_1 + R_{\rm H}}{R_1 + R_2}\right)(\mathbf{r}_2 - \mathbf{r}_1) \ . \tag{4}$$

Here, $R_1 = R_2 = 1.70$ Å and $R_{\rm H} = 1.1$ Å are atomic van der Waals radii for carbon and for hydrogen, respectively. More sophisticated capping strategies have been suggested,[26–29] but these are also more complicated and 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, $R_{\rm cut}$. In previous work on protein fragmentation,[14,24] we showed that $R_{\rm cut} = 8$ 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 with the protein data bank (PDB) code 181L, which is one of the systems considered below. In that case, $N_{\rm frag} = 164$ for the entire protein system but the use of $R_{\rm cut} = 8$ 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 FRAGME∩T code.[21,30] which drives all of the calculations reported here. FRAGME∩T implements both distance- and energy-based screening protocols[20–22] 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.[31] 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 $\Delta E_{\rm int}$ are computed according to

$$\Delta E_{\rm int} = E_{\rm P:L} - E_{\rm P} - E_{\rm L} \ , \tag{5}$$

with consistent application of MBE(n) to compute both the energy of the isolated protein ($E_{\rm P}$) and that of the protein–ligand complex ($E_{\rm P:L}$). The ligand energy ($E_{\rm L}$) is computed without fragmentation. Many of the n-body terms will cancel in eq. 5 and need not be computed. The present version of FRAGME∩T identifies these terms *a priori* and removes them, using the algorithm described in Ref. 22, which leads to considerable cost savings for $\Delta E_{\rm int}$ calculations. However, the present calculations were performed contemporaneously with that development and this savings was not exploited. As such, timing data reported here reflect the cost of all n-body terms in eq. 5, subject only to the distance cutoff $R_{\rm cut}$.

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 $\sim$ 300 atoms, for example, BSSE effects up to $\sim$ 50 kcal/mol have been documented when double-ζ basis sets are used,[32] as quantified by the difference between counterpoise-corrected and uncorrected values of $\Delta E_{\rm int}$. Versions of counterpoise correction designed for use with MBE(n) have been reported[33–35] but are not yet implemented in FRAGME∩T, 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 We selected two structures from the L99A and L99A/M102Q data sets of T4-lysozyme complexes:[36–38] 181L and 1LI2, where the ligands are benzene and phenol, respectively. Both complexes were also considered in our recent study of fragmentation protocols for protein–ligand interactions.[14] Each involves two Cl^- ions that were combined into a single monomer along with any residues within 2.5 Å of the ion. We also selected two complexes (1O48 and 1BOZ) from the large-inhibitor data set,[14] as examples where the ligands are much larger.

In the creation of QM models, we insist on automated methods that provide a reproducible, black-box approach to constructing structural models for QM calculations, without the use of any system-specific information beyond what is contained in a crystal structure. Structural models of the aforementioned protein–ligand complexes, containing anywhere from 120 to 1,726 atoms, were generated by one of several different approaches that are described in Section 2 .3.

Crystal structures were obtained from the PDB and protonated using the H++ web server,[39] specifying pH = 7.0, salinity of 0.15 M, $\varepsilon_{\rm in} = 10$, and $\varepsilon_{\rm out} = 80$. The large ligands were protonated separately using the PYMOL program.[40] As in previous work,[14] the geometry was then relaxed using the GFN2-xTB semi-empirical method,[41] in conjunction with a generalized Born/solvent-accessible surface area (GBSA/SASA) implicit solvation model for water.[42] Most crystallographic water molecules were removed after relaxation, although those coordinated with

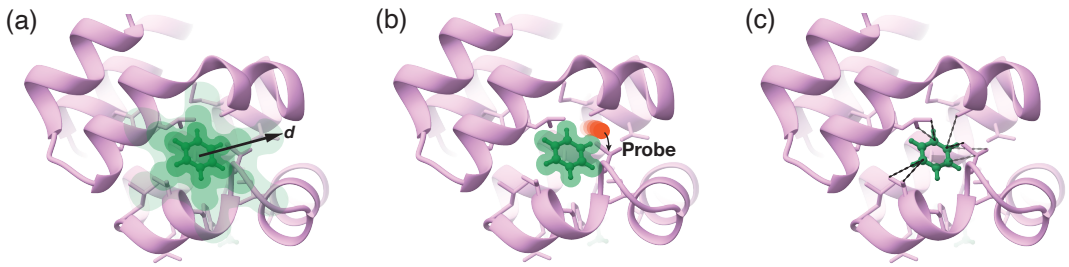

Fig. 1: Various methods for selecting amino acid residues around a benzene ligand that is shown atomistically, in green: (a) distance-based selection, using a cutoff distance d; (b) Probe selection, rolling a probe sphere over the van der Waals surface; and (c) Arpeggio selection, based on atomic information.

the Cl^- ions were retained, as were any crystallographic water molecules within 2.5 Å of the ligand.

2.3. Model Construction. The simplest approach to model construction uses a distance to select amino acid residues proximate to the ligand. Here, residue selection was performed using PyMol with various cutoffs, ranging from $d = 2.5$ to $d = 10.0$ Å. Any residue having at least one atom that is within the cutoff distance of any ligand atom is included in the QM model. This method is simple and systematic but its weakness lies in the fact that many biologically important active sites are highly aspherical. A good example is human catechol O-methyltransferase (COMT),[43–46] which has become something of a computational benchmark system[7,11,24,47] because it has a well-resolved crystal structure,[45] kinetics data,[43] and numerous known inhibitors.[7–9] It also has a catalytically important Mg^{2+} ion,[48] which engenders significant charge transfer and many-body polarization effects that cause especially slow convergence with respect to model size.[7,11,47]

For examples such as COMT, one might expect a "chemically informed" model-construction algorithm to converge more quickly than a brute-force distance-based approach. Therefore, as alternatives we examine models generated using the Residue Interaction Network Residue Selector (RINRUS) toolkit,[49] developed by DeYonker and co-workers.[11–13,15,16] We operate RINRUS in one of two modes: "Probe" or "Arpeggio".[50–53] These are illustrated in Fig. 1 alongside the distance-based method.

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 generate close-contacts.[50,51] 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 models.

The Arpeggio method operates similarly but uses uses atom types, interatomic distances, and angles to classify inter-residue interactions into 15 different categories,[52] 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 indicate the ligand. A model is then created by retaining all residues I for which

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

where τ_{2B} is a user-specified threshold. MBE(3) calculations are built upon this two-body screening by including $\Delta E_{I,J,\text{ligand}}$, for all residues J such that $\Delta E_{I,\text{ligand}}$ satisfies eq. 6.

2.4. Quantum Chemistry Calculations. The primary electronic structure method used in this work is HF-3c,[18] which starts from a minimal-basis Hartree-Fock (HF) calculation then add three empirical corrections. The latter are parameterized for use with a specific basis set ("MINIX"),[18] so in what follows we will not indicate the basis set for HF-3c calculations, as it is always MINIX. Some conventional DFT calculations are performed as well, using the ωB97X-V functional[54] as a representative example that performs well for small-molecule van der Waals complexes.[55,56] (For molecules with $\gtrsim 100$ atoms, DFT performs less well;[56] fragmentation may be a useful approach to exploit its better performance in small non-covalent complexes.)

Basis sets used for the ωB97X-V calculations are minimally augmented ("ma") versions of the standard Karlsruhe basis sets,[57,58] which are known as def2-ma-SVP, def2-ma-TZVP, and def2-ma-QZVP.[59] (These are proper subsets of the basis sets def2-SVPD, def2-TZVPD, and def2-QZVPD.[58,59]) Diffuse basis functions are important for non-covalent DFT calculations, even when triple-ζ basis sets are employed, but minimal augmentation appears to be sufficient.[32] Our preference for the simple MBE(n) fragmentation scheme that is described in Sec-

tion 2.1, 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.[19,60–63] For all calculations, the self-consistent field 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.[64]

Previous work applying MBE(n) to thermochemical calculations in large enzyme models has demonstrated that low-dielectric boundary conditions are necessary to make MBE(n) converge in the presence of ionic fragments, as will inevitably arise when native protonation states are considered. Therefore, all quantum chemistry calculations reported here use a conductor-like polarizable continuum model (C-PCM),[65,66] with a dielectric constant $\varepsilon = 4$. This is implemented using the switching/Gaussian discretization procedure.[66–69] C-PCM calculations use a van der Waals molecular cavity,[65] constructed from modified Bondi atomic radii,[70] with $R_{\mathrm{vdW}} = 1.2R_{\mathrm{Bondi}}$. The van der Waals surface is discretized using atom-centered Lebedev grids.[67] For the n-body DFT calculations, this discretization employs 110 Lebedev points for hydrogen and 194 points for other nuclei, whereas the HF-3c subsystem calculations and the HF-3c full-protein calculations use 50 points for hydrogen and 110 for other nuclei. A conjugate gradient PCM solver was used for the full-protein calculations.[69]

3 Results and Discussion

The primary goal of this work is to demonstrate that fragment-based semi-empirical 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 validate the use of HF-3c against conventional DFT, in Section 3.1. We then consider energy screening of the two-body HF-3c calculations in Section 3.2, which further improves the efficiency. The resulting method is used in Section 3.3 to evaluate the convergence of ΔE_{int} for various binding-site models. Comparisons to DFT results are presented in Section 3.4.

3.1. Comparing HF-3c to DFT. We first consider how two- and three-body corrections computed at the HF-3c level compare to the corresponding quantities obtained using ωB97X-V in basis sets through quadruple-ζ quality. Correlations between the two methods are illustrated in Fig. 2 for one particular protein–ligand complex (181L), and analogous plots for the other complexes considered in this work can be found in Figs. S1–S3.

Correlation between HF-3c and ωB97X-V is quite good for the two-body terms (Fig. 2a), and there is a

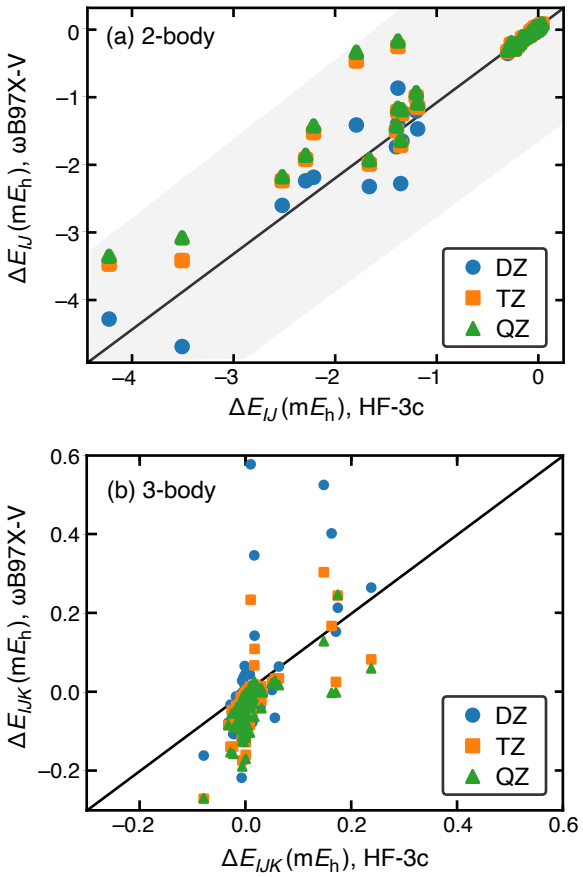

Fig. 2: Correlations between HF-3c and ωB97X-V (with the latter evaluated in several different basis sets), for (a) two-body corrections ΔE_{IJ} and (b) three-body corrections ΔE_{IJK}, for the protein–ligand complex 181L. The diagonal line indicates where the two methods predict the same value and gray area in (a) represents ±1 kcal/mol difference. Only terms that involve the ligand (benzene) are plotted, using $R_{\mathrm{cut}} = 8$ 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").

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 Fig. 2a 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-

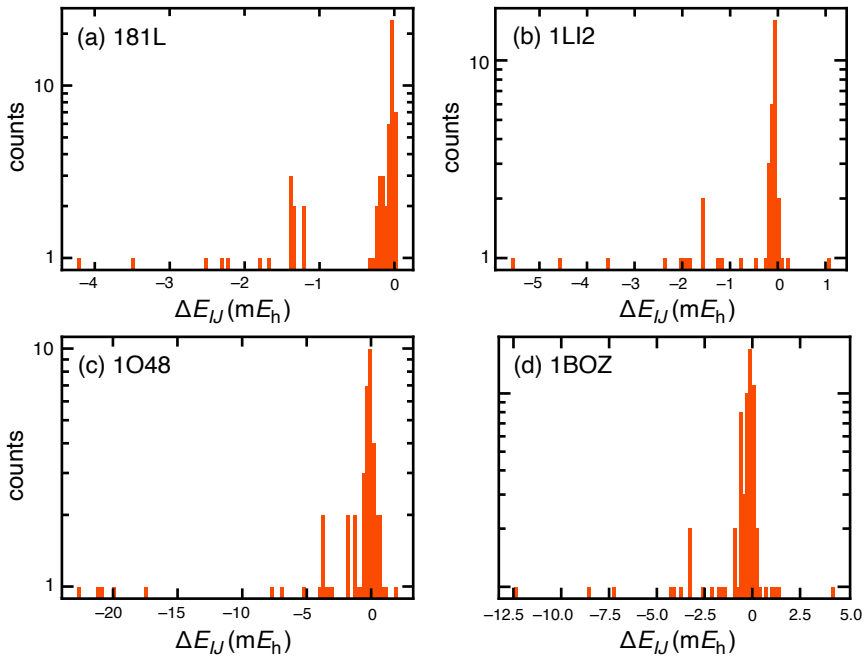

Fig. 3: Histograms of the two-body terms ΔE_{IJ}, where either I or J is the ligand, for protein–ligand complexes (a) 181L, (b) 1LI2, (c) 1O48, and (d) 1BOZ. All calculations were performed using HF-3c.

body corrections suggests that the BSSE is largely eliminated using def2-ma-TZVP, which is typical for small fragments.[32]

Correlations between HF-3c and ωB97X-V are much less pronounced for the three-body terms (Fig. 2b), 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,[19] although it is also possible that the three-body interactions are exaggerated by ωB97X-V calculations in a double-ζ basis set.[22] Setting aside the ωB97X-V/def2-ma-SVP results in Fig. 2b, 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 3 provides a closer look at the 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 Fig. 3 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.

3.2. 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, it is possible to obtain sub-kcal/mol fidelity with respect to a full-protein calculation using only MBE(2), 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 MBE(2) and MBE(3) 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

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

System	$\tau_{2B}/10^{-4}E_h$	No. atoms	MBE(2) error (kcal/mol)[a] absolute	MBE(2) error (kcal/mol)[a] per monomer	MBE(2) time[b] (h)	MBE(3) error (kcal/mol)[a] absolute	MBE(3) error (kcal/mol)[a] per monomer	MBE(3) 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	1,145	0.57	0.01	22	0.49	0.01	892
	1.25	1,637	1.85	0.02	45	2.85	0.03	3,512

[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 (Unfragmented) HF-3c Interaction Energies and Timings

System	ΔE_{int} (kcal/mol)	time[a] (hours)
181L	−19.4	4,156
1LI2	−18.8	5,542
1O48	−89.9	854
1BOZ	−31.3	5,018

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

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,[14,23,71,72] 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 per monomer has been suggested,[73] representing 10% of $k_{\text{B}}T$ 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 $0.1 \times k_{\text{B}}T$ 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, $\langle \Delta E_{\text{int}} \rangle$) and experimental binding affinities $\Delta G^{\circ}_{\text{bind}}$ is many times larger than $0.1k_{\text{B}}T$. (For a lengthy discussion of this point, see our recent work on fragmentation protocols for protein–ligand interaction energies.[14]) In addition, it is important to recognize the intrinsic limitations of semi-empirical quantum chemistry, as there is no sense in pushing for higher fidelity than is warranted by the intrinsic accuracy of the electronic structure method. To that end, we note that errors in HF-3c interaction energies average 4 kcal/mol for the combined L7[74] and S30L[75] data sets of large supramolecular complexes.[76] This is comparable to the performance of dispersion-corrected and dispersion-inclusive DFT methods applied to the same data sets.[77,78]

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 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.[19,79] Table 1 provides the aggregate computing time for the HF-3c MBE(n) calculations and Table 2

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

System	Model	No. atoms	MBE(2) error (kcal/mol)[b] absolute	MBE(2) error (kcal/mol)[b] per monomer	MBE(2) time[c] (h)	MBE(3) error (kcal/mol)[b] absolute	MBE(3) error (kcal/mol)[b] per monomer	MBE(3) 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 $R_{\text{cut}} = 8$ 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.

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.[79]

Even so, the cost reduction is significant for the MBE(2) calculations, amounting to no more than 1–2% of the supersystem cost, depending on model size. For the largest system considered here (1BOZ, with 3,124 atoms), and for the model constructed using $\tau_{\text{2B}} = 2.5 \times 10^{-4}$ 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 cost without fragmentation. For 1O48 (with 1,781 atoms), the cost of the $\tau_{\text{2B}} = 2.5 \times 10^{-4}$ 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.

3.3. 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 considered distance-based models ranging from $d = 2.5$ Å to $d = 10.0$ Å but only the 4 Å 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 3.2, increasing d (to increase model size) improves the MBE(2) and MBE(3) accuracy only up to a point; errors eventually reach a plateau such that 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 others 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 RINRUS 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 binding site is more spherical (and the ligand is much smaller) as compared to 1O48 and 1BOZ. Convergence to the supermolecular value of ΔE_{int} is slower for the latter two systems.

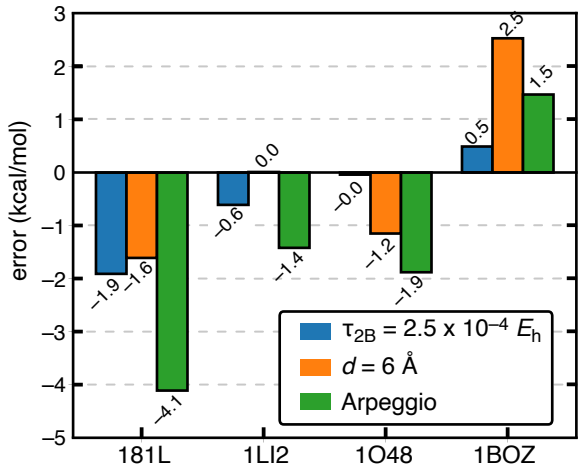

Fig. 4: Errors in ΔE_{int} for MBE(3) calculations at the HF-3c level, comparing comparing three different methods to construct a binding-site model for four different protein–ligand complexes.

Models generated using the RINRUS toolkit in either its "Probe" or "Arpeggio" configurations (Section 2.3) are generally similar to one another although Arpeggio includes a slightly larger number of residues. (In each case considered here, residues selected by the Probe model represent a proper subset of those selected using the Arpeggio construction.) While the Arpeggio models are slightly larger, they are not significantly or consistently more accurate.

At present, the RINRUS-generated models do not outperform the others but there are several avenues that could be used to improve the former. These include addition of a second interaction sphere, or the use of a seed that is larger than just the ligand, containing some nearest-neighbor residues. A recent development in RINRUS is an option to use a form of pairwise symmetry-adapted perturbation theory (SAPT),[80] as a means to decompose the interaction energy between a protein and individual residue main chains or side chains.[13] (It is not entirely clear why the "functional group" version of SAPT[80] is necessary in this capacity. Other pairwise forms of SAPT could probably be used instead.[78,81,82]) Thus, it is probably possible to further refine the RINRUS-based construction of binding-site models.

Models discussed in this section are generally smaller than the τ_{2B} models described in Section 3.2, 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 on 10 h for the largest protein–ligand complex, as compared to 22 h for our preferred τ_{2B}-derived model.

Figure 4 compares MBE(3) errors in across the data set, using three different paradigms to construct the binding-site model: 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.

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

System	ΔE_{int} (kcal/mol)			
	HF-3c	ωB97X-V		
		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. 14. [b]def2-ma-SVP. [c]def2-ma-TZVP. [d]def2-ma-QZVP.

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 eleven in 1BOZ. Apparently, these do not significantly improve the accuracy when compared to the τ_{2B} model construction, however. 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 but the model constructed using $\tau_{2B} = 2.5 \times 10^{-4}$ E_h includes 15 additional residues. 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.

3.4. 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 $R_{\text{cut}} = 8$ 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,[14]

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 $R_{\text{cut}} = 8$ Å. [b]Error is measured with respect to the full-system (fragmented) calculation at the same level of theory. [c]def2-ma-SVP. [d]def2-ma-TZVP. [e]def2-ma-QZVP.

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,[32] and convergence of the ωB97X-V interaction energies in Table 4 provides some measure of it. For the largest system considered here (1BOZ), the ωB97X-V/def2-SVP and ωB97X-V/def2-QZVP interaction energies differ by 23 kcal/mol.

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 $R_{\text{cut}} = 8$ Å, are listed in Table 5 for the best-performing model systems, as determined in Sections 3.2 and 3.3. For the small-ligand complexes 181L and 1LI2, the $d = 6$ Å model tends to exhibit the smallest errors, although the $\tau_{2B} = 2.5 \times 10^{-4}$ E_h model affords errors that are larger by only 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 likely to be 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 ≤ 1.5 kcal/mol for the $\tau_{2B} = 2.5 \times 10^{-4}$ E_h and the $d = 6$ Å models.

4 Conclusion

This work extends other recent work from our group,[14,24] 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[30] 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 semi-empirical HF-3c method,[18] correlate very well with results from high-quality DFT calculations (*e.g.*, 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. This is perhaps the most reliable algorithm for QM model construction of the ones examined here.

Energy-based model construction typically results in larger models as compared to algorithms implemented in the RINRUS program,[11,13,49] nevertheless the τ_{2B} models consistently deliver higher accuracy. Simple distance-based models with a 6 Å cutoff are also found to be effective. RINRUS models 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 semi-empirical 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 way achieve a fidelity of 1–2 kcal/mol in triple- or quadruple-ζ basis sets, as compared to MBE(2)-DFT calculations on larger, converged models of the protein. The combination of fast semi-empirical MBE(2) calculations, used to test convergence of ΔE_{int} with respect to model size, and convergent MBE(n) protocols for evaluating ΔE_{int},[14] represents a powerful tool chain for quantum-chemical studies of drug–protein interactions. The same semi-empirical 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.[24]

Data Availability Statement

All calculations were performed using the open-source FRAGME∩T code that is available at the URL in Ref. 30. In the present work, FRAGME∩T is interfaced with Q-

CHEM,[31] 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. The RINRUS software is available at the URL in Ref. 49. Coordinates for the protein–ligand complexes are provided in the Supporting Information.

5 Supporting Information

Additional data for fragmentation calculations (PDF)
List of residues included in the QM models (PDF)
Coordinates for the protein–ligand complexes (zip)

6 Notes

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

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.[83]

7 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] Bowling, P. E.; Broderick, D. R.; Herbert, J. M. Convergent protocols for protein–ligand interaction energies using fragment-based quantum chemistry. *ChemRxiv* **2024** (DOI: 10.26434/chemrxiv-2024-7v7pv).

[15] 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.

[16] 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.

[17] K.-S.Csizi,; Reiher, M. Universal QM/MM approaches for general nanoscale applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2023,** *13,* e1656.

[18] Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013,** *34,* 1672–1685.

[19] Herbert, J. M. Fantasy versus reality in fragment-based quantum chemistry. *J. Chem. Phys.* **2019,** *151,* 170901.

[20] 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.

[21] 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.

[22] Broderick, D. R.; Herbert, J. M. Delocalization error poisons the density-functional many-body expansion. *Chem. Sci.* (in press; preprint available at DOI: 10.26434/chemrxiv-2024-5tt53-v2).

[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] Vornweg, J. R.; Jacob, C. Protein-ligand interaction energies from quantum-chemical fragmentation methods: Upgrading the MFCC-scheme with many-body contributions. *ChemRxiv* **2024** (DOI: 10.26434/chemrxiv-2024-mt4nk).
[30] Broderick, D. R.; Bowling, P. E.; Shockey, J.; Higley, J.; Dickerson, H.; Ahmed, S.; Herbert, J. M. "FRAGME∩T", https://gitlab.com/fragment-qc/fragment.
[31] Epifanovsky, E. *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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005,** *48,* 3714–3728.
[37] 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.
[38] Mobley, D. L.; Gilson, M. K. Predicting binding free energies: Frontiers and benchmarks. *Annu. Rev. Biophys.* **2017,** *46,* 531–558.
[39] 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. *Nucl. Acids Res.* **2012,** *40,* 537–541.
[40] PyMOL molecular graphics system, v. 2.1, https://pymol.org (accessed 2024-10-25).
[41] 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.
[42] 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.
[43] Lotta, T.; Vidgren, J.; Tilgmann, C.; Ulmanen, I.; Melen, K.; Julkunen, I.; Taskinen, J. Kinetics of human soluble and membrane-bound catechol *O*-methyltransferase: A revised mechanism and description of the thermolabile variant of the enzyme. *Biochemistry* **1995,** *34,* 4202–4210.
[44] Bonifácio, M. J.; Archer, M.; Rodrigues, M. L.; Matias, P. M.; Learmouth, D. A.; Carrondo, M. A.; Soares-da-Silva, P. Kinetics and crystal structure of catechol *O*-methyltransferase complex with co-substrate and a novel inhibitor with potential therapeutic application. *Mol. Pharmacol.* **2002,** *62,* 795–805.
[45] Rutherford, K.; Le Trong, I.; Stenkamp, R. E.; Parson, W. W. Crystal structures of human 108V and 108M catechol *O*-methyltransferase. *J. Mol. Biol.* **2008,** *380,* 120–130.
[46] Bastos, P.; Gomes, T.; Ribeiro, L. Catechol-*O*-methyltransferase (COMT): An update on its role in cancer, neurological and cardiovascular diseases. *Rev. Physiol. Biochem. Pharmacol.* **2017,** *173,* 1–40.
[47] Jindal, G.; Warshel, A. Exploring the dependence of QM/MM calculations of enzyme catalysis on the size of the QM region. *J. Phys. Chem. B* **2016,** *120,* 9913–9921.
[48] Axelrod, J.; Tomchick, R. Enzymatic O-methylation of epinephrine and other catechols. *J. Biol. Chem.* **1958,** *233,* 702–705.
[49] Residue Network Interaction Residue Selector (RINRUS), https://github.com/natedey/RINRUS.
[50] 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.
[51] 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.
[52] 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.
[53] Arpeggio, https://biosig.lab.uq.edu.au/arpeggioweb (accessed 2024-10-24).
[54] 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.
[55] Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017,** *115,* 2315–2372.
[56] Gray, M.; Herbert, J. M. Density functional theory for van der Waals complexes: Size matters. *Annu. Rep. Comput. Chem.* **2024,** *20,* 1–61.
[57] 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.
[58] Rappoport, D.; Furche, F. Property-optimized Gaussian

basis sets for molecular response calculations. *J. Chem. Phys.* **2010,** *133,* 134105.
[59] 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.
[60] 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.
[61] 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.
[62] 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.
[63] 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.
[64] 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.
[65] Herbert, J. M. Dielectric continuum methods for quantum chemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021,** *11,* e1519.
[66] 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.
[67] 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.
[68] 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.
[69] 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.
[70] 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.
[71] 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.
[72] 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.
[73] Ouyang, J. F.; Bettens, R. P. A. Many-body basis set superposition effect. *J. Chem. Theory Comput.* **2015,** *11,* 5132–5143.
[74] 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.
[75] Sure, R.; Grimme, S. Comprehensive benchmark of association (free) energies of realistic host–guest complexes. *J. Chem. Theory Comput.* **2015,** *11,* 3785–3801.
[76] 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.
[77] 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.
[78] 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.
[79] Gavini, V. *et al.* Roadmap on electronic structure codes in the exascale era. *Model. Simul. Mater. Sci. Eng.* **2023,** *31,* 063301.
[80] 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.
[81] 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.
[82] 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.
[83] Ohio Supercomputer Center, http://osc.edu/ark:/19495/f5s1ph73 (accessed 2024-10-26).

TOC Graphic