JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

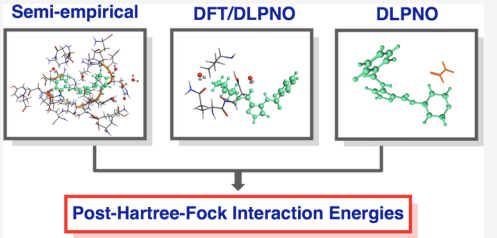
Toward Post-Hartree−Fock Accuracy for Protein−Ligand Affinities Using the Molecules-in-Molecules Fragmentation-Based Method

Ankur K. Gupta, Sarah Maier, Bishnu Thapa, and Krishnan Raghavachari*

Cite This: J. Chem. Theory Comput. 2024, 20, 2774−2785 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The complexity and size of large molecular systems, such as protein−ligand complexes, pose computational challenges for accurate post-Hartree−Fock calculations. This study delivers a thorough benchmarking of the Molecules-in-Molecules (MIM) method, presenting a clear and accessible strategy for layer/theory selections in post-Hartree−Fock computations on substantial molecular systems, notably protein−ligand complexes. An approach is articulated, enabling augmented computational efficiency by strategically canceling out common subsystem energy terms between complexes and proteins within the supermolecular equation. Employing DLPNO-based post-Hartree−Fock methods in conjunction with the three-layer MIM method (MIM3), this study demonstrates the achievement of protein−ligand binding energies with remarkable accuracy (errors <1 kcal mol^{-1}), while significantly reducing computational costs. Furthermore, noteworthy correlations between theoretically computed interaction energies and their experimental equivalents were observed, with R^2 values of approximately 0.90 and 0.78 for CDK2 and BZT-ITK sets, respectively, thus validating the efficacy of the MIM method in calculating binding energies. By highlighting the crucial role of diffuse or small Pople-style basis sets in the middle layer for reducing energy errors, this work provides valuable insights and practical methodologies for interaction energy computations in large molecular complexes and opens avenues for their application across a diverse range of molecular systems.



1. INTRODUCTION

Fragmentation methods in computational chemistry have emerged as a promising approach for achieving ab initio-level accuracy for large systems.[1−7] Nevertheless, to affirm the effectiveness of a fragmentation method in replicating post-Hartree−Fock (HF) level energies and properties, systematic performance analysis is required before its deployment for routine usage. A major challenge lies in the substantial computational demands of correlated methods in their canonical form, making even benchmarking tasks daunting with present-day hardware. While several studies have sought to reproduce post-HF level energies through fragmentation schemes, most have resorted to using small basis sets (primarily double-ζ Pople basis).[8−15] Correlated methods, unlike density functional theory (DFT) and HF methods, converge at a slower pace with respect to basis set size, making them notably susceptible to basis set incompleteness error (BSIE) and basis set superposition error (BSSE). This issue is exacerbated when using small basis sets. Consequently, larger basis sets are often favored to obtain reliable energies when using post-Hartree−Fock methods.

Recent advancements in *tensor decomposition methods*, including pair natural orbital (PNO),[16−22] orbital specific virtual (OSV),[23,24] and optimized virtual orbital space (OVOS) methods,[25−29] offer a paradigm shift in computational efficiency. The foundational concept underlying these methods is the strategic decomposition of high-dimensional tensors, e.g., two-electron integrals, into linear combinations of computationally amenable lower-dimensional tensors, facilitating large basis set correlated computations on extensive molecular systems within computationally acceptable timeframes. A case in point is the domain-based local pair natural orbital (DLPNO) theory, which has been widely adopted due to its seamless integration into established quantum chemistry software packages, namely, Orca.[30,31] In the DLPNO method, the core principle involves the use of projected atomic orbitals (PAOs) to span the virtual space while remaining localized, thus enabling the definition of specific correlation domains for each electron pair. The construction of natural orbitals (NOs) for every electron pair, called pair natural orbitals (PNOs), is achieved by expanding each PNO in a set of PAOs that are associated with a given electron pair's unique domain. DLPNO achieves a highly compact representation of the wavefunction by selectively including PNOs with an occupation number exceeding a predefined threshold (TCutPNO) in the correlation space. Consequently, this approach manages to reduce computational requirements significantly while minimizing the impact on accuracy. In fact, DLPNO variants,[32−37]

Received: November 24, 2023
Revised: February 29, 2024
Accepted: March 5, 2024
Published: March 26, 2024

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

including DLPNO-MP2 and DLPNO−CCSD(T) have demonstrated their capability to recover over 99.9% of the canonical correlation energy in test cases. This translates to minuscule errors—less than 0.1 kcal mol^{-1}—in DLPNO-derived relative energies compared to their canonical counterparts. However, it is worth noting that despite the time and computational efficiency of DLPNO methods, they come with hardware constraints, especially significant local disk space requirements. These demands can limit their application to even advanced computing systems. Therefore, we propose that integrating molecular fragmentation techniques can enhance the speed of DLPNO methods, making them viable across a broader spectrum of hardware setups.

This study seeks to devise a robust, and efficient computational protocol, hypothesizing that the incorporation of the Molecules-in-Molecules (MIM) fragmentation-based method[38,39] will facilitate the calculation for large molecules, which are often of practical interest, at post-Hartree−Fock precision, without a pronounced computational penalty. In particular, we demonstrate our approach by employing the MIM method to compute protein−ligand binding (or interaction) energies, utilizing the DLPNO level of theory as a surrogate for canonical correlated methods.[39] In a closely related effort, Ni et al.[40] recently performed benchmark studies on systems of up to approximately 1000 atoms, employing DLPNO methods at the complete basis set (CBS) limit through an extrapolation scheme and successfully achieved high accuracy in reproducing the energies using the Cluster-in-Molecule (CIM) approach.

Protein−ligand modeling is a widely researched area due to its key role in computer-aided drug discovery and is foundational for many computational chemistry challenges. Additionally, understanding interactions in biologically important systems is essential for developing treatments against prevailing and emerging life-threatening diseases. In this context, the MIM method has proven valuable for exploring various elements of protein−ligand chemistry, such as computing interaction energies,[41] decomposing total interaction energies into discrete contributions from individual protein moieties,[42] and understanding activity cliffs[43] at the DFT level of theory. This work extends the applicability of MIM to compute protein−ligand binding energies with wave function-based electron correlation theory and presents a method that is generally applicable to molecular systems with several hundred atoms.

2. METHODS

2.1. MIM Method. The MIM approach uses a fragmentation-based, multilayered composite energy methodology, wherein each layer, differing by subsystem size, captures various interatomic interactions through well-structured algorithms. The fragmentation process in MIM consists of sequential steps: initially, the molecule is split into nonoverlapping 'monomers' by cleaving nonpolar single bonds (e.g., C−C bonds), which could then be combined to form primary subsystems. The creation of a primary subsystem in MIM typically adheres to two primary criteria: number-based or distance-based formation. The former involves a predetermined number of neighboring monomers to create a subsystem, while the latter includes all monomers within a specified radial distance from a parent monomer fragment to build a primary subsystem. It is important to note that these primary subsystems inherently have overlapping areas; hence, derivative subsystems, shaped by the overlapping regions of primary subsystems, are generated to avoid atom and interaction overcounting. The energy (and other physical properties) of the full system, relevant to the employed fragmentation scheme, are subsequently derived using the inclusion-exclusion principle.

$$E_r = \sum_{i_1} E_{i_1} - \sum_{i_1<i_2} E_{i_1\cap i_2} + \ldots + (-1)^{n-1} \sum_{i_1<\ldots<i_n} E_{i_1\cap i_2\ldots\cap i_n} \quad (1)$$

Here, r acts as a flexible fragmentation parameter. The term E_{i_1} denotes the energy of the i_1th primary subsystem, while $E_{i_1\cap i_2}$ represents the energy of the derivative subsystem, derived from the intersection of two primary subsystems. Similarly, $E_{i_1\cap i_2\ldots\cap i_n}$ indicates the energy of the derivative subsystem originating from an n-body (or subsystem) overlap. Equation 1 facilitates the calculation of the energy of a large molecule at a high level of theory (E_r^{high}). It is crucial to acknowledge that, due to the molecule being partitioned into smaller subsystems, a minor fraction of the long-range interactions is neglected. Consequently, the error associated with the fragmentation scheme is expressed as,

$$\text{Error} = E_r^{\text{high}} - E_\infty^{\text{high}} \quad (2)$$

In this case, E_∞^{high} denotes the energy of the entire molecule computed at the designated high level of theory. Typically, executing high-level calculations for large systems is not practically feasible; therefore, this error is approximated by utilizing a more accessible, low level of theory. Thus, following the ONIOM formalism, the energy of the full molecule at a high level of theory (E_∞^{high}) can be computed as follows:

$$E_\infty^{\text{high}} = E_r^{\text{high}} - (E_r^{\text{high}} - E_\infty^{\text{high}}) \approx E_r^{\text{high}} - (E_r^{\text{low}} - E_\infty^{\text{low}}) \quad (3)$$

Through the application of eq 3, a structured hierarchy of schemes can be devised, aiming to diminish the Error (defined in eq 2) and enhance computational accuracy. As a practical example, the most fundamental scheme, labeled MIM1, does not incorporate any error correction and encompasses a singular layer. Therefore, the energy of the molecule is derived using eq 1, written as follows:

$$E_{\text{MIM1}} = E_r^{\text{high}} \quad (4)$$

In the subsequent model within the MIM hierarchy, MIM2, the Error (as detailed in eq 2) is approximated by utilizing a second layer and a full molecule calculation at a low level of theory, thereby accounting for all long-range interactions within the molecule. The corresponding energy equation is depicted as follows:

$$E_{\text{MIM2}} = E_r^{\text{high}} - (E_r^{\text{low}} - E_\infty^{\text{low}}) \quad (5)$$

Similarly, the formula for an n-layer fragmentation scheme may be expressed as,

$$E_{\text{MIM}n} = E_{r_1}^{\text{high}} - \sum_{i=2}^{n} (E_{r_{i-1}}^{t_i} - E_{r_i}^{t_i}) \quad (6)$$

where t_i and r_i denote the level of theory and fragmentation scheme, respectively. The level of theory implemented in each layer is judiciously selected in accordance with the fragment

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

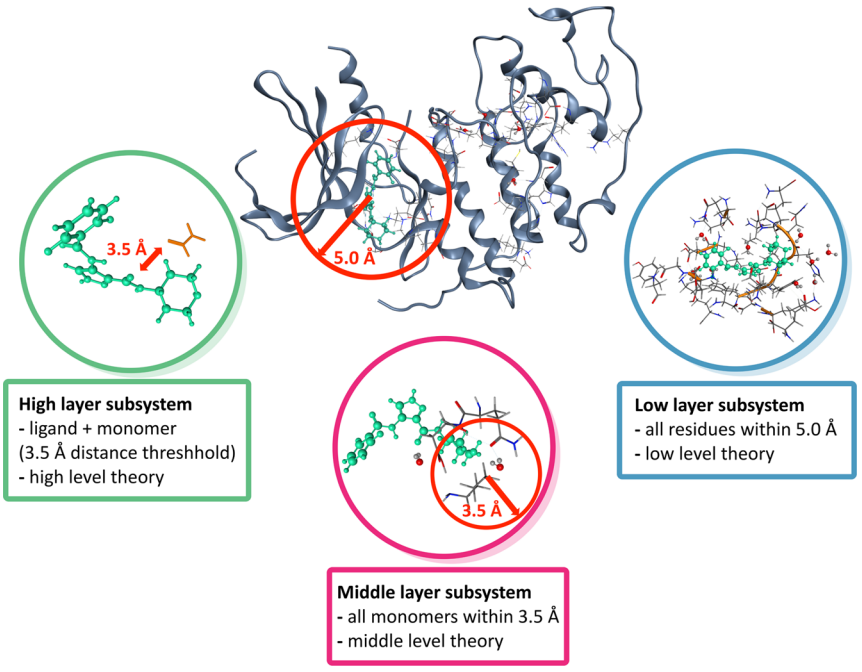


Figure 1. Schematic diagram of the Molecules-in-Molecules three-layer fragmentation scheme (MIM3). In the "high layer", we implement a number-based scheme enhanced with distance-based criteria. For the computation of binding energies, this approach simplifies to focus on interactions between the ligand and specific protein fragments within a 3.5 Å radius due to the cancellation of subsystems. The "middle layer" consists of subsystems formed around each fragment center, extending to a radius of 3.5 Å, based on distance criteria. The "low layer" encompasses full molecule calculations, including all residues within a 5.0 Å radius, utilizing computationally efficient, lower-level theories such as semiempirical methods.

size in that particular layer, with higher-accuracy methods applied to smaller fragments. In the *n*-layer model, various forms of interfragment and long-range interactions are segregated and addressed within distinct layers.

The MIM method extends its applicability beyond energy calculations, enabling the computation of energy derivatives and spectroscopic properties. Analogous to the ONIOM technique, all truncated bonds within primary and derivative subsystems are saturated with hydrogen atoms. This procedural adaptation necessitates slight adjustments to the strategies employed for computing energy derivatives within the MIM framework. Due to the exclusive presence of hydrogen link atoms within subsystems—absent in the full molecule—the derivatives related to these link atoms are projected onto the corresponding supporting and host atoms, a process previously elaborated upon in the existing literature.[44−46]

2.2. MIM for Protein−Ligand Binding Energies. In the primary topic of this investigation, the binding energy between a protein and ligand is estimated by employing the single snapshot structure approximation. Consequently, the interaction energy corresponding to a specific protein−ligand pair is calculated by using the supermolecular approach, expressed as follows:

$$E^{\text{interaction}} = E^{\text{complex}} - E^{\text{protein}} - E^{\text{ligand}} \tag{7}$$

Here, E^{complex}, E^{protein}, and E^{ligand} denote the energies of the interaction complex, protein, and ligand, respectively. Given the substantial size of protein molecules—typically ranging from 300 to 800 atoms when implementing a cutoff radius of 5 Å from the ligand—executing their energy computations at appreciably high levels of theory using traditional methods proves to be computationally inhibitive if not prohibitive. As such, fragmentation methods have evolved into invaluable assets for quantum mechanically deciphering protein−ligand interactions. Notably, a three-layer MIM model (MIM3) has demonstrated considerable success in calculating protein−ligand binding energies with DFT accuracy in an economically efficient manner. The energy of the molecule in the MIM3 model, derived from eq 6, is computed as follows:

$$E_{\text{MIM3}} = E_{r_1}^{\text{high}} - (E_{r_1}^{\text{medium}} - E_{r_2}^{\text{medium}}) - (E_{r_2}^{\text{low}} - E_{\infty}^{\text{low}}) \tag{8}$$

By substituting eq 8 into eq 7, a mathematical expression for calculating interaction energy using the MIM3 model is obtained:

$$E_{\text{MIM3}}^{\text{interaction}} = E_{\text{MIM3}}^{\text{complex}} - E_{\text{MIM3}}^{\text{protein}} - E^{\text{ligand(high)}} \tag{9}$$

For simplicity, we assume in eq 9 that the ligand itself does not have to be fragmented, although generalizations including ligand fragmentation are easily derived.

$$\begin{aligned} E_{\text{MIM3}}^{\text{interaction}} = {} & E_{r_1}^{\text{complex(high)}} - (E_{r_1}^{\text{complex(medium)}} \\ & - E_{r_2}^{\text{complex(medium)}}) - (E_{r_2}^{\text{complex(low)}} \\ & - E_{\infty}^{\text{complex(low)}}) - E_{r_1}^{\text{protein(high)}} \\ & + (E_{r_1}^{\text{protein(medium)}} - E_{r_2}^{\text{protein(medium)}}) \\ & + (E_{r_2}^{\text{protein(low)}} - E_{\infty}^{\text{protein(low)}}) \\ & - (E^{\text{ligand(high)}} - E^{\text{ligand(medium)}} + E^{\text{ligand(medium)}} \\ & - E^{\text{ligand(low)}} + E^{\text{ligand(low)}}) \end{aligned} \tag{10}$$

Rearranging eq 10, we obtain the following simplified expression,

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

$$E_{\mathrm{MIM3}}^{\mathrm{interaction}} = E_{r_1}^{\mathrm{interaction(high)}} - (E_{r_1}^{\mathrm{interaction(medium)}} - E_{r_2}^{\mathrm{interaction(medium)}}) - (E_{r_2}^{\mathrm{interaction(low)}} - E_{\infty}^{\mathrm{interaction(low)}}) \quad (11)$$

Consequently, as demonstrated in eq 11, interaction energies can be calculated independently for each MIM layer, which, as described below, provides additional speed-up in intermolecular energy computations.

According to methodologies outlined in our previous articles, we typically utilize a number-based scheme, augmented with distance-based nonbonded dimers (i.e., N4D, N5D models in ref 36. for the high layer, and a distance-based scheme (employing a cutoff radius of 3.5 Å) for the middle layer (Figure 1). Specifically, in the high layer, we employed the N4D scheme (denoted as the r_1 scheme in eq 11), which combines physically bonded monomeric fragments into tetramers and, for effective capture of interfragment noncovalent interactions, forms dimeric subsystems between monomers within a 3.5 Å cutoff radius. In the middle layer, we applied a distance-based fragmentation scheme (denoted as the r_2 scheme in eq 11), with a cutoff radius of 3.5 Å. This method involves incorporating all monomers within this specified distance to form subsystems, thereby enabling a more comprehensive calculation of interaction energies within the protein−ligand complexes, including interactions beyond the immediate bonding environments. While MIM already demonstrates computational efficiency in obtaining absolute energies of large protein molecules, the calculation of binding energies offers an opportunity to further accelerate computations by omitting all subsystem calculations common to both complex and protein molecules. This feature directly stems from the use of the same protein structure snapshot for both complex and protein energy computations. Elaborating further, when calculating interaction energy using MIM3 (eq 11), a MIM layer might include subsystems originating from the complex that are the same as those derived from the protein, thereby canceling their energies in the expanded MIM energy expression. Specifically, the number-based subsystems from the complex cancel those originating from the protein in the high layer. Extending the same concept, numerous distance-based dimers can also be removed from the high layer (r_1) interaction energy expressions ($E_{r_1}^{\mathrm{interaction(high)}}$ and $E_{r_1}^{\mathrm{interaction(medium)}}$); consequently, the high layer consists solely of distance-based dimers incorporating the ligand and a protein fragment. Such subsystem cancellations also occur in the middle layer, especially among subsystems that do not include any ligand moiety.

This work aims to utilize post-Hartree−Fock methods within the MIM framework to compute protein−ligand binding energies. The high level of theory in the MIM model, or the target level, is selected to be a DLPNO method (i.e., DLPNO-MP2, DLPNO−CCSD(T)) paired with a large polarized triple-ζ basis set (i.e., def2-TZVP, cc-pVTZ). Given that the high layer subsystems are relatively small, including roughly 20−50 atoms, they can be easily computed with such high-level theories. On the other extreme, the MIM low layer, requiring full molecule computations, limits the suitable level of theory options due to computational cost, generally confining them to semiempirical theories (i.e., PM6-D3, PM7, xTB, etc.) or the 3c methods[47] (viz., B97−3c, HF-3c, etc.). However, the middle layer subsystem sizes offer some flexibility in method selection. As such, we have tested DLPNO and DFT (B97-D3) methods with a range of small- and medium-sized basis sets to determine the dependency of the error in MIM interaction energies on the middle layer theory. Therefore, we strive to fine-tune the middle layer theory while keeping the high and low levels of theory fixed. Finally, we reiterate that as each subsystem calculation is independent, MIM energy computations can be efficiently parallelized across multiple cores or compute nodes.

3. COMPUTATIONAL DETAILS AND STRUCTURE PREPARATION

All the DLPNO and B97−3c[48] computations were carried out using Orca 4.2.1.[30,31] RIJCOSX[49,50] approximation was employed in every DLPNO computation to speed up the self-consistent field (SCF) convergence. It is important to note that while RIJCOSX accelerates the SCF process by approximating the calculation of Coulomb and exchange energies, it introduces a level of approximation. Our preliminary analysis for the cyclin-dependent kinase 2 (CDK2) data set revealed that the mean absolute error (MAE) in binding energies, when comparing calculations with and without the RIJCOSX option for the DLPNO-MP2 def2-TZVP method, is approximately 1.3 kcal mol^{-1}. This error magnitude is relatively small in the context of the absolute binding energies of the complexes in this data set, indicating that the RIJCOSX approach, while introducing a computational approximation, maintains a reasonable balance between efficiency and accuracy in our interaction energy calculations. All the DFT (B97-D3[51−53]) and PM6-D3[54] calculations were done using the Gaussian 16 software,[55] except those involving ma-def2-SVP and ma-def2-SV(P) basis sets, which were done using Orca 4.2.1.

Preparation of complexed protein−ligand structures followed a protocol detailed in one of our previous works.[41] The cyclin-dependent kinase 2 (CDK2) data set is comprised of 13 complexes, featuring a series of 13 unique ligand molecules, each bound to a common kinase receptor.[56,57] Reported cocrystal PDB structures exist for all 13 complexes (PDB IDs: 2VTA, 2VTH, 2VTI, 2VTJ, 2VTM, 2VTN, 2VTO, 2VTQ, 2VTR, 2VU3, 2VTT, 2VTS) and were used as the starting structures for all subsequent calculations. Cocrystal structures are not reported for all ligands in the benzothiazole (BZT)-based data set of interleukin-2-inducible T-cell kinase inhibitors, and so an altered approach was adopted.[58,59] In this case, a congeneric set of BZT ligands was generated from a single published cocrystal structure (PDB ID: 4MF0). Appropriate modifications were made to the published ligand crystal structure; subsequently, the modified ligand was placed into the binding pocket of the published receptor structure. During ligand modification, the flexible alignment module in MOE was employed to preserve key binding features observed in the crystal structure of 4MF0.[60] Here we assume that similar ligands bind similarly to a common receptor.

Missing hydrogen atoms were added to the crystal structure using the Protonate 3D tool as implemented in the Molecular Operating Environment (MOE) program (version 2019.01).[61] Each complex was then minimized using MOE with the AMBER10:EHT force field, employing a generalized Born/volume integral implicit solvation model with an internal dielectric of 2 and an external dielectric of 80.[62,63] During the minimization, restraints were used, whereby a harmonic potential was placed on each atom with a force constant

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

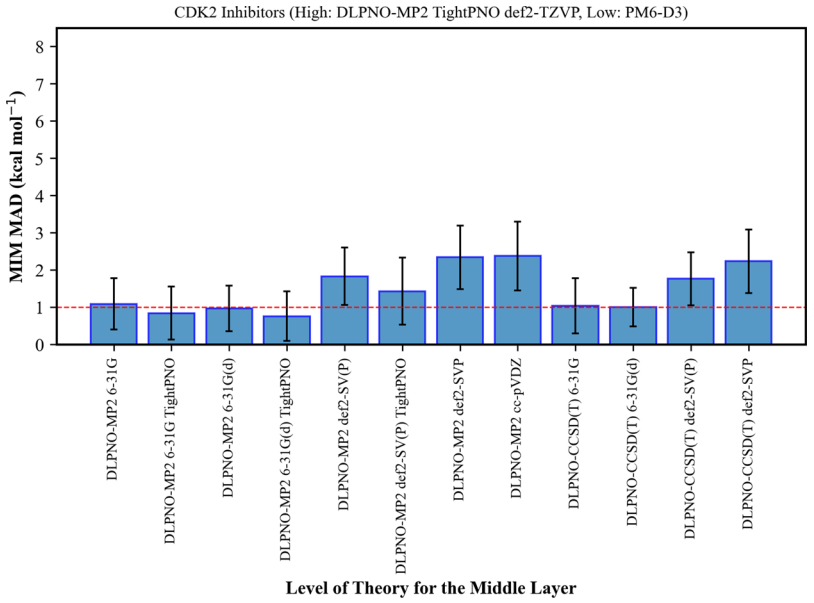


Figure 2. Bar plot showing mean absolute deviation (MAD) in MIM3 interaction energies compared to unfragmented full molecule calculations for the CDK2 inhibitors data set. The plot illustrates the variation in MAD as a function of the middle layer theory, employing only post-Hartree−Fock methods, with the high and low layer theories held constant at DLPNO-MP2 TightPNO def2-TZVP and PM6-D3, respectively.

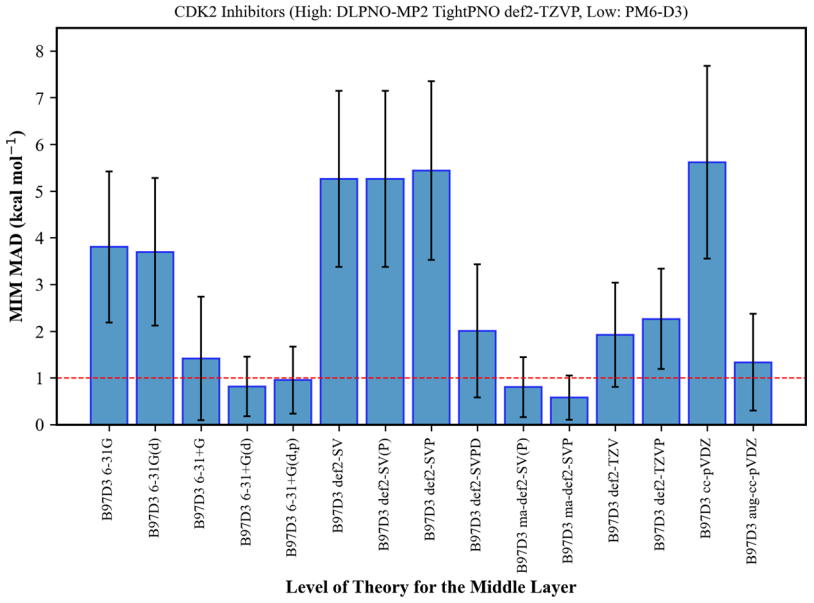


Figure 3. Bar plot showing mean absolute deviation (MAD) in MIM3 interaction energies compared to unfragmented full molecule calculations for the CDK2 inhibitors data set. The plot illustrates the variation in MAD as a function of the middle layer theory, employing only DFT (B97D3) methods, with the high and low layer theories held constant at DLPNO-MP2 TightPNO def2-TZVP and PM6-D3, respectively.

defined by the standard deviation (0.5 Å) from the starting coordinates at 300 K.

4. RESULTS AND DISCUSSION

To evaluate the proficiency of the MIM method in accurately replicating protein−ligand binding energies using high-fidelity correlated methods, our primary focus was on the CDK2 series, which encompasses 13 experimentally documented complex crystal structures. In an effort to further substantiate our findings, we also computed MIM3 interaction energies for the BZT-ITK set, contrasting them with those derived from full molecule computations. For quantum mechanical (QM) calculations, structures were prepared by incorporating residues and water molecules within a 5.0 Å radius of the ligand. In addition, all dangling bonds are capped with hydrogen atoms to maintain the integrity of the atomic valence. This approach resulted in systems sized between 250−500 and 350−450 atoms for the CDK2 and BZT-ITK sets, respectively.

To evaluate the efficacy of a specific theoretical combination within the MIM3 model, we calculated the difference between interaction energies derived from MIM3 and those computed using the full molecular complex, providing a quantifiable metric for the MIM error. This error quantification is reminiscent of the *S*-Value test, ubiquitously used for validation and theory selection in ONIOM calculations.[64,65]

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

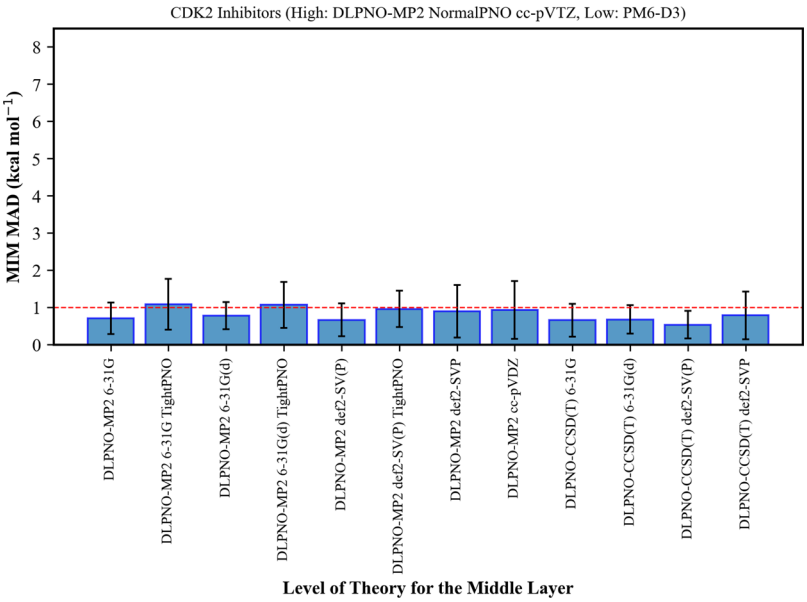


Figure 4. Bar plot showing mean absolute deviation (MAD) in MIM3 interaction energies compared to unfragmented full molecule calculations for the CDK2 inhibitors data set. The plot illustrates the variation in MAD as a function of the middle layer theory, employing only post-Hartree−Fock methods, with the high and low layer theories held constant at DLPNO-MP2 NormalPNO cc-pVTZ and PM6-D3, respectively.

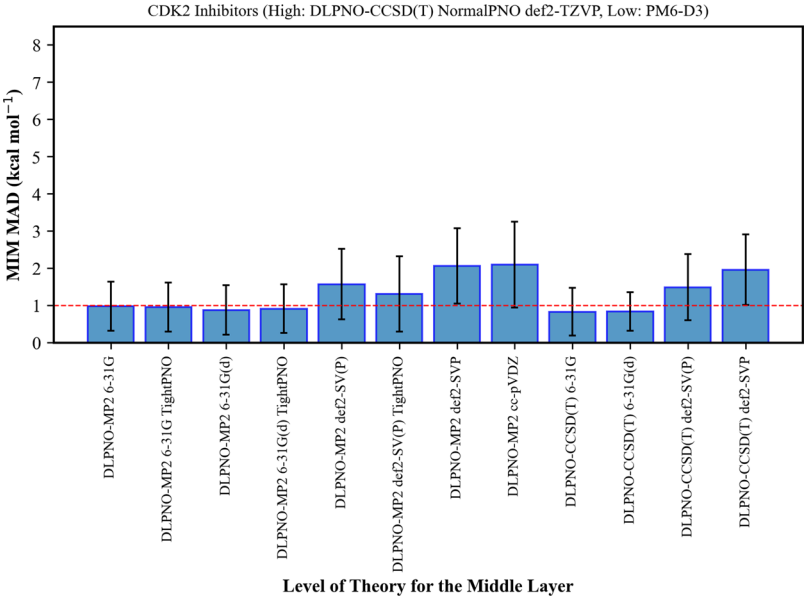


Figure 5. Bar plot showing mean absolute deviation (MAD) in MIM3 interaction energies compared to unfragmented full molecule calculations for the CDK2 inhibitors data set. The plot illustrates the variation in MAD as a function of the middle layer theory, employing only post-Hartree−Fock methods, with the high and low layer theories held constant at DLPNO−CCSD(T) NormalPNO def2-TZVP and PM6-D3, respectively.

Given that a protein−ligand set (e.g., CDK2, BZT-ITK) is composed of multiple protein−ligand complexes, we computed the mean of the MIM errors (derived from each sample within the set) to act as a performance indicator of the MIM3 model for a given combination of theories. Specifically, we employed the mean absolute deviation (MAD)—a metric attained by computing the mean of the absolute values of the deviations, which are, in turn, obtained by subtracting the systematic error (mean signed error) from the signed error of each data point. Consequently, MAD, being inherently free from systematic errors, serves as an optimal metric for this benchmark study.

Initially, we tested the DLPNO-MP2/def2-TZVP theory under TightPNO settings as the target (or high) level of theory to calculate binding energies for the CDK2 set. Figure 2 demonstrates the testing of DLPNO-MP2 and DLPNO−CCSD(T) theories, paired with various small- and medium-sized basis sets for the middle layer, yielding MADs of approximately 2 kcal mol^{-1} across all theories—a figure relatively modest in contrast to the computed binding energies, which range between 30 and 100 kcal mol^{-1}. Interestingly, small Pople-style basis sets (6-31G and 6-31G(d)) in the middle layer offered the most efficient performance with deviations less than 1 kcal mol^{-1}. Furthermore, our experimentation with DFT (B97-D3) in the middle layer

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

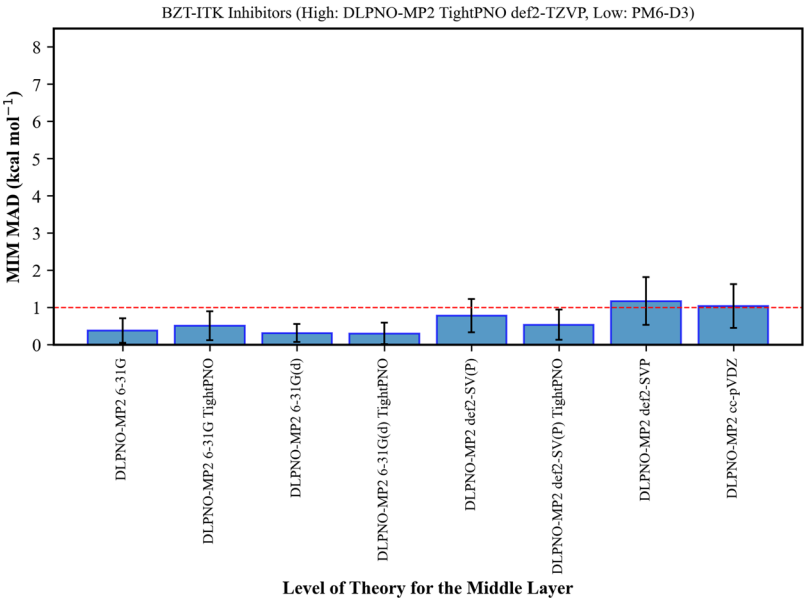


Figure 6. Bar plot showing mean absolute deviation (MAD) in MIM3 interaction energies compared to unfragmented full molecule calculations for the BZT-ITK inhibitors data set. The plot illustrates the variation in MAD as a function of the middle layer theory, employing only post-Hartree−Fock methods, with the high and low layer theories held constant at DLPNO-MP2 TightPNO def2-TZVP and PM6-D3, respectively.

(Figure 3) indicated a trend in MADs, suggesting the crucial role of diffuse basis functions for achieving reduced MIM errors when DFT is employed in the middle layer. In contrast, the use of nondiffuse basis sets in the middle layer was sufficient for achieving excellent results with the correlated (or DLPNO) methods. This observation may be ascribed to disparity in the theoretical methods applied across different layers in multilayer fragmentation methods.

In our multilayered computational approach, the high layer employs a correlated ab initio method, inherently incorporating a degree of diffuseness through the inclusion of virtual orbitals in the wave function. This plays a key role in the description of long-range dispersion interactions. Consequently, to ensure methodological consistency across MIM layers, it becomes crucial for the middle layer to incorporate similar physical effects. B97D3, while being a dispersion-corrected functional, does not account for nonlocal orbital correlation, as in post-Hartree−Fock methods or double hybrid functionals. In the realm of DFT, one of the ways to introduce diffuseness is via a basis set. Therefore, employing diffuse basis functions in the middle layer is critical to ensuring compatible treatment of long-range dispersion interactions across layers. In future work, we plan to explore the use of range-separated functionals, especially those incorporating nonlocal dispersion effects in the middle layer, as an alternative approach to model long-range interactions. Our overall goal is to enhance MIM's ability to model both short-range and long-range interactions accurately with correlated wave function methods. To achieve this, the successful application of the MIM method will depend on a nuanced treatment of dispersion interactions across layers—particularly between the high and middle layers—to ensure the accuracy of the results.

We conducted additional tests using the Dunning cc-pVTZ basis, combined with DLPNO-MP2, in the high layer (Figure 4). Generally, the observed trends mirrored those identified using the def2-TZVP basis for correlated methods in the middle layer, yet notable was that all errors remained below 1 kcal mol^{-1}. This may be partially attributed to Dunning basis sets exhibiting higher compatibility with correlated methods. Interaction energies were also computed utilizing the highly accurate DLPNO−CCSD(T)/def2-TZVP theory, where once again smaller Pople basis sets marginally surpassed others in performance, as illustrated in Figure 5. Furthermore, the MIM3 errors, when employing DFT methods in the middle layer, paired with DLPNO-MP2/cc-pVTZ and DLPNO−CCSD(T)/def2-TZVP methods in the high layer, are illustrated in Figures S3 and S6 of the Supporting Information. We once again observed lower errors with diffuse basis sets, a result that not only aligns with but also reinforces a consistent trend noted in our previous findings, underscoring the vital role of such basis functions in elevating computational accuracy in the middle layer. The TightPNO setting was employed exclusively for the smallest basis sets while evaluating various theories for the middle layer, primarily due to its substantial computational cost. All MIM results displayed in Figures 2−5 were achieved with PM6-D3 in the low layer. Nonetheless, exploratory use of B97−3c (see Figures S1, S2, S4, S5, S7, and S8 in the Supporting Information) yielded results comparable to those with PM6-D3 and thus emerges as a viable alternative for low layer theory.

We conducted computations using the BZT-ITK set for added validation, in which the ligands differ by only two substituents, resulting in relatively consistent complex sizes (ranging between 350 and 450 atoms per complex). Consequently, the MAD of the binding energies—relative to the majority of middle-layer theories using correlated methods—remains notably below or close to 1 kcal mol^{-1} (see Figure 6). In contrast, the CDK2 protein−ligand set presents complexes varying more substantially in size, from approximately 250−500 atoms, owing to significant variations in ligand structures. This discrepancy not only potentially explains the slightly larger variations in MIM MADs for CDK2 but also strengthens our benchmark against molecular size variability.

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

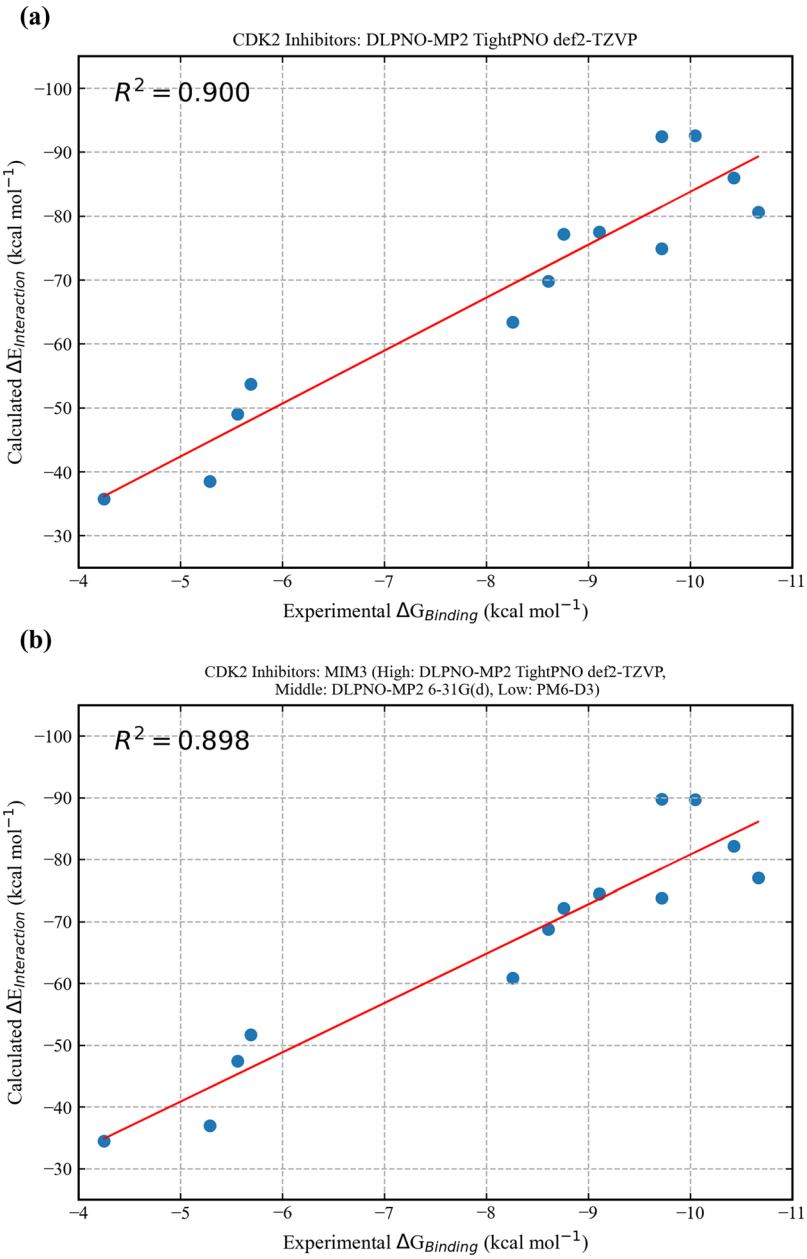


Figure 7. Correlation plots comparing experimental binding free energies ($\Delta G_{binding}$) with theoretical gas-phase interaction energies ($\Delta E_{interaction}$) for the CDK2 inhibitors data set,[56,66] using (a) DLPNO-MP2 TightPNO def2-TZVP and (b) MIM3 model (DLPNO-MP2 TightPNO def2-TZVP: DLPNO-MP2 6-31G(d): PM6-D3).

In our study, we also demonstrate the correlation (R^2) between theoretically computed DLPNO-MP2 interaction energies and their respective experimental counterparts, as depicted in Figures 7a and 8a. In this context, we note that the calculated *gas phase* interaction energies are plotted directly against experimentally determined binding free energies *in solution*. Nevertheless, the correlations for both protein−ligand sets are noteworthy—approximately 0.90 for CDK2 and 0.78 for BZT-ITK, demonstrating the validity of our premise that binding trends for different ligands at the same protein binding site can be reliably reproduced. Furthermore, the interaction energies computed using fragmentation via the MIM3 method (Figures 7b and 8b) yielded virtually identical correlations, demonstrating that fragmentation introduces very little error and affirming the utility of our MIM method for calculating protein−ligand binding energies. Additional correlation plots pertaining to other tested theories, namely, DLPNO-MP2/cc-pVTZ and DLPNO−CCSD(T)/def2-TZVP, can be found in the Supporting Information (refer to Figures S9 and S10).

In our study, we also performed a focused analysis to assess the effect of the TightPNO setting in the middle layer of our MIM calculations. This analysis revealed that including TightPNO in the middle layer can lead to marginal improvements in MIM3 MADs, especially when TightPNO is also used in the high layer or when DLPNO−CCSD(T) is the high-level method, as demonstrated in Figures 2, 5, 7, S1, and S7. In contrast, when the NormalPNO setting is used in the high layer in conjunction with DLPNO-MP2, introducing TightPNO in the middle layer results in slightly higher MIM3 MADs than those obtained without the TightPNO option, as shown in Figures 4 and S4. These findings underscore the importance of compatibility between the PNO settings in

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

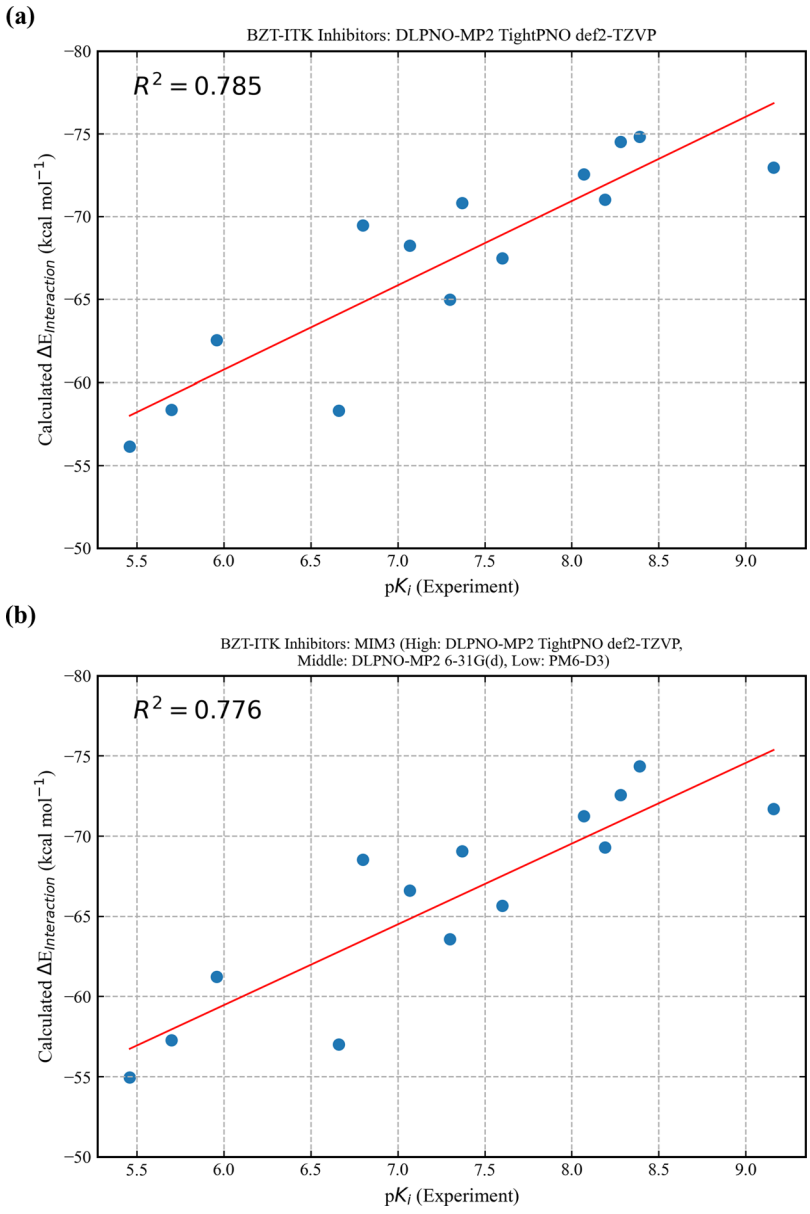


Figure 8. Correlation plots comparing experimental binding affinities (pK_i) with theoretical gas-phase interaction energies ($\Delta E_{interaction}$) for the BZT-ITK inhibitors data set,[59,67] using (a) DLPNO-MP2 TightPNO def2-TZVP and (b) MIM3 model (DLPNO-MP2 TightPNO def2-TZVP: DLPNO-MP2 6-31G(d): PM6-D3).

different layers. They suggest that a synergistic effect exists between the high and middle layer settings, where consistent levels of theory can enhance the overall accuracy of the MIM method. Our study indicates that while TightPNO can improve accuracy under certain conditions, its application must be carefully balanced with the computational settings in other layers to ensure optimal accuracy. Therefore, achieving the best MIM errors requires the high and middle layers to be compatible, particularly in terms of correlation energies.

To analyze the systematic errors, we have also included bar plots depicting the mean absolute errors (MAEs) in MIM3 binding energies in the Supporting Information (Figures S13–S25). These plots indicate that while the trends in MAEs generally mirror those observed for MADs, the error magnitudes are predictably larger and the patterns less consistent. Employing smaller Pople basis sets (6-31G and 6-31G(d)) yields lower MAEs, especially when paired with def2-TZVP in the high layer for both DLPNO-MP2 and DLPNO–CCSD(T) methods. Additionally, DLPNO-MP2 with cc-pVTZ in the high layer aligns well with DLPNO–CCSD(T) in the middle layer across all tested basis sets. When DFT (B97D3 functional) is used in the middle layer, the impact on the MAEs is even more pronounced. However, a notable improvement is observed when diffuse basis sets are employed with DFT in the middle layer, consistent with the earlier discussion in this context. This underscores the significance of ensuring compatibility between the MIM layer theories for accurate binding energy predictions.

5. CONCLUSIONS

This study meticulously benchmarks the MIM fragmentation-based method for executing post-Hartree–Fock computations on large molecules, presenting a clear strategy for selecting an appropriate level of theory for each MIM layer. Moreover, the

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774–2785

MIM method reveals a capacity to speed up interaction energy computations by canceling out subsystem energy terms shared between complex and protein molecules in the supermolecular equation. Our results indicate that MIM3 can be used to compute protein−ligand binding energies with minimal error (<1 kcal mol^{-1} compared to DLPNO-based post-Hartree−Fock methods), while significantly reducing computational costs. A key observation is that the smallest MIM interaction energy errors occurred when utilizing DLPNO methods with smaller Pople-style basis sets or the B97D3 DFT functional with diffuse basis sets in the middle layer. While our findings are specific to B97D3, we posit that similar trends could emerge with other dispersion-corrected functionals due to shared approaches in dispersion treatment. Hence, it is plausible that the insights gained from B97D3 might apply to functionals with similar dispersion corrections. Further research, however, is essential to confirm the broader applicability of our results. Importantly, we observed strong correlation between theoretically computed interaction energies and their experimental counterparts, with correlation coefficients of ∼0.90 and 0.78 for CDK2 and BZT-ITK protein−ligand sets, respectively. This underscores the practicality and precision of our MIM method in computing protein−ligand binding energies. The insights and recommendations from this benchmark study have the potential to be applicable to other large molecular systems, enabling further scientific investigations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01293.

Bar plots displaying MADs with alternative theoretical approaches, including B97−3c in the low layer and DFT in the middle layer; correlation plots comparing theoretical and experimental data; correlation plots illustrating the relationship between MIM3 and full molecule calculation results; a table detailing the full molecule energies for the data sets; and bar plots presenting MAEs for the MIM calculations (PDF)

Cartesian coordinates of the optimized geometries for protein−ligand complexes within the BZT-ITK data set (ZIP)

Cartesian coordinates of the optimized geometries for protein−ligand complexes within the CDK2 data set (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Krishnan Raghavachari − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0003-3275-1426; Email: kraghava@indiana.edu

Authors

Ankur K. Gupta − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States;* orcid.org/0000-0002-3128-9535

Sarah Maier − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States*

Bishnu Thapa − *Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01293

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Science Foundation grant CHE-2102583 at Indiana University.

■ REFERENCES

(1) Raghavachari, K.; Saha, A. Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev.* **2015**, *115* (12), 5643−5677.

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

(3) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112* (1), 632−672.

(4) Collins, M. A.; Bettens, R. P. A. Energy-Based Molecular Fragmentation Methods. *Chem. Rev.* **2015**, *115* (12), 5607−5642.

(5) Li, W.; Dong, H.; Ma, J.; Li, S. Structures and Spectroscopic Properties of Large Molecules and Condensed-Phase Systems Predicted by Generalized Energy-Based Fragmentation Approach. *Acc. Chem. Res.* **2021**, *54* (1), 169−181.

(6) Li, W.; Wang, Y.; Ni, Z.; Li, S. Cluster-in-Molecule Local Correlation Method for Dispersion Interactions in Large Systems and Periodic Systems. *Acc. Chem. Res.* **2023**, *56* (23), 3462−3474.

(7) Liu, J.; He, X. Recent advances in quantum fragmentation approaches to complex molecular and condensed-phase systems. *WIREs Comput. Mol. Sci.* **2023**, *13* (3), No. e1650.

(8) Rahalkar, A. P.; Katouda, M.; Gadre, S. R.; Nagase, S. Molecular tailoring approach in conjunction with MP2 and Ri-MP2 codes: A comparison with fragment molecular orbital method. *J. Comput. Chem.* **2010**, *31* (13), 2405−2418.

(9) Saha, A.; Raghavachari, K. Analysis of Different Fragmentation Strategies on a Variety of Large Peptides: Implementation of a Low Level of Theory in Fragment-Based Methods Can Be a Crucial Factor. *J. Chem. Theory Comput.* **2015**, *11* (5), 2012−2023.

(10) Sheng, Y.; Watanabe, H.; Maruyama, K.; Watanabe, C.; Okiyama, Y.; Honma, T.; Fukuzawa, K.; Tanaka, S. Towards good correlation between fragment molecular orbital interaction energies and experimental IC(50) for ligand binding: A case study of p38 MAP kinase. *Comput. Struct Biotechnol J.* **2018**, *16*, 421−434.

(11) Choi, J.; Kim, H.-J.; Jin, X.; Lim, H.; Kim, S.; Roh, I.-S.; Kang, H.-E.; No, K. T.; Sohn, H.-J. Application of the fragment molecular orbital method to discover novel natural products for prion disease. *Sci. Rep.* **2018**, *8* (1), 13063.

(12) Otsuka, T.; Okimoto, N.; Taiji, M. Assessment and acceleration of binding energy calculations for protein−ligand complexes by the fragment molecular orbital method. *J. Comput. Chem.* **2015**, *36* (30), 2209−2218.

(13) Okimoto, N.; Otsuka, T.; Hirano, Y.; Taiji, M. Use of the Multilayer Fragment Molecular Orbital Method to Predict the Rank Order of Protein−Ligand Binding Affinities: A Case Study Using Tankyrase 2 Inhibitors. *ACS Omega* **2018**, *3* (4), 4475−4485.

(14) Ricard, T. C.; Haycraft, C.; Iyengar, S. S. Adaptive, Geometric Networks for Efficient Coarse-Grained Ab Initio Molecular Dynamics with Post-Hartree−Fock Accuracy. *J. Chem. Theory Comput.* **2018**, *14* (6), 2852−2866.

(15) Yamamoto, S.; Saito, R.; Nakamura, S.; Sogawa, H.; Karpov, P.; Shulga, S.; Blume, Y.; Kurita, N. Proposal of Potent Inhibitors for a Bacterial Cell Division Protein FtsZ: Molecular Simulations Based on Molecular Docking and ab Initio Molecular Orbital Calculations. *Antibiotics (Basel)* **2020**, *9* (12), 846 DOI: 10.3390/antibiotics9120846.

(16) Meyer, W. Theory of self-consistent electron pairs. An iterative method for correlated many-electron wavefunctions. *J. Chem. Phys.* **1976**, *64* (7), 2901−2907.
(17) Meyer, W. PNO−CI Studies of electron correlation effects. I. Configuration expansion by means of nonorthogonal orbitals, and application to the ground state and ionized states of methane. *J. Chem. Phys.* **1973**, *58* (3), 1017−1035.
(18) Ahlrichs, R.; Kutzelnigg, W. Direct Calculation of Approximate Natural Orbitals and Natural Expansion Coefficients of Atomic and Molecular Electronic Wavefunctions. II. Decoupling of the Pair Equations and Calculation of the Pair Correlation Energies for the Be and LiH Ground States. *J. Chem. Phys.* **1968**, *48* (4), 1819−1832.
(19) Edmiston, C.; Krauss, M. Pseudonatural Orbitals as a Basis for the Superposition of Configurations. I. He2+. *J. Chem. Phys.* **1966**, *45* (5), 1833−1839.
(20) Edmiston, C.; Krauss, M. Configuration-Interaction Calculation of H3 and H2. *J. Chem. Phys.* **1965**, *42* (3), 1119−1120.
(21) Meyer, W.; Jakubetz, W.; Schuster, P. Correlation effects on energy curves for proton transfer. The cation [H5O2]+. *Chem. Phys. Lett.* **1973**, *21* (1), 97−102.
(22) Meyer, W. PNO-CI and CEPA studies of electron correlation effects. *Theoretica chimica acta* **1974**, *35* (4), 277−292.
(23) Yang, J.; Chan, G. K.-L.; Manby, F. R.; Schütz, M.; Werner, H.-J. The orbital-specific-virtual local coupled cluster singles and doubles method. *J. Chem. Phys.* **2012**, *136* (14), 144105.
(24) Schütz, M.; Yang, J.; Chan, G. K.-L.; Manby, F. R.; Werner, H.-J. The orbital-specific virtual local triples correction: OSV-L(T). *J. Chem. Phys.* **2013**, *138* (5), No. 054109.
(25) Adamowicz, L.; Bartlett, R. J. Optimized virtual orbital space for high-level correlated calculations. *J. Chem. Phys.* **1987**, *86* (11), 6314−6324.
(26) Adamowicz, L.; Bartlett, R. J.; Sadlej, A. J. Optimized virtual orbital space for high-level correlated calculations. II. Electric properties. *J. Chem. Phys.* **1988**, *88* (9), 5749−5758.
(27) Kraus, M.; Pitoňák, M.; Hobza, P.; Urban, M.; Neogrády, P. Highly correlated calculations using optimized virtual orbital space with controlled accuracy. Application to counterpoise corrected interaction energy calculations. *Int. J. Quantum Chem.* **2012**, *112* (4), 948−959.
(28) Neogrády, P.; PitoňáK, M.; Urban, M. Optimized virtual orbitals for correlated calculations: an alternative approach. *Mol. Phys.* **2005**, *103* (15−16), 2141−2157.
(29) Adamowicz, L. Optimized virtual orbital space (OVOS) in coupled-cluster calculations. *Mol. Phys.* **2010**, *108* (21−23), 3105−3112.
(30) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **2020**, *152* (22), 224108.
(31) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8* (1), No. e1327.
(32) Pinski, P.; Riplinger, C.; Valeev, E. F.; Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. I. An efficient and simple linear scaling local MP2 method that uses an intermediate basis of pair natural orbitals. *J. Chem. Phys.* **2015**, *143* (3), No. 034108.
(33) Pavošević, F.; Pinski, P.; Riplinger, C.; Neese, F.; Valeev, E. F. SparseMaps—A systematic infrastructure for reduced-scaling electronic structure methods. IV. Linear-scaling second-order explicitly correlated energy with pair natural orbitals. *J. Chem. Phys.* **2016**, *144* (14), 144109.
(34) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method [DLPNO-CCSD(T)]. *J. Chem. Phys.* **2018**, *148* (1), No. 011101.
(35) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144* (2), No. 024109.
(36) Pavošević, F.; Peng, C.; Pinski, P.; Riplinger, C.; Neese, F.; Valeev, E. F. SparseMaps—A systematic infrastructure for reduced scaling electronic structure methods. V. Linear scaling explicitly correlated coupled-cluster method with pair natural orbitals. *J. Chem. Phys.* **2017**, *146* (17), 174108.
(37) Guo, Y.; Sivalingam, K.; Valeev, E. F.; Neese, F. SparseMaps—A systematic infrastructure for reduced-scaling electronic structure methods. III. Linear-scaling multireference domain-based pair natural orbital N-electron valence perturbation theory. *J. Chem. Phys.* **2016**, *144* (9), No. 094111.
(38) Gupta, A. K.; Thapa, B.; Raghavachari, K. Exploring Reaction Energy Profiles Using the Molecules-in-Molecules Fragmentation-Based Approach. *J. Chem. Theory Comput.* **2019**, *15* (7), 3991−4002.
(39) 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* (5), 1336−1343.
(40) Ni, Z.; Guo, Y.; Neese, F.; Li, W.; Li, S. Cluster-in-Molecule Local Correlation Method with an Accurate Distant Pair Correction for Large Systems. *J. Chem. Theory Comput.* **2021**, *17* (2), 756−766.
(41) 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* (10), 5143−5155.
(42) Thapa, B.; Raghavachari, K. Energy Decomposition Analysis of Protein−Ligand Interactions Using Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Inf. Model.* **2019**, *59* (8), 3474−3484.
(43) Thapa, B.; Erickson, J.; Raghavachari, K. Quantum Mechanical Investigation of Three-Dimensional Activity Cliffs Using the Molecules-in-Molecules Fragmentation-Based Method. *J. Chem. Inf. Model.* **2020**, *60* (6), 2924−2938.
(44) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives1Dedicated to Professor Keiji Morokuma in celebration of his 65th birthday.1. *Journal of Molecular Structure: THEOCHEM* **1999**, *461−462*, 1−21.
(45) Jose, K. V. J.; Raghavachari, K. Evaluation of Energy Gradients and Infrared Vibrational Spectra through Molecules-in-Molecules Fragment-Based Approach. *J. Chem. Theory Comput.* **2015**, *11* (3), 950−961.
(46) Gupta, A. K.; Gamoke, B. C.; Raghavachari, K. Interaction−Deletion: A Composite Energy Method for the Optimization of Molecular Systems Selectively Removing Specific Nonbonded Interactions. *J. Phys. Chem. A* **2021**, *125* (21), 4668−4682.
(47) Brandenburg, J. G.; Hochheim, M.; Bredow, T.; Grimme, S. Low-Cost Quantum Chemical Methods for Noncovalent Interactions. *J. Phys. Chem. Lett.* **2014**, *5* (24), 4275−4284.
(48) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97−3c: A revised low-cost variant of the B97-D density functional method. *J. Chem. Phys.* **2018**, *148* (6), No. 064104.
(49) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree−Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree−Fock exchange. *Chem. Phys.* **2009**, *356* (1), 98−109.
(50) Feyereisen, M.; Fitzgerald, G.; Komornicki, A. Use of approximate integrals in ab initio theory. An application in MP2 energy calculations. *Chem. Phys. Lett.* **1993**, *208* (5), 359−363.
(51) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32* (7), 1456−1465.
(52) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27* (15), 1787−1799.
(53) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *J. Chem. Phys.* **1997**, *107* (20), 8554−8560.

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785

(54) 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* (12), 1173−1213.
(55) *Gaussian 16 Rev. C.01*; Gaussian, Inc.: Wallingford, CT, 2016.
(56) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51* (13), 3661−3680.
(57) Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E.; et al. Identification of N-(4-piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a novel cyclin dependent kinase inhibitor using fragment-based X-ray crystallography and structure based drug design. *J. Med. Chem.* **2008**, *51* (16), 4986−4999.
(58) Heifetz, A.; Trani, G.; Aldeghi, M.; MacKinnon, C. H.; McEwan, P. A.; Brookfield, F. A.; Chudyk, E. I.; Bodkin, M.; Pei, Z. H.; Burch, J. D.; et al. Fragment Molecular Orbital Method Applied to Lead Optimization of Novel Interleukin-2 Inducible T-Cell Kinase (ITK) Inhibitors. *J. Med. Chem.* **2016**, *59* (9), 4352−4363.
(59) MacKinnon, C. H.; Lau, K.; Burch, J. D.; Chen, Y.; Dines, J.; Ding, X.; Eigenbrot, C.; Heifetz, A.; Jaochico, A.; Johnson, A.; et al. Structure-based design and synthesis of potent benzothiazole inhibitors of interleukin-2 inducible T cell kinase (ITK). *Bioorg. Med. Chem. Lett.* **2013**, *23* (23), 6331−6335.
(60) *Molecular Operating Environment (MOE)*; Chemical Computing Group: Montreal, QC, 2018.
(61) Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins* **2009**, *75* (1), 187−205.
(62) Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. *J. Chem. Theory Comput* **2014**, *10* (10), 4515−4534.
(63) Gerber, P. R.; Muller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput. Aided Mol. Des* **1995**, *9* (3), 251−268.
(64) 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.; et al. The ONIOM Method and Its Applications. *Chem. Rev.* **2015**, *115* (12), 5678−5796.
(65) Morokuma, K.; Musaev, D. G.; Vreven, T.; Basch, H.; Torrent, M.; Khoroshun, D. V. Model studies of the structures, reactivities, and reaction mechanisms of metalloenzymes. *IBM J. Res. Dev.* **2001**, *45* (3.4), 367−395.
(66) Wyatt, P. G.; Woodhead, A. J.; Berdini, V.; Boulstridge, J. A.; Carr, M. G.; Cross, D. M.; Davis, D. J.; Devine, L. A.; Early, T. R.; Feltell, R. E.; et al. Identification of N-(4-Piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), a Novel Cyclin Dependent Kinase Inhibitor Using Fragment-Based X-Ray Crystallography and Structure Based Drug Design. *J. Med. Chem.* **2008**, *51* (16), 4986−4999.
(67) Heifetz, A.; Trani, G.; Aldeghi, M.; MacKinnon, C. H.; McEwan, P. A.; Brookfield, F. A.; Chudyk, E. I.; Bodkin, M.; Pei, Z.; Burch, J. D.; et al. Fragment Molecular Orbital Method Applied to Lead Optimization of Novel Interleukin-2 Inducible T-Cell Kinase (ITK) Inhibitors. *J. Med. Chem.* **2016**, *59* (9), 4352−4363.

https://doi.org/10.1021/acs.jctc.3c01293
J. Chem. Theory Comput. 2024, 20, 2774−2785