Convergent Protocols for Computing Protein–Ligand Interaction Energies Using Fragment-Based 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

Fragment-based quantum chemistry methods offer a means to sidestep the steep nonlinear scaling of electronic structure calculations so that large molecular systems can be investigated using high-level methods. Here, we use fragmentation to compute protein–ligand interaction energies in systems with several thousand atoms, using a new software platform for managing fragment-based calculations that implements a screened many-body expansion. Convergence tests using a minimal-basis semi-empirical method (HF-3c) indicate that two-body calculations, with single-residue fragments and simple hydrogen caps, are sufficient to reproduce interaction energies obtained using conventional supramolecular electronic structure calculations, to within 1 kcal/mol at about 1% of the computational cost. We also demonstrate that the HF-3c results are illustrative of trends obtained with density functional theory in basis sets up to augmented quadruple-ζ quality. Strategic deployment of fragmentation facilitates the use of converged biomolecular model systems alongside high-quality electronic structure methods and basis sets, bringing *ab initio* quantum chemistry to systems of hitherto unimaginable size. This will be useful for generation of high-quality training data for machine learning applications.

1 Introduction

There is an urgent and growing need to provide high-accuracy training data for machine learning (ML) applications. This is especially true for biological systems, where understanding protein–ligand interactions is crucial for advancing drug discovery and where ML-based screening is playing an increasingly prominent role.[1–16] Integration of quantum chemistry with ML has the potential to revolutionize computational biology and to reduce the cost of drug discovery by enabling the use of non-empirical screening tools.

Encoding biomolecular systems requires a large, inconsistently-sized parameter space that is intractable to train and use when considered as a whole. ML approaches commonly reduce complex systems into their component parts ("tokens"), then infer properties of the system as a whole based on relationships between tokens. This approach is complementary to fragmentation methods in quantum chemistry,[17,18] which approximate supersystem properties by systematically partitioning that system into numerous fragments, for which it is relatively inexpensive to perform high-quality calculations. This provides a hierarchy of well-defined fragments and a database can be used to train ML model for large systems. Generating high-quality training data for protein–ligand binding is complicated, however, by the requisite size of the models involved. Furthermore, the non-covalent nature of many protein–ligand interactions means that the electronic structure model must be chosen carefully.[19]

Fragment-based quantum chemistry leverages distributed computing by means of physics-based approximations, as an alternative to parallelization of conventional algorithms.[20,21] In this way, $\mathcal{O}(N^p)$ computational scaling (where N measures system size and the exponent p depends on the electronic structure model) is reduced to $N_{\text{sub}} \times \mathcal{O}(n^p)$, where n is a fixed subsystem size that does not grow with N, and N_{sub} is the number of subsystems, which increases with N. This is an attractive approach to parallelization, in part because the storage footprint (*i.e.*, memory and/or disk space), which is often the practical limitation, is reduced to that of the largest subsystem and checkpointing can be organized at the level of individual subsystem calculations. However, the number of fragments can be prohibitive for large systems, and must be culled via some kind of screening algorithm.[22–26] This is necessary not only to reduce cost but also to forestall precision issues associated with calculations that might involve 10^5 or more individual separate subsystems.[27–29]

To this end, we have recently introduced a new software framework, FRAGME∩T,[26,30] with inherent database management, parallelization, and screening capabilities. It is built upon a generalized many-body expansion (MBE)[17,31–33] and interfaced with numerous quantum chemistry codes. In recent applications, FRAGME∩T has been used to investigate enzyme thermochemistry in large active-site models,[25] and to perform high-order n-body calculations on water clusters and ion–water clusters.[26,34] In the present work, we aim to apply fragmentation to protein–ligand interaction energies using enzyme models that include not just nearest-neighbor residues but which afford energetically converged interaction energies. Even at the level of density functional theory (DFT), there have been few studies with converged results for full-protein models of ligand binding.[35,36]

*herbert@chemistry.ohio-state.edu

There has been other work applying fragment-based quantum chemistry to calculate protein–ligand interaction energies,[37–48] mostly using DFT although a few studies using second-order Møller-Plesset perturbation theory in small basis sets have been published.[45–48] The purpose of the present work is to establish protocols that are robust and reliable, which could eventually be used at better levels of theory. Crucially, we aim to compute interaction energies ($\Delta E_{\rm int}$) that are faithful to a supramolecular calculation performed at the same level of theory (method and basis set), and to use sufficiently large molecular models so that $\Delta E_{\rm int}$ is converged with respect to further increases in system size. Our approach is based on the MBE truncated at n-body interactions [MBE(n)] and we examine convergence for n = 2–4 using single-residue fragments, in models containing up to 3,124 atoms. This represents unprecedented size and scope for application of MBE(n).

The present calculations use DFT and semi-empirical quantum chemistry but extension to correlated wave function models can be envisaged. Even for DFT calculations, our goal is to reach the basis-set limit. For that purpose, the widely-used "fragment molecular orbital method" (FMO)[49] is inadequate. As applied to protein–ligand interaction energies,[40–43,46,47,50–52] FMO is tantamount to MBE(2) with an electrostatic embedding scheme that is known to be unstable in large basis sets.[17] Much of the FMO literature on drug discovery is focused on pairwise analysis of interaction energies, rather than the absolute value of $\Delta E_{\rm int}$,[47,51–54] but for ML applications we desire a scheme that is robust enough to target $\Delta E_{\rm int}$ itself.

We also wish to avoid complicated capping schemes, as in "molecular fragmentation with conjugate caps" (MFCC).[55] The "conjugate caps" amount to the backbone of the neighboring amino acid residue, the size of which makes MFCC difficult to generalize to arbitrary n-body interactions.[44,56] In contrast, we have found that MBE(n) with simple hydrogen-atom caps can be used to obtain converged thermochemical quantities for enzymatic reactions.[25] Lastly, we desire a method that can be applied to enzymes in their native protonation states, so that the ability to describe ions (and to use diffuse basis functions) is required. In fragment-based calculations, ionizable side chains are often protonated so as to obtain charge-neutral fragments,[37–39,57] as this minimizes many-body polarization effects. However, there is no guarantee that the neutralized enzyme remains functional.

In previous work, MBE(n) has been successfully applied to enzymatic thermochemistry with all of the aforementioned caveats.[25] Inclusion of ionic residues required low-dielectric boundary conditions to eliminate spurious many-body effects, which is likely a consequence of delocalization error in DFT, as discussed elsewhere.[34] The present work extends the thermochemical protocols developed in Ref. 25 to the case of protein–ligand binding. We introduce a set of four T4-lysozyme complexes with small aromatic ligands and four other complexes with large ligands. These systems are then used to assess both the accuracy and the cost of various fragment-based methods to compute $\Delta E_{\rm int}$.

2 Methods

2.1. Data Sets. Bacteriophage T4 lysozyme promotes the release of phage particle from the wall of a cell by breaking down peptidoglycan, allowing for the injection of genomic DNA into the host *Escherichia coli* cell.[58] The class of enzymes considered here have apolar and polar binding sites and are known as L99A and L99A/M102Q, respectively.[59–61] For these systems, calculation of protein–ligand interaction energies has proven challenging for classical molecular dynamics methods.[61] Benchmark data sets of crystal structures and binding energies for both sites, with a variety of non-covalent ligands, were introduced in a recent review.[61] These examples having binding energies within a narrow range from 4.0–6.7 kcal/mol,[61] with estimated uncertainties that are < 0.2 kcal/mol.[60–65]

We selected two representative systems from the L99A data set, with protein databank (PDB) codes 181L[66] and 4W54.[65] The L99A/M102Q data set introduces a point mutation at one side of the binding site, replacing methionine residue 102 with the polar side chain of glutamine to serve as a hydrogen-bond acceptor. From this data set we selected representative systems 1LI2[60] and 3HUA.[64] The ligands for these four T4-lysozyme complexes are benzene (for 181L), ethylbenzene (for 4W54), phenol (for 1LI2), and indole (for 3HUA); see Fig. 1a.

In addition to this T4 lysozyme data set, an additional set of proteins with fewer than 200 residues but much larger ligands was chosen for additional tests. This data set ranges from the compact tyrosine kinase structure (PDB: 1O48)[67] to a large inhibitor of dihydrofolate reductase (PDB: 1BOZ).[68] In ascending order of size, they are 1O48,[67] 1ZP5,[69] 1MMQ,[70] and 1BOZ.[68] All of the ligands, which are depicted in Fig. 1b, serve as inhibitors and we refer to this set of complexes as the "large inhibitor data set" (LIDS).

The ligand of 1O48 binds to the SH2 domain of $^{\rm pp60}$Src kinase,[67] which is important in the control of cell proliferation, differentiation, motility, and adhesion.[71] This site serves as a potential target as $^{\rm pp60}$Src kinase has been linked to bone resorption.[72] The ligands for 1ZP5 and 1MMQ serve as inhibitors for metalloproteases; the ligand in 1ZP5 serves as an inhibitor for N-hydroxyurea and 1MMQ's inhibitor binds to matrilysin (uterine metalloproteinase).[69,70] Enzymes 1ZP5 and 1MMQ each contain two Zn^{2+} and two Ca^{2+} ions. In both of these metalloproteinases, over-regulation can lead to uncontrolled degradation of the extracellular matrix, which is seen in diseases including cancer, arthritis, and multiple sclerosis.[69,70] The design of the ligand in 1BOZ was meant to serve as an inhibitor of *Toxoplasma*

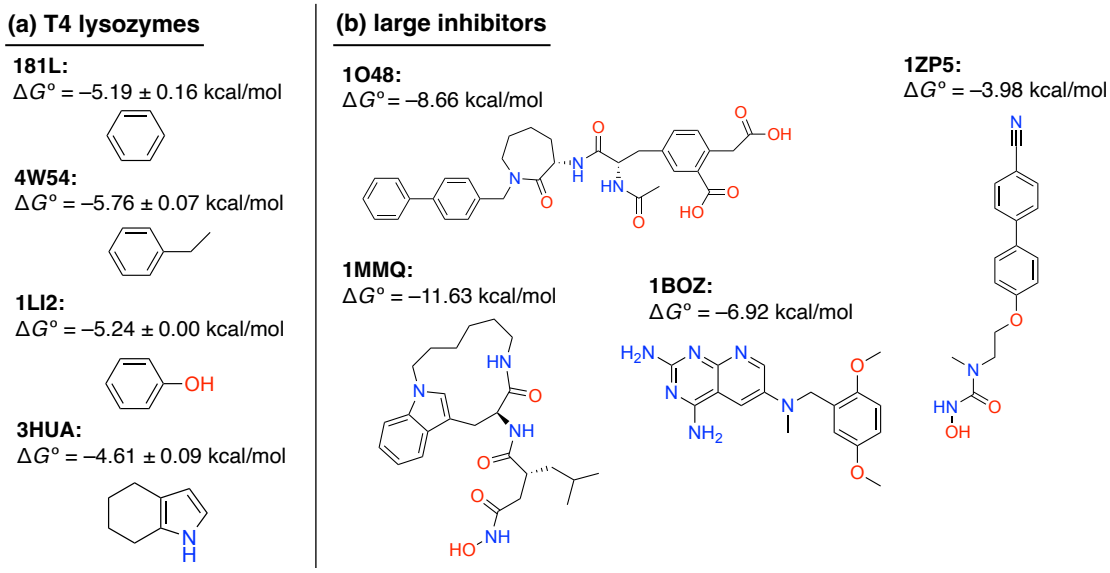

Fig. 1: Structures and binding energies for the ligands used in this work along with the PDB code for the protein–ligand crystal structure: (a) T4 lysozymes with small ligands and (b) large inhibitors ("LIDS").

gondii dihydrofolate reductase and a potential antitumor agent.[68] In addition to the inhibitor, 1BOZ also contains NADPH as a cofactor.

2.2. System Preparation. Crystal structures were obtained from the PDB and protonated using the H++ web server,[73] for pH = 7.0, salinity of 0.15 M, and dielectric constants $\varepsilon_{\text{in}} = 10$ and $\varepsilon_{\text{out}} = 80$. Ligands were protonated separately using PyMOL.[74] Except as noted below, the resulting structures were relaxed using the semi-empirical GFN2-xTB model[75] with a generalized Born/solvent-accessible surface area (GB/SASA) implicit solvation model for water.[76] GFN2-xTB affords reasonable protein geometries as compared to crystal structures,[77] whereas direct calculation of protein–ligand interaction energies with GFN2-xTB affords mixed results as compared to DFT calculations.[78–80] Both metalloenzymes (1ZP5 and 1MMQ) proved difficult to relax using GFN2-xTB, as the Zn^{2+} ion was repeatedly expelled from its binding site in numerous attempts to optimize the geometry. For these two systems, we relaxed the geometry using GFN-FF,[81] a polarizable force field designed for biological macromolecules.

Following structure relaxation, most crystallographic water molecules were removed except for those that were directly coordinated to the ligand or to ionic moieties. (All of the ligands are charge neutral, but most of the proteins contain at least one charged moiety.) Structures for 181L and 1LI2 contain two Cl^- ions each, and 3HUA contains a charged phosphate group. Within LIDS, the 1MMQ and 1ZP5 structures each contain four charged metal ions (Zn^{2+} and Ca^{2+}), with Zn^{2+} loosely coordinated to the ligand. Ionic cofactors were combined into a monomer with their nearest residues (within 2.5 Å) to improve monomer stability of the MBE(n) calculations and to reduce the number of fragments. For 1MMQ and 1ZP5, however, Zn^{2+} cannot be combined with the ligand because that would be incompatible with computing the interaction energy for removing the ligand. This has implications for the magnitude of the many-body effects in these systems, as discussed in Section 3.2.

2.3. Fragmentation. The MBE is a telescoping expansion for the total ground-state energy E, starting from fragment energies $\{E_I\}$ (for $I = 1, \ldots, N_{\text{frag}}$):

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

Two-body corrections are

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

where E_{IJ} is the energy of a dimer formed from fragments I and J. Similarly, the three-body corrections are

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

If eq. 1 is truncated at n-body terms, then we refer to the resulting method as MBE(n).

Following previous work,[25] we deconstruct proteins into single-residue fragments although we do not sever the polar peptide (C–N) bond.[22,57] Instead, fragments are constructed by cutting the C–C bond at C_α–C(=O).

https://doi.org/10.26434/chemrxiv-2024-7v7pv **ORCID:** https://orcid.org/0000-0002-1663-2278

We refer to the fragments as "monomers" and each consists of one amino acid along with the carbonyl moiety from the neighboring residue. More complicated algorithms for partitioning a protein into fragments have been suggested[22,82] but have not proven necessary to obtain the accuracy that we seek. Hydrogen-atom caps are used to saturate the severed valencies, as in previous work.[22,25] These capping atoms are positioned at

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

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the locations of the two carbon atoms in the original C–C bond. The quantities $R_1 = R_2 = 1.70$ Å and $R_{\text{H}} = 1.1$ Å are atomic van der Waals radii for carbon and hydrogen. This procedure is performed using FRAGME∩T.[30]

Each of the ligands considered in this work is retained as a single fragment. For the large ligands in Fig. 1b this may prove to be cost-prohibitive at levels of theory beyond DFT and will be revisited at a later time. In other applications of fragmentation to protein–ligand interaction energies, relatively small fragments have been used for both ligand and protein,[37–39,48] so there is reason to expect that ligand fragmentation is viable. In the present work, however, our goal is to establish that the enzymatic host can be effectively fragmented in a systematic manner that is conducive to obtaining interaction energies that are converged with respect to both the size of the enzyme model (N) and to the level of n-body interactions that are included. Fragmentation of the protein is relatively systematic whereas fragmentation of the ligand is less so, and we choose not to intermingle these two issues in the present work.

Absent some method to cull the number of subsystems, the combinatorial nature of MBE(n) quickly leads to an intractable number of small calculations since

$$N_{\text{sub}} \sim \binom{N_{\text{frag}}}{n} \sim N^n \ . \tag{5}$$

This combinatorial growth can lead to catastrophic loss-of-precision under some circumstances,[27–29] and in some cases fragment-based calculations may be more expensive than the supramolecular calculation they are intended to approximate.[17,23,29] In the present work, we use distance-based screening to reduce the number of subsystems for $n > 2$. Subsystems are eliminated if the minimum interatomic distance between any two fragments exceeds a specified threshold (R_{cut}) that we will vary systematically to test for convergence. In the case of 181L, where $N_{\text{frag}} = 164$, setting $R_{\text{cut}} = 8$ Å for MBE(3) reduces the number of subsystem calculations from 708,561 to 16,016, a 97.7% reduction. This makes higher-order n-body expansions feasible in large systems.[25,26] Enzymatic reaction energies and barrier heights converge quickly with respect to R_{cut}.[25] Energy-based screening (*e.g.*, with GFN2-xTB) can be even more efficient than distance-based screening[24] but was not fully implemented in FRAGME∩T when this work was undertaken.

Protein–ligand (P:L) interaction energies ΔE_{int} are computed via a supramolecular approach,

$$\Delta E_{\text{int}} = E_{\text{P:L}} - E_{\text{P}} - E_{\text{L}} \ , \tag{6}$$

by applying MBE(n) consistently to $E_{\text{P:L}}$ and E_{P}. A large number of subsystem calculations cancel in eq. 6 and can be eliminated *a priori*, as described elsewhere.[34] In principle, eq. 6 should be combined with counterpoise correction to eliminate basis-set superposition error (BSSE), which can be quite large for sizable protein–ligand models, especially if double-ζ basis sets are used.[83] Many-body counterpoise corrections that are consistent order-by-order with MBE(n) have been developed for this purpose[84,85] but are not yet available in FRAGME∩T. As a result, and because we are interested in demonstrating that our protocols are robust in large basis sets, we opt to push our calculations to the complete basis-set (CBS) limit using triple- and even quadruple-ζ basis sets.

Because we allow the amino acids to inhabit their native protonation states, leading to ionic side chains in some cases, there may be concern about long-range polarization interactions. The Zn^{2+} ion that is present in two of the LIDS proteins leads to especially large three- and four-body terms as discussed in Section 3.2. FRAGME∩T has the ability to add a low-level, ONIOM-style[86] supersystem correction for long-range polarization, with the subsystem MBE(n) calculations described at a higher level of theory.[25,26] Elsewhere, this procedure has been called a two-layer "molecules-in-molecules" approach (MIM2),[87] and it has been used by others under various names.[17,88–90] Applying this correction, the total energy for any given calculation, meaning any of the three terms in eq. 6, is

$$E_{\text{total}} = E_{\text{high}}^{\text{MBE}(n)} \underbrace{-E_{\text{low}}^{\text{MBE}(n)} + E_{\text{low}}^{\text{super}}}_{\delta_{\text{frag}}} \ . \tag{7}$$

The first two terms represent MBE(n) calculations at either the target (high) level of theory or else the affordable (low) level of theory. The final term ($E_{\text{low}}^{\text{super}}$) is the supersystem energy evaluated at the low level of theory with no fragmentation, thus δ_{frag} can be viewed as a low-level correction for the effects of fragmentation. In previous work on enzyme thermochemistry, the Hartree-Fock (HF)/6-31G method (sans polarization functions) was shown to be an adequate choice for $E_{\text{low}}^{\text{super}}$.[25] Use of the 6-31G basis set keeps the cost relatively low as compared to other double-ζ basis sets, especially if the electronic structure program can take advantage of compound *sp* shells used in Pople basis sets.[91,92] Due to the size of the enzyme models considered here, however, we will use HF-3c for the low level of theory in eq. 7; see Section 3.2.3.

2.4. Quantum Chemistry Calculations. Calculations were performed using FRAGME∩T[26,30] interfaced to Q-CHEM v. 6.0.[93] For timing data, calculations were

run on 28-core nodes (Dell Intel Xeon E5-2680 v4) using 7 worker processes per node, with each individual Q-Chem calculation employing 4 cores. Supersystem calculations were performed using a single 48-core node (Intel Xeon Platinum 8268). Timings will be reported in terms of aggregate computer time across all processors. The self-consistent field convergence threshold was set to 10^{-8} E_h for all calculations. Integral screening and shell-pair drop tolerances were both set to 10^{-12} a.u., consistent with recommendations for large-molecule calculations using diffuse basis sets.[92]

We use the ωB97X-V functional[94] as our target level of theory, as it performs well across a wide range of benchmarks including non-covalent interaction energies for small molecules,[19,95] where the benchmarks are well established. (For molecules with 100+ atoms, benchmarks are more uncertain.[19]) Minimally-augmented versions[91] of the Karlsruhe basis sets[96,97] are used for the ωB97X-V calculations. Diffuse functions can be important for non-covalent interaction energies but minimal augmentation is sufficient for this purpose.[83,91] The semi-empirical HF-3c model is used to evaluate convergence and all HF-3c calculations use the minimal "MINIX" basis set.[98] Since HF-3c is specifically parameterized for MINIX that basis set will not be mentioned in the discussion that follows, whereas for ωB97X-V we will systematically test the basis-set convergence.

A dielectric constant in the range $\varepsilon = 2$–4 is often used to represent the hydrophobic interior of a protein.[99–104] In previous work,[25] we found that a continuum solvation model with $\varepsilon \approx 4$ helps to avoid spurious oscillations in MBE(n) calculations, even for protein models with numerous charged residues where MBE(n) with vacuum boundary conditions does oscillate.[25] All calculations reported here use the conductor-like polarizable continuum model (C-PCM) with $\varepsilon = 4$.[105,106] The interface with the continuum region is represented using a van der Waals cavity,[106] constructed from atomic radii that are 1.2× larger than values in the modified Bondi set,[106,107] then discretized using the switching/Gaussian procedure.[105,108–110] For ωB97X-V calculations, 110 Lebedev points were used for hydrogen and 194 points for other nuclei. For HF-3c, we used 50 points for hydrogen and 110 points for other nuclei. For supersystem calculations involving the entire protein, a conjugate gradient implementation of C-PCM was used.[110]

3 Results and Discussion

3.1. T4 Lysozyme Data Set. The primary goal of this work is to develop reliable and reproducible protocols that afford energetically converged protein–ligand models that are usable across different levels of electronic structure theory. In the present work, we use HF-3c to test convergence with respect to model size but we demonstrate ΔE_{int} calculations using ωB97X-V in basis sets up to def2-ma-QZVP.[91]

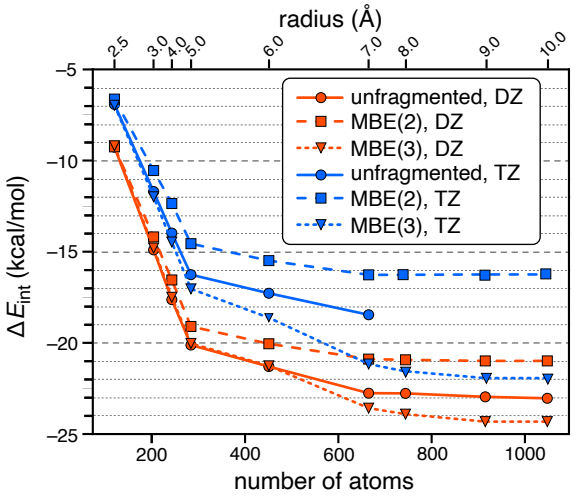

Fig. 2: Interaction energy ΔE_{int} for the benzene ligand in 181L, computed using radial enzyme models of increasing size. Model size in number of atoms is indicated along the bottom axis while the top axis shows the radius used to generate each model. Calculations were performed using ωB97X-V in conjunction with either the def2-ma-SVP basis set (labeled "DZ" in the figure) or else the the def2-ma-TZVP basis set ("TZ"). All MBE(n) calculations use $R_{\text{cut}} = 8$ Å. The "unfragmented" result is a conventional supramolecular calculation of ΔE_{int}.

3.1.1. Radial Enzyme Models. First, we study convergence of ΔE_{int} with respect to the size of the enzyme model using both MBE(n) and conventional supramolecular calculations. We created reduced models of 181L (which has 2,636 total atoms) based on a simple radial cutoff around the benzene ligand, then computed ΔE_{int} using each model for comparison to the result obtained using the complete 2,636-atom protein. (In follow-up work, we may consider the use of automated construction of binding-site models using residue interaction networks,[111–115] but here we use unsophisticated radial models as a simple means to establish protocols.)

Figure 2 shows how the results converge with respect to the size of the enzyme model, for both conventional supramolecular DFT and also for MBE(2) and MBE(3) approximations to it. Conventional calculations at the ωB97X-V/def2-ma-SVP level are converged using a model with 665 atoms, corresponding to all residues within 7 Å of the ligand. Resource limitations preclude ωB97X-V/def2-ma-TZVP calculations for models larger than this, but the convergence behavior for smaller models seems to mirror that obtained using the double-ζ basis set so we expect that ωB97X-V/def2-ma-TZVP calculations are also converged for the 665-atom model. Analogous testing was completed for 1LI2, which has phenol as a ligand, and the convergence behavior is very similar (Fig. S2). In that case, a 7 Å (617-atom) model affords affords a converged value of ΔE_{int} at the ωB97X-V/def2-

ma-SVP level, while convergence of the ωB97X-V/def2-ma-TZVP calculations looks similar.

Examining the MBE(n) results in Fig. 2, we observe that MBE(2) consistently underestimates $|\Delta E_{\text{int}}|$ in both basis sets, due to missing nonadditive polarization. MBE(3) calculations overestimate $|\Delta E_{\text{int}}|$. In the def2-ma-SVP basis set where we are able to demonstrate convincing convergence with respect to system size, the two-body result is underbound by about 2 kcal/mol while the three-body result is overbound by about 1 kcal/mol, as compared to a conventional calculation at the same level of theory. For models larger than 700 atoms, convergence of the MBE(2) and MBE(3) approximations to ωB97X-V/def2-ma-SVP track the conventional result quite well, albeit with constant offsets. For small models, however, that offset is masked and both the MBE(2) and MBE(3) results are in fortuitously good agreement with conventional supramolecular calculations. Unless these studies are pushed to the $N \to \infty$ limit, one might erroneously conclude that three-body effects are unimportant. Furthermore, a 200-atom model affords an interaction energy $|\Delta E_{\text{int}}|$ that is 10 kcal/mol smaller than the converged value!

If the triple-ζ calculations are indeed converged at the 665-atom model, then the MBE(2) calculations are are underbound in that case by about 2 kcal/mol while MBE(3) calculations are overbound by perhaps 3 kcal/mol. Regardless of where the converged triple-ζ result for ΔE_{int} may lie, we can state that MBE(2) and MBE(3) estimates bracket the conventional value by about 3.5 kcal/mol at the double-ζ level versus ≈ 5.5 kcal/mol at the triple-ζ level. These are large ranges by the standards of benchmark accuracy in small-molecule quantum chemistry calculations and it is not immediately clear whether this level of agreement is acceptable. That issue is taken up in Section 3.1.4, where we discuss the appropriate level of accuracy for large-scale electronic structure calculations of protein–ligand interaction energies. Before that, however, we examine convergence of the radial enzyme models with respect to both model size (N) and level of approximation (n), in Section 3.1.2. Basis-set convergence is examined in Section 3.1.3.

3.1.2. Convergence with N and n. To understand what is required in order to obtain converged values of ΔE_{int}, we first study the behavior of MBE(2) as a function of R_{cut}. Results for the 2,636-atom 181L system that was considered in Section 3.1.1 are shown in Fig. 3, examining how MBE(2) converges with respect to R_{cut}, the distance threshold for discarding subsystems. Previously,[25] we showed that $R_{\text{cut}} = 8$ Å affords converged thermochemistry for a 632-atom model of a different enzyme, exploring several functional and basis-set combinations. (This 632-atom model affords converged results with respect to larger radial models of the same enzyme.[116]) The value $R_{\text{cut}} = 8$ Å also works well here. Increasing it to

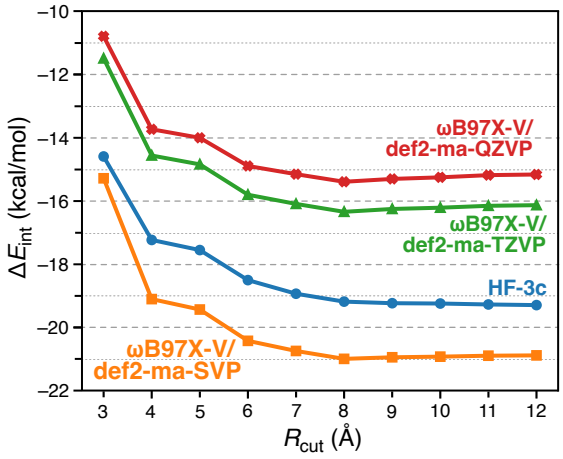

Fig. 3: Interaction energies ΔE_{int} for the benzene ligand in 181L, computed using MBE(2) at the indicated levels of theory, using a distance cutoff (R_{cut}) to cull the dimers that are included in the calculation.

10 Å changes ΔE_{int} by ∼ 0.1 kcal/mol but increases the requisite number of subsystem calculations from 2,534 to 3,496. Convergence behavior for the other three T4-lysozyme systems is quite similar (see Fig. S1 and Table S2) and converged MBE(2) interaction energies are obtained using $R_{\text{cut}} = 8$ Å in those cases as well.

Convergence behavior as a function of R_{cut} is also quite similar for the minimal-basis HF-3c method as compared to ωB97X-V calculations in basis sets ranging from def2-ma-SVP to def2-ma-QZVP, although the converged values of ΔE_{int} certainly differ in each case. Interaction energies computed at the ωB97X-V/def2-ma-QZVP level, which ought to lie close to the ωB97X-V/CBS limit, differ by about 6 kcal/mol from double-ζ results. Most of the overbinding in the latter case is likely a BSSE artifact and is clearly not negligible even though DFT/double-ζ interaction energies are still (much too) widely used in biomolecular ΔE_{int} calculations.[46,117] Implementation of many body counterpoise corrections[84,85] within FRAGME∩T is underway, and should provide better-converged results in smaller basis sets.

For now, we can use def2-ma-QZVP to establish the basis-set limit.[83] This reveals that HF-3c results are closer to ωB97X-V/def2-ma-SVP than they are to ωB97X-V/CBS. However, the minimal-basis HF-3c method can be run in a tiny fraction of the computational cost, meaning tens of hours for HF-3c versus hundreds of hours for ωB97X-V/def2-ma-SVP, or ∼ 10^4 h for ωB97X-V/def2-ma-QZVP.

The smallest of the T4-lysozyme systems contains 2,636 atoms, which taxes our ability to perform supersystem benchmarks using high-quality basis sets. In an effort to obtain more convergence data, we turn to HF-3c where supersystem calculations are more feasible. Results in Table 1 demonstrate that MBE(2) estimates of ΔE_{int} consistently fall within 1 kcal/mol of the conventional supramolecular result obtained at the same level

Table 1: HF-3c Results for T4 Lysozymes.

System	Method	energy (kcal/mol)		CPU time (hours)[a]
		ΔE_{int}	error per monomer	
181L	supersys.[b]	−19.4	—	4,156
	MBE(2)[c]	−19.1	0.002	21
	MBE(3)[c]	−18.8	0.004	298
4W54	supersys.[b]	−25.6	—	2,576
	MBE(2)[c]	−26.6	0.006	21
	MBE(3)[c]	−26.2	0.004	300
1LI2	supersys.[b]	−18.8	—	5,542
	MBE(2)[c]	−19.8	0.006	21
	MBE(3)[c]	−19.5	0.004	303
3HUA	supersys.[b]	−30.0	—	2,313
	MBE(2)[c]	−30.5	0.003	25
	MBE(3)[c]	−30.2	0.002	351

[a]Hardware is described in Section 2.4. [b]Conventional (unfragmented) calculation. Timings include $E_{\text{P:L}}$, E_{P}, and E_{L}. [c]MBE(n) calculations use $R_{\text{cut}} = 8$ Å.

Table 2: MBE(2) Calculations for T4 Lysozymes Using ωB97X-V in Various Basis Sets.[a]

System	Basis Set	ΔE_{int} (kcal/mol)	CPU time (hours)
181L	def2-ma-SVP	−20.99	763
	def2-ma-TZVP	−16.34	2,543
	def2-ma-QZVP	−15.39	34,474
4W54	def2-ma-SVP	−30.52	778
	def2-ma-TZVP	−24.93	2,749
	def2-ma-QZVP	−23.69	36,030
1LI2	def2-ma-SVP	−23.14	771
	def2-ma-TZVP	−18.00	2,585
	def2-ma-QZVP	−16.76	34,941
3HUA	def2-ma-SVP	−34.73	924
	def2-ma-TZVP	−27.25	3,246
	def2-ma-QZVP	−25.42	42,449

[a]All calculations use $R_{\text{cut}} = 8$ Å.

of theory. MBE(2) and MBE(3) values of ΔE_{int} differ by only 0.3 kcal/mol on average, but the latter are more than 10× more expensive, even with $R_{\text{cut}} = 8$ Å. Timing data in Table 1 suggest that the MBE(2) cost is 1% or less of the conventional cost to compute ΔE_{int}. This overstates the MBE(2) cost somewhat, because the full-system calculations were performed on slightly newer hardware as described in Section 2.4.

It is common in fragment-based calculations to report errors on a per-monomer basis, recognizing that overall errors may be size-extensive. A target accuracy of $0.1 \times (3/2)k_{\text{B}}T$, or ≈ 0.1 kcal/mol at $T = 298$ K, has been suggested.[118] This threshold represents 10% of the available thermal energy per fragment, with the idea that fragmentation errors should be rendered negligible in comparison to thermal fluctuations in the energy. While it's not clear that this is the right target accuracy for biomolecular ΔE_{int} calculations, our T4-lysozyme calculations do achieve this stringent criterion: the largest errors in Table 1 are only 0.006 kcal/mol/monomer.

For these examples. which involve very small ligands, a two-body calculation with no electrostatic embedding at all meets the highest-fidelity standard for fragmentation, while keeping the cost extremely low in comparison to fully-converged supramolecular calculations. Avoiding embedding renders these calculations stable in large basis sets, including basis sets that contain diffuse functions. This will be important for future work, where we intend to push fragmentation to levels of theory beyond DFT. As such, we next take a closer look at basis-set convergence.

3.1.3. Basis-Set Convergence. Having used HF-3c to establish that MBE(2) provides reliably converged interaction energies (which is not the same as *accurate* interaction energies), we next examine MBE(2) calculations using ωB97X-V in various basis sets; see Table 2. The value of $|\Delta E_{\text{int}}|$ is reduced as the basis set is enlarged, consistent with a reduction in BSSE, and we expect that DFT/def2-ma-QZVP results lie near the DFT/CBS limit even without counterpoise correction.[83] Smaller models of 181L and 1LI2 were examined a previous study,[83] where it was concluded that ωB97M-V/def2-ma-QZVP calculations without counterpoise correction were within 0.2 kcal/mol of the ωB97M-V/CBS limit. For comparison, uncorrected ωB97M-V/def2-ma-TZVP calculations were 1.1–1.7 kcal/mol from the CBS limit, erring towards overbinding, while uncorrected ωB97M-V/def2-ma-SVP calculations overestimated $|\Delta E_{\text{int}}|$ 5.0–6.1 kcal/mol as compared to the ωB97M-V/CBS limit.[83] These results from Ref. 83 use a different functional (ωB97M-V) as compared to that used here (ωB97X-V), which is unlikely to affect convergence to the CBS limit, but they correspond to small (5 Å) models with less than 300 atoms so that the BSSE is likely somewhat smaller than it is in the present calculations.

3.1.4. Discussion. The def2-ma-QZVP results in Table 2 are certainly converged well enough to conclude that single-pose interaction energies (ΔE_{int}) obtained with high-quality DFT are considerably larger in magnitude than the *free* energies of binding ($\Delta G^{\circ}_{\text{bind}}$) that are measured experimentally, the latter of which range from $\Delta G^{\circ}_{\text{bind}} = -4.6$ kcal/mol to $\Delta G^{\circ}_{\text{bind}} = -5.8$ kcal/mol for the T4-lysozyme data set. The same observation has been made in full-protein DFT calculations.[35,36] In particular, single-pose interaction energies for 1LI2, computed at the PBE+D level, are on the order of $\Delta E_{\text{int}} = -28$ kcal/mol,[36] somewhat larger in magnitude than the ωB97X-V/def2-ma-SVPD value reported in Table 2.

The difference between a single-pose ΔE_{int} and $\Delta G^{\circ}_{\text{bind}}$ can be partitioned into several different contributions.[36]

These include conformational averaging (which is not included in the present work), the differential solvation energy between the protein–ligand complex and its separated constituents (denoted $\Delta G^{\circ}_{\text{solvn}}$ below), and finally the change intramolecular vibrational entropy ($-T\Delta S_{\text{vib}}$ where ΔS_{vib} is the change in vibrational entropy upon complexation). Following Ref. 36, one may express the free energy according to

$$\langle G \rangle = \langle E \rangle + \langle G_{\text{solvn}} \rangle - T\Delta S_{\text{vib}} \tag{8}$$

where $\langle \cdots \rangle$ represents conformational averaging. Then the free energy for ligand binding can be expressed as

$$\Delta G^{\circ}_{\text{bind}} = \langle \Delta E_{\text{int}} \rangle + \Delta G^{\circ}_{\text{solvn}} - T\Delta S_{\text{vib}} \; . \tag{9}$$

The quantity $\langle \Delta E_{\text{int}} \rangle$ represents the interaction energy averaged over a molecular dynamics trajectory, and results for 1LI2 indicate that $\langle \Delta E_{\text{int}} \rangle$ converges in fewer than 100 snapshots.[35,36] The correction $\Delta G^{\circ}_{\text{solvn}}$ be estimated using implicit solvation models that are compatible with large-scale electronic structure calculations.[106,119] Finally, ΔS_{vib} can be computed from DFT (or perhaps semi-empirical) vibrational frequency calculations.[36,120,121] These corrections are not included in the present work, however, as our focus is to establish fragment-based protocols to compute ΔE_{int}. As such, we do not expect to recover $\Delta G^{\circ}_{\text{bind}}$ in these calculations.

In selecting between DFT and semi-empirical methods, or between double- and triple-ζ DFT methods, it is worth considering what level of accuracy is required from the calculations at hand. Convergence of ensemble averages $\langle \Delta E_{\text{int}} \rangle$ using single-pose interaction energies ΔE_{int} appears to be rapid, using classical molecular dynamics to sample structures,[35,36] yet the result will not approximate $\Delta G^{\circ}_{\text{bind}}$ without a calculation of S_{vib}. The latter requires vibrational frequency calculations, as SASA-dependent corrections are insufficient to bridge the quantitative gap between $\langle \Delta E_{\text{int}} \rangle$ and $\Delta G^{\circ}_{\text{bind}}$.[38,39] As compared to the disparity between these two values, changes in ΔE_{int} with respect to n-body order are small.

That said, prior fragment-based DFT calculations of single-pose interaction energies (as in the present work) have established that these ΔE_{int} values exhibit remarkably good *correlations* with experimental $\Delta G^{\circ}_{\text{bind}}$ values,[38,39,48] even while they differ by an order-of-magnitude in absolute value. In some cases, very simple SASA-dependent entropy corrections[122,123] have been added to fragment-based DFT calculations of ΔE_{int}.[38,39] In other cases, however, good correlations are observed even without such a correction.[48] For the purpose of obtaining training data for ML, direct correlation with experiment is not the most important consideration; sampling, solvation, and entropic corrections can be added later, using a low-cost ML force field trained on ΔE_{int} values from electronic structure calculations.[124,125] What is more important is obtaining high-quality quantum-chemical benchmark data.

For that purpose, the computational efficiency of MBE(2) at the DFT/def2-ma-TZVP level presents a compelling advantage. Such calculations constitute less than 10% of the cost of MBE(2) using def2-ma-QZVP, yet afford interaction energies that differ by ~ 1 kcal/mol from what is likely the DFT/CBS limit. Even that difference may very well disappear once many-body counterpoise corrections are incorporated.[83] Moreover, the ωB97X-V/def2-ma-TZVP calculations using MBE(2) require only about half the computer time that is required for an *unfragmented* (full-system) calculation at the minimal-basis HF-3c level. The former do require 2,500–3,200 h of computer time, which is not a trivial investment. However, wall times can be significantly reduced by exploiting the inherent parallelizability of the pairwise MBE(2) approach. For example, the 181L system consists of 2,534 dimers when $R_{\text{cut}} = 8$ Å, and each of these calculations is completely independent of the others. This makes MBE(2)-based DFT/triple-ζ calculations an attractive choice if a realistic value of ΔE_{int} is sought.

At the same time (and for the same reason), fragmentation enables large-scale quantum chemistry calculations using modest hardware, which is an important consideration in making these approaches accessible to investigators at under-resourced institutions. As an example, consider that a full-system calculation on 181L (2,636 atoms) means 50,558 basis functions for def2-ma-TZVP, or 116,483 basis functions for def2-ma-QZVP. Even the triple-ζ calculation lies outside the realm of single-node (workstation) computing, requiring supercomputer resources that are not available to everyone. In contrast, low-order MBE(n) remains feasible on workstation hardware even for the large enzyme models considered here. The present calculations represent some of the largest applications to date of DFT used to compute protein–ligand binding using full (or at least, converged) protein models. Such studies have been carried out recently using semilocal DFT and a full T4-lysozyme protein,[36] using a linear-scaling DFT code.[126] This requires supercomputer resources,[36] whereas all calculations reported here exploit only single-node parallelism.

The target fidelity of $0.1 \times (3/2)k_{\text{B}}T$ that was discussed in Section 3.1.2 is a stringent criterion posited with an eye toward *ab initio* molecular dynamics studies using fragmentation.[118] That approach is complicated by the complexity of analytic gradients in the presence of charge embedding,[17,127] and likely unnecessary since force fields or semi-empirical quantum chemistry can be used to better and (much) more efficiently sample the conformational space. Thus, it is worth asking what eventual purpose fragment-based *ab initio* calculations of protein–ligand binding will serve, and what level of accuracy and convergence is necessitated by that application. We do not have a simple answer to that question but it's probably safe to assume that for ML, one desires a method that accurately reflects the interaction potential for short-range protein–ligand interactions, *e.g.*, to replace docking models[8–14] or classical force fields.[125,128] For that purpose, the highest-quality *ab initio* interac-

tion energies may not be necessary and DFT or even semi-empirical calculations might suffice.

That said, we do worry that the def2-ma-SVP affords interaction energies that are too far removed from those obtained in higher-quality basis sets, and that the BSSE inherent in double-ζ calculations may skew the conformational landscape towards compact structures, which exhibit larger BSSE and thus ostensibly stronger interactions in small-basis calculations.[129–132] It is also worth considering that the accuracy of DFT for small-molecule van der Waals complexes does not seem to extend to complexes in the 150-atom size regime,[19] so the quality of supramolecular DFT "benchmarks" is uncertain in sizable protein–ligand models. These are important questions to explore in future work. For now, we simply note that the conventional "chemical accuracy" standard of 1 kcal/mol may be overly conservative for the present purpose.

3.2. Large Inhibitor Data Set. The T4-lysozyme data set was a useful starting point to establish best practices for large systems with small ligands. Good accuracy for very small ligands is important in order to meet the requirements of fragment-based approaches to drug design and discovery,[133–138] which search for "hits" based on small-molecule probes rather than larger ligand models that resemble existing drugs. This strategy has been suggested as a salve to remedy a slow drift towards drug candidates with larger and larger molecular weight,[139] a trend that has been blamed for increased attrition rates in clinical trials.[140–143] That said, with an eye towards computational investigation of *existing* drug molecules, or structure-based drug design, it is important to understand how fragmentation protocols fare for much larger ligands, epitomized by those in Fig. 1b. Each of these ligands is larger than 40 atoms and there is also more variety in the enzymatic targets as compared to the T4-lysozyme data set.

3.2.1. Radial Enzyme Models. In the T4 lysozymes, errors associated with fragmentation appeared to stabilize as the size of the model system increased (Fig. 2). Here, we perform analogous testing using 1O48 where the ligand is much larger. Figure 4 plots the results for a sequence of radial models of increasing size, comparing MBE(2) and MBE(3) calculations to unfragmented (supramolecular) values of ΔE_{int} computed at the same level of theory, namely, DFT in either a double- or a triple-ζ basis set. The largest model in Fig. 4 uses a 6 Å radius and contains 619 atoms but convergence to within 1 kcal/mol is achieved using a 3 Å model with 381 atoms.

In the smallest models, MBE(3) yields a marginally smaller interaction energy as compared to a full-system calculation but converges to the full-system result in larger models. For the largest model (619 atoms), the difference in ΔE_{int} with respect to the full-system calculation is 1.3 kcal/mol for MBE(2) and 0.4 kcal/mol for MBE(3), at the ωB97X-V/def2-ma-SVP level. If the def2-ma-TZVP basis set is used instead, the error is 1.0 kcal/mol for MBE(2) and 1.7 kcal/mol for MBE(3). Strictly speaking, MBE(3) is less accurate than MBE(2) for the largest enzyme model and basis set, but the difference with respect to MBE(2) is only a tiny fraction ($< 1\%$) of $\Delta E_{\text{int}} = -80$ kcal/mol. For all practical purposes, we conclude that MBE(3) provides negligible improvement upon MBE(2) in this case.

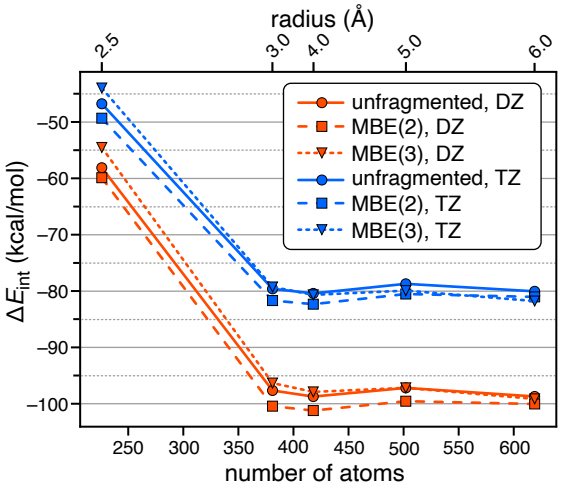

Fig. 4: Interaction energy ΔE_{int} for the benzene ligand in 1O48, computed using radial enzyme models of increasing size. Model size in number of atoms is indicated along the bottom axis while the top axis shows the radius used to generate each model. Calculations were performed using ωB97X-V with either the def2-ma-SVP basis set ("DZ") or else the def2-ma-TZVP basis set ("TZ"). MBE(n) calculations used $R_{\text{cut}} = 8$ Å. The "unfragmented" result is a conventional supramolecular calculation of ΔE_{int}.

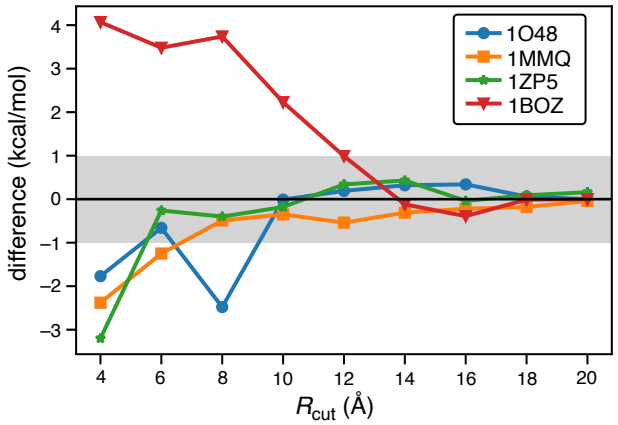

Fig. 5: Difference in ΔE_{int} for the large-inhibitor complexes, based on MBE(2) calculations at the HF-3c level as a function of R_{cut}. The baseline calculation is MBE(2) with no cutoff and the shaded region indicates ±1 kcal/mol with respect to that baseline.

Table 3: Summary of HF-3c Calculations for the Large-Ligand Systems.

System	Method	energy (kcal/mol)		CPU time (h)
		ΔE_{int}	error per monomer	
1O48	supersys.[a]	−89.9	—	854
	MBE(2)[b]	−94.8	0.05	21
	MBE(3)[b]	−91.3	0.01	366
	MBE(4)[b]	−88.9	0.01	3,045
1MMQ	supersys.[a]	−178.6	—	3,138
	MBE(2)[c]	−157.2	0.13	41
	MBE(3)[c]	−179.6	0.01	534
	MBE(4)[c]	−178.4	0.00	3,437
1ZP5	supersys.[a]	−108.7	—	5,619
	MBE(2)[c]	−108.7	0.00	45
	MBE(3)[c]	−80.0	0.18	654
	MBE(4)[c]	−105.0	0.02	4,595
1BOZ	supersys.[a]	−31.3	—	5,018
	MBE(2)[d]	−34.1	0.01	93
	MBE(3)[d]	−34.8	0.02	2,857
	MBE(4)[d]	−11.2	0.11	46,276

[a] Conventional supramolecular calculation. [b] $R_{\text{cut}} = 9$ Å. [c] $R_{\text{cut}} = 8$ Å. [d] $R_{\text{cut}} = 12$ Å.

3.2.2. Convergence of the MBE. Calculations on the T4 lysozymes reveal that HF-3c and ωB97X-V exhibit similar convergence behavior as a function of R_{cut}, so in what follows we use the much cheaper HF-3c method to examine convergence for the large-inhibitor models. Figure 5 plots the convergence behavior of MBE(2) calculations as a function of R_{cut}, relative to a baseline where all dimers are retained. That full-MBE(2) limit is obtained, to within 0.2 kcal/mol, when $R_{\text{cut}} = 20$ Å. Adopting a more permissive 1 kcal/mol tolerance, we can use $R_{\text{cut}} = 9$ Å for 1O48, $R_{\text{cut}} = 8$ Å for 1MMQ and 1ZP5, and $R_{\text{cut}} = 12$ Å for 1BOZ. (See Table S3 for the numerical data.) Notably, the 1MMQ and 1ZP5 systems are approximately the same size as the T4 lysozymes that were also converged by $R_{\text{cut}} = 8$ Å, whereas 1BOZ is the largest system considered here, and it exhibits the slowest convergence as measured by R_{cut}. In future studies of new enzymes, two-body screening at a semi-empirical level of theory may offer a way to test convergence at only modest cost, and this is an avenue that we are currently pursuing.

Using the aforementioned system-specific R_{cut} values, we next examine how MBE(n) converges towards the supersystem result, again using HF-3c calculations, with results up to $n = 4$ presented in Table 3. We extended these calculations to the four-body level because the two-body accuracy is inferior to what we observed for the T4-lysozyme data set. For example, in the case of 1MMQ the accuracy of MBE(2) lies outside of our strict 0.1 kcal/mol/monomer tolerance but that is rectified at the three-body level, and MBE(3) calculations also noticeably improve the result for 1O48 as well. The

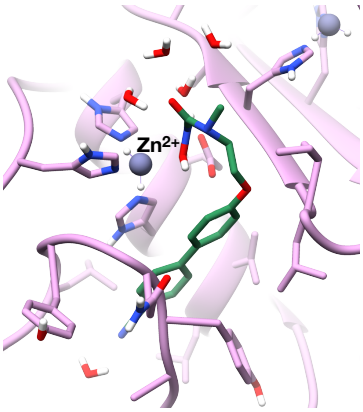

Fig. 6: Active site of 1ZP5 with Zn^{2+} labeled. The two hydrogen atoms shown as coordinated to Zn^{2+} are from the neighboring histidine residues. A second Zn^{2+} (not in the active site) is visible at the top right.

three-body terms provide little change for 1BOZ, where the MBE(2) result is already within 3 kcal/mol of the supersystem calculation. However, MBE(3) calculations for 1ZP5 and MBE(4) calculations for 1BOZ afford very large errors, which we next examine.

For 1ZP5, MBE(2) is fortuitously accuracy but MBE(3) is much worse, deviating from the full-system value of ΔE_{int} by 29 kcal/mol. That observation, combined with the dramatic failure of MBE(2) for 1MMQ (21 kcal/mol error) hints at additional complexities for these two metalloproteins. Both systems feature Zn^{2+} coordinated to the inhibitor ligand (as shown in Fig. 6 for 1ZP5), leading to sizable many-body polarization.

Histograms of all three-body interactions 1ZP5 (Fig. S4) exhibit a few significant outliers where $|\Delta E_{IJK}| \gtrsim 0.01\ E_h$. The largest of these three-body contributions is about 30 kcal/mol. The metalloenzyme 1MMQ has one very large three-body interaction of 23 kcal/mol (Fig. S5), whereas for 1O48 and 1BOZ the three-body corrections are much smaller (Figs. S3 and S6). Inspecting the fragments that give rise to the outliers in 1ZP5 and 1MMQ confirms that each contains the Zn^{2+} cation, the ligand, and a residue whose side chain is coordinated to the ligand. For example, the large three body term for 1MMQ contains GLU121, which is coordinated to the carboxylic acid group in the binding site. While it is not surprising that a divalent cation engenders significant nonadditive polarization, these results underscore the fact that MBE(2) is not always a good approximation for protein–ligand interaction energies.

Inclusion of four-body terms affords notable improvement in the case of 1ZP5, reducing the error from almost 30 kcal/mol to less than 4 kcal/mol, albeit at significant computational expense. Nevertheless, this demonstrates what seems to be convergent MBE(n) results for the two Zn^{2+}-containing enzymes although MBE(4) calculations for 1MMQ are more expensive than the conventional supersystem calculation. Elsewhere, we have addressed this

problem using energy-based screening,[24,26] but that has not been attempted for these systems. MBE(3) remains significantly cheaper than a full-system calculation for these systems, and MBE(2) is at least one order of magnitude less expensive still. This points to the need for future work in which a limited set of energetically important four-body terms might be included in order to recover comparable accuracy at greatly reduced cost, and we are presently implementing this capability within the FRAGME∩T code.

The need for additional screening is dramatically underscored by MBE(4) results for 1BOZ, the largest enzyme considered here. For this system, the difference between a 8 Å cutoff and a 12 Å cutoff increases the number of tetramers from 82,975 to 758,495 and increases the error from 0.5 to 20.1 kcal/mol. The MBE(4) entry in Table 3 represents the larger cutoff radius, which two-body calculations suggest is required to reach convergence, but using that cutoff results in a MBE(4) estimate of ΔE_{int} that is substantially less accurate than the MBE(3) estimate. Meanwhile, MBE(4) results with an 8 Å cutoff afford $\Delta E_{\text{int}} = -31.8$ kcal/mol for 1BOZ, which within 0.5 kcal/mol of the supersystem result.

Thus, 1BOZ appears to be a case where cumulative errors from a enormous number of subsystem calculations skews the result, a possibility that we have noted previously.[27,28] A histogram of four-body terms for 1BOZ with 12 Å cutoff can be found in Fig. S7. Energetically, these terms are two orders of magnitude smaller than the three-body terms (shown in Fig. S6), but considerably more numerous. Incorporation of bottom-up energy-based screening[26] into our enzymatic fragmentation protocol will be essential for routine application of MBE(4) to model systems of this size.

3.2.3. Two-Layer Approach. The need for four-body terms drastically increases the requisite number of subsystem calculations. Automatic energy screening using GFN2-xTB,[26,34] which was developed in parallel with this work, should ultimately be useful in this regard. As an alternative, we examine the use of a two-layer supersystem correction (eq. 7). For the calculation of thermochemical quantities in enzymes, this approach worked quite well when used with only a two-body expansion at the higher level of theory, and low-level methods such as HF/6-31G.[25] Versions of this "MIM2" procedure (and also a three-layer "MIM3") have been used in other fragment-based calculations of protein–ligand interaction energies,[37–39,48] typically using the semi-empirical PM6+D3[144,145] method for the low-level correction, and with smaller (sub-residue) fragments as compared to the present work. Here, we examine the efficacy of using HF-3c to compute δ_{frag} in eq. 7, since we know that HF-3c is computationally feasible in very large enzyme models.

Although the overall errors in ΔE_{int} are larger for the large-ligand data set, the system sizes are also larger so error per monomer becomes a useful point of comparison. With the incorporation of four-body terms that metric achieves the strict criterion of 0.1 kcal/mol/monomer, but at the same time the number of subsystem calculations becomes nearly intractable for any level of theory beyond semi-empirical calculations. Large three-body terms in 1ZP5 cause require higher-order expansions but in all of the remaining systems, the error per monomer is below the target accuracy already at the MBE(2) level. In other fragment-based studies of metalloproteins, the fragment that includes the metal ion cofactor typically contains all proximal molecules (side chains and crystallographic water molecules).[25,146] However, this is not possible in the case of the metalloproteins investigated here because the ligand coordinates to the metal ion, and since the ligand must be removed in order to compute ΔE_{int}, no atoms from the enzyme can be included in the ligand fragment(s).

With this in mind, a supersystem HF-3c correction has been applied to all MBE(n) calculations for the large inhibitors, with results listed in Table 4. These calculations use the full enzyme although MBE(n) calculations are applied with $R_{\text{cut}} = 8$ Å rather than the system-specific cutoffs used in Table 3. The number of subsystems generally doubles for every 2 Å that is added to R_{cut}; for example, in the case of 1BOZ at the MBE(3) level we obtain 22,001 subsystems for $R_{\text{cut}} = 8$ Å, 45,433 subsystems for $R_{\text{cut}} = 10$ Å, and 88,811 subsystems for $R_{\text{cut}} = 12$ Å. To perform MBE(3) at the ωB97X-V/def2-ma-QZVP level requires $\sim 1.1 \times 10^6$ h of computer time to complete for the largest of these systems, 1BOZ.

Supersystem-corrected MBE(2) + δ_{frag} interaction energies for 1MMQ are quite close to those obtained using uncorrected MBE(3), but at a fraction of the cost. Results for the other systems do not align quite as well. The average difference between MBE(3) and MBE(2) + δ_{frag} estimates of ΔE_{int} is ≈ 20 kcal/mol across all basis sets, although the difference increases marginally with increasing basis set size. The supersystem correction actually increases the disparity between MBE(2) and MBE(3) in several cases. In these cases, it seems that the two-layer approach is no substitute for MBE(3), at least when the supersystem correction is performed using the minimal-basis HF-3c model.

In view of the success of MIM3 methods using PM6+D3 as the lowest level of theory (and B97+D3/6-311++G** as the highest),[37–39] it is worth considering whether alternative semi-empirical models would fare better. In previous work on enzyme thermochemistry,[25] supersystem correction computed using either HF-3c or HF/6-31G afforded nearly identical results. The performance of the PBEh-3c model,[147] which uses a double-ζ basis set rather than a minimal one, was also comparable.[25] As such, it is perhaps more beneficial to work on ways to reduce the cost of MBE(3) calculations via screening, rather than cycle through a long list of low-cost electronic structure methods that could be used for the supersystem correction, with no clear physical reason why one performs better than others.

Table 4: Interaction Energies for the Large-Ligand Complexes Computed using ωB97X-V.[a]

System	Basis Set	MBE(2)		MBE(3)		MBE(2) + δ_{frag}[b]	
		ΔE_{int} (kcal/mol)	CPU time (hours)	ΔE_{int} (kcal/mol)	CPU time (hours)	ΔE_{int} (kcal/mol)	CPU time (hours)[c]
1O48	def2-ma-SVP	−101.6	629	−103.6	8,736	−97.8	1,500
	def2-ma-TZVP	−82.5	2,440	−86.3	35,642	−78.8	3,311
	def2-ma-QZVP	−80.6	32,032	−85.6	470,973	−76.8	32,903
1MMQ	def2-ma-SVP	−132.1	1,150	−154.0	15,885	−153.4	4,329
	def2-ma-TZVP	−118.7	4,350	−142.4	62,962	−140.0	7,529
	def2-ma-QZVP	−118.0	53,052	−142.9	776,176	−139.4	56,231
1ZP5	def2-ma-SVP	−110.1	1,285	−77.5	19,827	−110.1	4,939
	def2-ma-TZVP	−94.2	5,005	−61.5	81,679	−94.1	8,658
	def2-ma-QZVP	−91.8	60,585	−59.9	1,004,143	−91.8	64,238
1BOZ	def2-ma-SVP	−53.6	1,349	−68.0	22,942	−48.1	6,405
	def2-ma-TZVP	−34.1	5,351	−51.1	92,890	−28.6	10,407
	def2-ma-QZVP	−30.4	68,310	−48.9	1,128,189	−24.9	73,366

[a]MBE(n) calculations use $R_{\text{cut}} = 8$ Å . [b]Using HF-3c to evaluate δ_{frag} in eq. 7. [c]Includes the cost to compute δ_{frag} (Table S4).

Basis set trends are similar to what is observed for the T4-lysozyme data set, with a reduction in BSSE as the basis-set quality improves leading to a reduction in $|\Delta E_{\text{int}}|$. Numerical values change much more dramatically than they did for the T4 lysozymes, however, because the much larger LIDS ligands engender larger BSSE, which increases with system size because the number of neighbor atoms increases.[83] For example, swapping def2-ma-SVP for def2-ma-QZVP changes ΔE_{int} by an average of 7.0 kcal/mol for the T4 lysozyme data set but for LIDS the change is a staggering 19.2 kcal/mol. For the large ligands, results obtained using the def2-ma-SVP basis set seem inappropriate to use. This is important information given that many DFT calculations of protein–ligand interactions continue to use double-ζ basis sets for reasons of cost. Many-body counterpoise corrections may facilitate the use of double-ζ basis sets and we intend to explore this in future work.

4 Conclusions

Fragment-based approximations provide the means to address dramatically larger systems sizes using quantum chemistry calculations. In recent work,[25,26,34] we have pursued a stripped-down n-body expansion that makes no attempt at classical electrostatic embedding, as a robust means to converge fragment-based calculations to a well-defined supersystem limit, at essentially arbitrary levels of electronic structure theory including arbitrary basis sets. In the present work, we have extended this approach to protein–ligand interaction energies ΔE_{int}, exploring the affect of distance cutoffs (both for the n-body terms and for the enzymatic model itself), and considering up to four-body terms. These considerations are unprecedented in studies of this kind. As in previous work on enzyme thermochemistry,[25] we aim to present results that are fully converged with respect to the size of the enzyme model, while systematically testing the effects of basis set and higher n-body interactions. Our goal is to present robust protocols that can be widely deployed with relatively minor modifications, using a new software framework called FRAGME∩T.[26,30]

For non-covalent binding of small ligands with fewer than 20 atoms to T4 lysozyme proteins, we are able to achieve remarkable accuracy for ΔE_{int}, as assessed by comparison to a conventional supramolecular calculation at the same level of theory. Fragmentation errors are smaller than 0.01 kcal/mol/fragment, in converged enzyme models with 1,000+ atoms. This level of fidelity is an order-of-magnitude better than the very conservative standard of $0.1 \times (3/2)k_{\text{B}}T$ that has been suggested for fragment-based *ab initio* molecular dynamics simulations.[118] This can be achieved in a total computing time (aggregated across all processors) that is only a tiny fraction of what is required for a conventional supersystem calculation at the DFT level. The cost is also small in comparison to supramolecular calculations using semi-empirical models.

For significantly larger ligands, exemplified by the "LIDS" data set assembled for this work, we are able to obtain tightly converged results in some cases but two metalloenzymes prove to be problematic due to the presence of Zn^{2+} near the ligand binding site. In these cases, we were unable to obtain results that were converged to sub-kcal/mol accuracy at reasonable cost. However, if the intent of these calculations is to generate *ab initio* data sets for ML approaches, then there is some question as to whether sub-kcal/mol accuracy is a reasonable standard given significant disparities between ΔE_{int} (computed for a single binding pose) and $\Delta G^{\circ}_{\text{bind}}$. Although further testing is needed, the protocols developed here may already be sufficiently accurate to generate *ab initio*

training data that do not rely on experimental inhibition constants, thus can be trusted for novel ligands that do not resemble existing drugs.

Ours is the first systematically improvable fragmentation protocol to be applied to systems of this size. Our approach is robust in high-quality basis sets (up to augmented quadruple-ζ quality) and can be extended beyond the two-body level should the desired accuracy prove to be unobtainable using MBE(2). In future work, we will consider the use of counterpoise corrections that are compatible with MBE(n),[84,85] and will implement energy screening to identify the most important subset of three-body terms, as we showed that these are few in number even for the problematic Zn^{2+}-containing enzymes. Our scheme holds the potential to enable rigorous electronic structure theory calculations for large-scale computational biochemistry applications.

5 Supporting Information

Additional calculations and data (PDF)
Coordinates for the relaxed structural models (zip)

6 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. 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.[148]

References

[1] Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. *Drug Discov. Today* **2017,** *22,* 1680–1685.

[2] Lavecchia, A. Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discov. Today* **2019,** *24,* 2017–2032.

[3] Klambauer, G.; Hochreiter, S.; Rarey, M. Machine learning in drug discovery. *J. Chem. Inf. Model.* **2019,** *59,* 945–946.

[4] Patel, L.; Shukla, T.; Huang, X.; Ussery, D. W.; Wang, S. Machine learning methods in drug discovery. *Molecules* **2020,** *25,* 5277.

[5] Wei, B.; Zhang, Y.; Gong, X. DeepLPI: A novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. *Sci. Rep.* **2022,** *12,* 18200.

[6] Di Palma, F.; Abate, C.; Decherchi, S.; Cavalli, A. Ligandability and druggability assessment via machine learning. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2023,** *13,* e1676.

[7] Kim, J.; Chang, W.; Ji, H.; Joung, I. Quantum-informed molecular representation learning enhancing ADMET property prediction. *J. Chem. Inf. Model.* **2024,** *64,* 5028–5040.

[8] Nguyen, D. D.; Wei, G.-W. AGL-score: Algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *J. Chem. Inf. Model.* **2019,** *59,* 3291–3304.

[9] Yasuo, N.; Sekijima, M. Improved method of structure-based virtual screening via interaction-energy-based learning. *J. Chem. Inf. Model.* **2019,** *59,* 1050–1061.

[10] Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. Machine-learning scoring functions for structure-based drug lead optmization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020,** *10,* e1465.

[11] Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020,** *10,* 1429.

[12] Guedes, I. A.; Barreto, A. M. S.; Marinho, D.; Krempser, E.; Kuenemann, M. A.; Sperandio, O.; Dardenne, L. E.; Miteva, M. A. New machine learning and physics-based scoring functions for drug discovery. *Sci. Rep.* **2021,** *11,* 3198.

[13] Bucinsky, L.; Gall, M.; Matúška, J.; Pitoňák, M.; Štekláč, M. Advances and critical assessment of machine learning techniques for prediction of docking scores. *Int. J. Quantum Chem.* **2023,** *123,* e27110.

[14] Liu, H.; Hu, B.; Chen, P.; Wang, X.; Wang, H.; Wang, S.; Wang, J.; Lin, B.; Cheng, M. Docking score ML: Target-specific machine learning models improving docking-based virtual screening in 155 targets. *J. Chem. Inf. Model.* **2024,** *64,* 5413–5426.

[15] Harren, T.; Gutermuth, T.; Grebner, C.; Hessler, G.; Rarey, M. Modern machine-learning for binding affinity estimation of protein–ligand complexes: Progress, opportunities, and challenges. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024,** *14,* e1716.

[16] Moon, S.; Zhung, W.; Kim, W. Y. Toward generalizable structure-based deep learning models for protein–ligand interaction prediction: Challenges and strategies. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024,** *14,* e1705.

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

[18] 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.

[19] Gray, M.; Herbert, J. M. Density functional theory for van der Waals complexes: Size matters. *Annu. Rep. Comput. Chem.* **2024,** *20,* 1–61.

[20] Gavini, V. *et al.* Roadmap on electronic structure codes in the exascale era. *Model. Simul. Mater. Sci. Eng.* **2023,** *31,* 063301.

[21] Byrd, J. N.; Lotrich, V. F.; Sanders, B. A. Massively parallel computational chemistry with the super instruction architecture and ACES4. *J. Phys. Chem. A* **2024,** *128,* 7498–7509.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[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.

[31] Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012,** *137,* 064113.

[32] Richard, R. M.; Herbert, J. M. The many-body expansion with overlapping fragments: Analysis of two approaches. *J. Chem. Theory Comput.* **2013,** *9,* 1408–1416.

[33] Jacobson, L. D.; Richard, R. M.; Lao, K. U.; Herbert, J. M. Efficient monomer-based quantum chemistry methods for molecular and ionic clusters. *Annu. Rep. Comput. Chem.* **2013,** *9,* 25–58.

[34] 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).

[35] Fox, S. J.; Dziedzic, J.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Density functional theory calculations on entire proteins for free energies of binding: Application to a model polar binding site. *Proteins* **2014,** *82,* 3335–3346.

[36] Gundelach, L.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Protein-ligand free energies of binding from full-protein DFT calculations: Convergence and choice of exchange-correlation functional. *Phys. Chem. Chem. Phys.* **2021,** *23,* 9381–9393.

[37] 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.

[38] 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.

[39] Maier, S.; Thapa, B.; Erickson, J.; Raghavachari, K. Comparative assessment of QM-based and MM-based models for prediction of protein–ligand binding affinity trends. *Phys. Chem. Chem. Phys.* **2022,** *24,* 14525–14537.

[40] Fukuzawa, K.; Tanaka, S.; Yagi, Y.; Kurita, N.; Kawashita, N.; Takaba, K.; Honma, T. FMO drug design consortium. In *Recent Advances of the Fragment Molecular Orbital Method: Enhanced Performance and Applicability*; Mochizuki, Y.; Tanaka, S.; Fukuzawa, K., Eds.; Springer Nature: Singapore, 2021; pages 127–181.

[41] Ozawa, T.; Ozawa, M. Application of FMO to ligand design: SBDD, FBDD, and protein–protein interactions. In *Recent Advances of the Fragment Molecular Orbital Method: Enhanced Performance and Applicability*; Mochizuki, Y.; Tanaka, S.; Fukuzawa, K., Eds.; Springer Nature: Singapore, 2021; pages 205–252.

[42] Takimoto-Kamimura, M.; Kurita, N. Drug discovery screening by combination of x-ray crystal structure analysis and FMO calculations. In *Recent Advances of the Fragment Molecular Orbital Method: Enhanced Performance and Applicability*; Mochizuki, Y.; Tanaka, S.; Fukuzawa, K., Eds.; Springer Nature: Singapore, 2021; pages 253–266.

[43] Takaba, K. Application of FMO for protein–ligand binding affinity prediction. In *Recent Advances of the Fragment Molecular Orbital Method: Enhanced Performance and Applicability*; Mochizuki, Y.; Tanaka, S.; Fukuzawa, K., Eds.; Springer Nature: Singapore, 2021; pages 93–126.

[44] 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).

[45] Söderhjelm, P.; Aquilante, F.; Ryde, U. Calculation of protein–ligand interaction energies by a fragmentation approach combining high-level quantum chemistry with classical many-body effects. *J. Pys. Chem. B* **2009,** *113,* 11085–11094.

[46] Mazanetz, M. P.; Chudyk, E.; Fedorov, D. G.; Alexeev, Y. Applications of the fragment molecular orbital method to drug research. In *Computer-Aidd Drug Discovery*; Zhang, W., Ed.; Methods in Pharmacology and Toxicology Springer Science+Business Media: New York, 2016; pages 217–255.

[47] Fukuzawa, K.; Watanabe, C.; Okiyama, Y.; Nakano, T. How to perform FMO calculation in drug discovery. In *Recent Advances of the Fragment Molecular Orbital Method: Enhanced Performance and Applicability*; Mochizuki, Y.; Tanaka, S.; Fukuzawa, K., Eds.; Springer Nature: Singapore, 2021; pages 93–126.

[48] Gupta, A.; Maier, S.; Thapa, B.; Raghavachari, K. Towards post-Hartree–Fock accuracy for protein–ligand affinities using the molecules-in-molecules fragmentation-based method. *J. Chem. Theory Comput.* **2024,** *20,* 2774–2785.

[49] Fedorov, D. G. The fragment molecular orbital method: Theoretical development, implementation in GAMESS,

and applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2017,** *7,* e1322.

[50] Heifetz, A.; James, T.; Southey, M.; Bromidge, M. J. B. S. Guiding medicinal chemistry with fragment molecular orbital (FMO) method. In *Quantum Mechanics in Drug Discovery*, Vol. 2114; Heifetz, A., Ed.; Springer Science+Business Media: New York, 2020; Chapter 3, pages 37–48.

[51] Fukuzawa, K.; Tanaka, S. Fragment molecular orbital calculations for biomolecules. *Curr. Opin. Struc. Biol.* **2022,** *72,* 127–134.

[52] Takaya, D. Computer-aided drug design using the fragment molecular orbital method: Current status and future applications for SBDD. *Chem. Pharm. Bull.* **2024,** *72,* 781–786.

[53] Fedorov, D. Analyzing interactions with the fragment molecular orbital method. In *Quantum Mechanics in Drug Discovery*, Vol. 2114; Heifetz, A., Ed.; Springer Science+Business Media: New York, 2020; Chapter 4, pages 49–74.

[54] Kawashita, N. Interaction analysis by fragment molecular orbital method for drug discovery research. *Chem. Pharm. Bull.* **2024,** *72,* 787–793.

[55] 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.

[56] 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.

[57] 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.

[58] Kanamaru, S.; Ishiwata, Y.; Suzuki, T.; Rossman, M. G.; Arisaka, F. Control of bateriophage T4 tail lysozyme activity during the infection process. *J. Mol. Biol.* **2005,** *346,* 1013–1020.

[59] Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005,** *48,* 3714–3728.

[60] 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.

[61] Mobley, D. L.; Gilson, M. K. Predicting binding free energies: Frontiers and benchmarks. *Annu. Rev. Biophys.* **2017,** *46,* 531–558.

[62] Morton, A.; Baase, W. A.; Matthews, B. W. Energetic origins of specificity of ligand binding in an interior nonpolar cavity of T4 lysozyme. *Biochem.* **1995,** *34,* 8564–8575.

[63] Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring docking hit lists for model cavity sites: Predictions and experimental testing. *J. Mol. Biol.* **2008,** *377,* 914–934.

[64] Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.* **2009,** *394,* 747–763.

[65] Merski, M.; Fischer, M.; Balius, T. E.; Shoichet, B. K. Homologous ligands accommodated by discrete conformations of a buried cavity. *Proc. Natl. Acad. Sci. USA* **2015,** *112,* 5039–5044.

[66] Morton, A.; Matthews, B. W. Specificity of ligand binding in a buried nonpolar cavity of T4 lysozyme: Linkage of dynamics and structural plasticity. *Biochemistry* **1995,** *34,* 8576–8588.

[67] Lange, G.; Lesuisse, D.; Deprez, P.; Schoot, B.; Loenze, P.; Benard, D.; Marquette, J.; Broto, P.; Sarubbi, E.; Mandine, E. Requirements for specific binding of low affinity inhibitor fragments to the SH2 domain of ^{pp60}Src are identical to those for high affinity binding of full length inhibitors. *J. Med. Chem.* **2003,** *46,* 5184–5195.

[68] Gangjee, A.; Vidwans, A. P.; Vasudevan, A.; Queener, S. F.; Kisliuk, R. L.; Cody, V.; Li, R.; Galitsky, N.; Luft, J. R.; Pangborn, W. Structure-based design and synthesis of lipophilic 2,4-diamino-6-substituted quinazolines and their evaluation as inhibitors of dihydrofolate reductases and potential antitumor agents. *J. Med. Chem.* **1998,** *41,* 3426–3434.

[69] Campestre, C.; Agamennone, M.; Tortorella, P.; Preziuso, S.; Biasone, A.; Gavuzzo, E.; Pochetti, G.; Mazza, F.; Hiller, O.; Tschesche, H.; Consalvi, V.; Gallina, G. *N*-hydroxyurea as zinc binding group in matrix metalloproteinase inhibition: Mode of binding in a complex with MMP-8. *Bioorg. Med. Chem. Lett.* **2006,** *16,* 20–24.

[70] Browner, M. F.; Smith, W. W.; Castelhano, A. L. Matrilysin-inhibitor complexes: Common themes among metalloproteases. *Biochemistry* **1995,** *34,* 6602–6610.

[71] Bjorge, J.; Jakymiw, A.; Fujita, D. J. Selected glimpses into the activation and function of Src kinase. *Oncogene* **2000,** *19,* 5620–5635.

[72] Soriano, P.; Montgomery, C.; Geske, R.; Bradley, A. Targeted disruption of the c-*src* proto-oncogene leads to osteopetrosis in mice. *Cell* **1991,** *64,* 693–702.

[73] 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.

[74] Schrödinger, L. "The PyMOL molecular graphics system, v. 2.1", .

[75] 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.

[76] 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.

[77] Řezáč, J.; Stewart, J. J. P. How well do semiempirical QM methods describe the structure of proteins? *J. Chem. Phys.* **2023,** *158,* 044118.

[78] Boz, E.; Stein, M. Accurate receptor-ligand binding free energies from fast QM conformational chemical space sampling. *Int. J. Mol. Sci.* **2021,** *22,* 3078.

[79] Chen, Y.; Sheng, Y.; Ma, Y.; Ding, H. Efficient calculation of protein–ligand binding free energy using GFN methods: The power of the cluster model. *Phys. Chem. Chem. Phys.* **2022,** *24,* 14339–14347.

[80] Chan, B.; Dawson, W.; Nakajima, T. Sorting drug conformers in enzyme active sites: The XTB way. *Phys. Chem. Chem. Phys.* **2024,** *26,* 12610–12618.

[81] Spicher, S.; Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem. Int. Ed. Engl.* **2020,** *59,* 15665–15673.

[82] Wolter, M.; von Looz, M.; Meyerhenke, H.; Jacob, C. R.

Systematic partitioning of proteins for quantum-chemical fragmentation methods using graph algorithms. *J. Chem. Theory Comput.* **2021,** *17,* 1355–1367.
[83] 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.
[84] 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.
[85] 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.
[86] Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M.; Petrova, G. P.; Harris, T. V.; Li, X.; Ke, Z.; Liu, F.; Li, H.-B.; Ding, L.; Morokuma, K. The ONIOM method and its applications. *Chem. Rev.* **2015,** *115,* 5678–5796.
[87] Mayhall, N. J.; Raghavachari, K. Molecules-in-molecules: An extrapolated fragment-based approach for accurate calculations on large molecules and materials. *J. Chem. Theory Comput.* **2011,** *7,* 1336–1343.
[88] Tschumper, G. S. Multicentered integrated QM:QM methods for weakly bound clusters: An efficient and accurate 2-body:many-body treatment of hydrogen bonding and van der Waals interactions. *Chem. Phys. Lett.* **2006,** *427,* 185–191.
[89] Wen, S.; Nanda, K.; Huang, Y.; Beran, G. J. O. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. *Phys. Chem. Chem. Phys.* **2012,** *14,* 7578–7590.
[90] Sahu, N.; Gadre, S. R. Molecular tailoring approach: A route for *ab initio* treatment of large clusters. *Acc. Chem. Res.* **2014,** *47,* 2739–2747.
[91] 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.
[92] 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.
[93] 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.
[94] 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.
[95] 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.
[96] 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.
[97] Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **2010,** *133,* 134105.
[98] Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013,** *34,* 1672–1685.
[99] Gilson, M. K.; Honig, B. H. The dielectric constant of a folded protein. *Biopolymers* **1986,** *25,* 2097–2119.
[100] 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.
[101] Rodgers, K. K.; Silgar, S. G. Surface electrostatics, reduction potentials, and internal dielectric constant of proteins. *J. Am. Chem. Soc.* **1991,** *113,* 9419–9421.
[102] Nakamura, H. Roles of electrostatic interaction in proteins. *Q. Rev. Biophys.* **1996,** *29,* 1–90.
[103] 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.
[104] 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.
[105] 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.
[106] Herbert, J. M. Dielectric continuum methods for quantum chemistry. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021,** *11,* e1519.
[107] 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.
[108] 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.
[109] 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.
[110] 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.
[111] 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.
[112] 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.
[113] 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.
[114] 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.
[115] 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.

[116] 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.

[117] Heifetz, A., Ed.; *Quantum Mechanics in Drug Discovery;* volume 2114 of *Methods in Molecular Biology* Springer Science+Business Media: New York, 2020.

[118] Ouyang, J. F.; Bettens, R. P. A. Many-body basis set superposition effect. *J. Chem. Theory Comput.* **2015,** *11,* 5132–5143.

[119] Dziedzic, J.; Helal, H. H.; Skylaris, C.-K.; Mostofi, A. A.; Payne, M. C. Minimal parameter implicit solvent model for ab initio electronic-structure calculations. *Europhys. Lett.* **2011,** *95,* 43001.

[120] Grimme, S. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chem. Eur. J.* **2012,** *18,* 9955–9964.

[121] Pracht, P.; Grimme, S. Calculation of absolute molecular entropies and heat capacities made simple. *Chem. Sci.* **2021,** *12,* 6551–6568.

[122] Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **1994,** *98,* 1978–1988.

[123] Tan, C.; Tan, Y.-H.; Luo, R. Implicit nonpolar solvent models. *J. Phys. Chem. B* **2007,** *111,* 12263–12274.

[124] Glick, Z. L.; Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials. *J. Chem. Phys.* **2020,** *153,* 044112.

[125] Glick, Z.; Metcalf, D.; Sargent, C.; Spronk, S.; Koutsoukas, A.; Cheney, D.; Sherrill, C. D. A physics-aware neural network for protein-ligand interactions with quantum chemical accuracy. *Chem. Sci.* **2024,** *15,* 13313–13324.

[126] Prentice, J. C. A. *et al.* The ONETEP linear-scaling density functional theory program. *J. Chem. Phys.* **2020,** *152,* 174111.

[127] 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.

[128] Grassano, J. S.; Pickering, I.; Roitberg, A. E.; Lebrero, M. C. G.; Estrin, D. A.; Semelak, J. A. Assessment of embedding schemes in a hybrid machine learning/classical potentials (ML/MM) approach. *J. Chem. Inf. Model.* **2024,** *64,* 4047–4058.

[129] Valdés, H.; Klusák, V.; Pitoňák, M.; Exner, O.; Starý, I.; Hobza, P.; Rulíšek, L. Evaluation of the intramolecular basis set superposition error in the calculations of larger molecules: [*n*]helicenes and Phe-Gly-Phe tripeptide. *J. Comput. Chem.* **2007,** *29,* 861–870.

[130] Shields, A. E.; van Mourik, T. Comparison of ab initio and DFT electronic structure methods for peptides containing an aromatic ring: Effect of dispersion and BSSE. *J. Phys. Chem. A* **2007,** *111,* 13272–13277.

[131] van Mourik, T. Determining potential energy surfaces for flexible peptides. Problems caused by intramolecular BSSE and dispersion. In *Molecular Potential Energy Surfaces in Many Dimensions*; Law, M. M.; Ernesti, A., Eds.; Collaborative Computational Project on Molecular Quantum Dynamics (CCP6): Daresbury Laboratory, Daresbury, Warrington, UK, 2009.

[132] Hameed, R.; Khan, A.; van Mourik, T. Intramolecular BSSE and dispersion affect the structure of a dipeptide conformer. *Mol. Phys.* **2017,** *116,* 1236–1244.

[133] Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008,** *51,* 3661–3680.

[134] Murray, C. W.; Rees, D. C. The rise of fragment-based drug discovery. *Nat. Chem.* **2009,** *1,* 187–192.

[135] Kumar, A.; Voet, A.; Zhang, K. Y. J. Fragment based drug design: From experimental to computational approaches. *Curr. Med. Chem.* **2012,** *19,* 5128–5147.

[136] Efremov, I. V.; Erlanson, D. A. Fragment-based lead generation. In *Lead Generation: Methods, Strategies, and Case Studies*, First ed.; Holenz, J., Ed.; Wiley-VCH: Weinheim, 2016; Chapter 6, pages 133–157.

[137] Doak, B. C.; Norton, R. S.; Scanlon, M. J. The ways and means of fragment-based drug design. *Pharmacol. Therapeut.* **2016,** *167,* 28–37.

[138] Kirsch, P.; Hartman, A. M.; Hirsch, A. K. H.; Empting, M. Concepts and core principles of fragment-based drug design. *Molecules* **2019,** *24,* 4309.

[139] 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 Discov.* **2014,** *13,* 105–121.

[140] Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **2004,** *3,* 711–715.

[141] Bunnage, M. E. Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* **2011,** *7,* 335–339.

[142] Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **2014,** *32,* 40–51.

[143] Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* **2015,** *14,* 475–486.

[144] Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007,** *13,* 1173–1213.

[145] 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.

[146] Xu, M.; He, X.; Zhu, T.; Zhang, J. Z. H. A fragment quantum mechanical method for metalloproteins. *J. Chem. Theory Comput.* **2019,** *15,* 1430–1439.

[147] Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015,** *143,* 054107.

[148] Ohio Supercomputer Center, http://osc.edu/ark:/19495/f5s1ph73 (accessed 2024-10-23).

TOC Graphic