

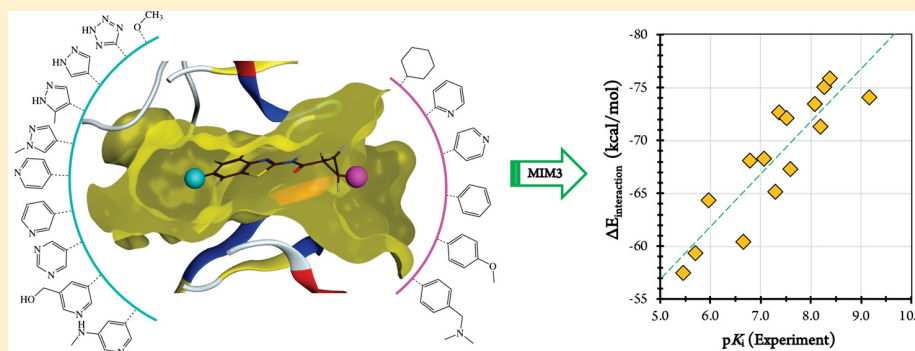
Theoretical Study of Protein–Ligand Interactions Using the Molecules-in-Molecules Fragmentation-Based Method

Bishnu Thapa,[†] Daniel Beckett,[†] Jon Erickson,[‡] and Krishnan Raghavachari^{*,†}

[†]Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States

[‡]Lilly Research Laboratories, Eli Lilly & Co., Indianapolis, Indiana 47285, United States

Supporting Information



ABSTRACT: We have recently significantly expanded the applicability of our Molecules-in-Molecules (MIM) fragmentation method to large proteins by developing a three-layer model (MIM3) in which an accurate quantum-mechanical method is used in conjunction with a cost-effective, dispersion-corrected semiempirical model to overcome previous computational bottlenecks. In this work, we develop MIM3 as a structure-based drug design tool by application of the methodology for the accurate calculation of protein–ligand interaction energies. A systematic protocol is derived for the determination of the geometries of the protein–ligand complexes and to calculate their accurate interaction energies in the gas phase using MIM3. We also derive a simple and affordable procedure based on implicit solvation models and the ligand solvent-accessible surface area to approximate the ligand desolvation penalty in gas-phase interaction energy calculations. We have carefully assessed how closely such interaction energies, which are based on a single protein–ligand conformation, display correlations with the experimentally determined binding affinities. The performance of MIM3 was evaluated on a total of seven data sets comprising 89 protein–ligand complexes, all with experimentally known binding affinities, using a binding pocket involving a quantum region ranging in size from 250 to 600 atoms. The dispersion-corrected B97-D3BJ density functional, previously known to perform accurately for calculations involving non-covalent interactions, was used as the target level of theory for this work, with dispersion-corrected PM6-D3 as the semiempirical low level to incorporate the long-range interactions. Comparing directly to the experimental binding potencies, we obtain impressive correlations over all seven test sets, with an R^2 range of 0.74–0.93 and a Spearman rank correlation coefficient (ρ) range of 0.83–0.93. Our results suggest that protein–ligand interaction energies are useful in predicting binding potency trends and validate the potential of MIM3 as a quantum-chemical structure-based drug design tool.

1. INTRODUCTION

The main goal of drug development is to design a candidate molecule that strongly and stereospecifically binds to a target receptor without interfering with the biological activities of other systems.^{1,2} Screening several thousand drug candidates to obtain a few tens of potential drugs with wet-lab experiments alone is virtually impossible due to the associated time and cost. Complementing the experiments with virtual database screening and simulations using computers is thus an indispensable tool in reducing cost and accelerating the drug discovery process.^{3,4} Consequently, numerous computational protocols have been developed to accurately predict drug–receptor binding affinities.⁵ The most common protocols involve ranking of ligands on the basis of scores obtained from

knowledge-based or empirical scoring functions or from energy functions based on molecular mechanics (MM) force fields within a single, target, binding pocket.⁶ Although such methods are very efficient and extensively used in predicting ligand binding modes, the accuracy of the estimated binding affinities may be limited by the difficulties in parametrization of inherently quantum-mechanical effects.^{6–9}

The most accurate theoretical approaches for estimating protein–ligand (P–L) binding affinities are based on free-energy perturbation (FEP), thermodynamic integration (TI), and related approaches.^{10,11} The application of such methods,

Received: June 1, 2018

Published: September 28, 2018

however, is limited by the high computational cost required to extensively sample all of the intermediate configurational states.^{12–14} Consequently, many approximate but more efficient methods (e.g., the linear-response approximation (LRA), the semimacroscopic protein–dipole Langevin–dipole approach (PDL/D-S-LRA), linear interaction energy (LIE) and molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) approaches) that also employ sampling, but only around the reactants and products, have been suggested.^{15–20} In addition, these methods have the advantage of fast energy evaluation using molecular mechanics force fields that can be parametrized for any specific system incorporating solvation effects. Although such methods are extensively used in estimating P–L binding affinities, a rigorous description of electronic effects that play a critical role in intermolecular interactions (e.g., charge-transfer, π – π interactions and many-body effects) is still difficult to capture with force field parameters.^{21,22} In this context, a high-level quantum-mechanical (QM) treatment can be very useful to correct for the deficiencies in the classical description of such interactions.^{23–25}

While the QM treatment of a ligand–receptor complex provides a great deal of information about the nature of their interactions, the large size of protein active sites with several hundreds of atoms makes it challenging, if not impossible, to perform a direct computation on the entire system using accurate ab initio QM methods.²⁶ Moreover, the ideal goal of performing such computations for several thousand snapshots from a simulation, that is necessary to correctly sample the protein and ligand conformations during binding, is still impossible even when steps are taken to reduce the scaling of expensive methods.

The computational cost associated with brute force ab initio calculations can be significantly reduced by employing more approximate yet quantum-based methods.^{27–30} For example, a number of semiempirical approaches exploiting the AM1, PM3, and PM6 Hamiltonians as well as density-functional-based tight-binding (DFTB) models have been applied to the calculation of P–L binding affinities.^{30–35} In addition, several studies have used the QM/MM approach to calculate P–L binding affinities.^{6,36,37} In the QM/MM methodology, the ligand and a few of the closest residues are treated with a QM method, leaving the rest of the protein to be calculated using MM.^{27,34,38} The QM region is mostly calculated using density functional theory (DFT) with small basis sets or with semiempirical methods. The agreement between the QM/MM and experimental binding energies is somewhat mixed: correlation coefficients (*R*) range from 0.17 to 0.96.⁶ Some recent studies have also employed ONIOM as a QM:QM or QM:QM:QM approach to calculate binding affinities.^{39–42} An enhanced correlation with the experiments can be expected with an increase in the size of the QM region in QM/MM binding energy calculations; the same effect can be seen when increasing the size of the high level in an ONIOM QM:QM calculation. As the region including the most critical residues near the ligand is typically inside a radius of 4 to 6 Å and contains more than 300 atoms, it is too large, in general, for full QM treatment at a highly correlated level of theory. With some occasional success in computing accurate binding affinities, the QM/MM approach is mostly limited to the qualitative description of ligand–receptor interactions.

A more attractive and fruitful strategy to calculate accurate binding energies by including a larger section of the P–L

complex quantum mechanically involves the use of linear-scaling fragmentation-based methods.^{6,43} In fragmentation-based methods, a large molecule is decomposed into many smaller subsystems, QM calculations are performed independently on each subsystem, and the energies from the subsystem calculations are combined to obtain the total energy. Fragmentation-based approaches have been used to explore P–L complexes, and several encouraging reports on binding energy calculations and quantitative descriptions of interaction energies using such methods have been published.^{44–58} The solvation free energy contribution to the total binding energy has been included in some cases by using implicit solvation (e.g., generalized Born (GB), polarizable continuum model (PCM), and conductor-like screening model for real solvents (COSMO-RS)). Recently, a few attempts have also been made to include entropic contributions to the binding free energy via conformational sampling, normal-mode analysis, and/or other approximations (e.g., number of rotatable single bonds).^{48,59} Most of these studies calculate the interaction energies using a single P–L complex obtained either from a short MD simulation or via geometry minimization, although proper conformational sampling may be required to obtain reliable results for absolute binding affinities.⁶ Only a few studies have considered multiple conformations of P–L complexes in the calculations due to the high cost associated with such sampling.^{55,58}

Among the many fragmentation-based methods, the fragment molecular orbital (FMO) technique, molecular fragmentation with conjugate caps (MFCC), and related approaches have been used in studying P–L interactions.^{6,43,60} The FMO method has been used both as a predictive tool for the determination of binding affinities and as an analytical tool in understanding the nature of the interactions within the active site.^{6,43} The MFCC-related approaches have been primarily applied to describe intermolecular interactions. For a comprehensive description of the current developments in fragmentation-based methods and their application in P–L interactions, we refer readers to some recent reviews.^{6,43,60} Interestingly, most fragmentation-based methods calculate the total interaction energy as the sum of contributions from individual residue–ligand interactions and do not account for many-body interactions. Hence, they contain some inherent errors due to nonadditive effects. To address this issue in the MFCC-related approaches, Antony and Grimme⁶¹ suggested including five or six residues per fragment in routine applications to reduce the associated fractionation errors. Such an increase, however, in fragment size (five or six residues plus the ligand, per fragment) steeply increases the associated computational cost. In the FMO method, interaction energies are calculated using Kitaura–Morokuma energy decomposition analysis (KM-EDA).⁶² A common issue associated with the KM-EDA approach is that it does not work well with large, diffuse basis sets and at short intermolecular distances due to neglect of the Pauli exclusion principle in the external potential.^{63,64} Hence, most P–L interaction studies using the FMO method have commonly employed small basis sets (e.g., 6-31G) with or without polarization functions.^{6,43} It is important to note the necessity of large basis sets are needed to correctly describe nonbonded interactions that typically determine the majority of a ligand's binding affinity. Overall, the errors associated with the accuracy of a QM method or with fragmentation protocols can be significant and need to be

properly taken into account when computing absolute binding energies.

We recently reported a rigorous calibration of our multilayer molecules-in-molecules (MIM) method using various fragmentation strategies and combinations of levels of theory specifically applicable for large, biologically relevant polypeptides containing many nonlocal interactions.⁶⁵ By carefully selecting the fragmentation schemes and combinations of high and low levels of theory, we showed that the MIM method is capable of predicting the total energy to within 0–2 kcal/mol of the full, unfragmented energy for molecules containing up to 700+ atoms. More importantly, since the largest subsystems in the high layer were only pentamers (vide infra), accurate theoretical methods can be used with significantly larger basis sets without making the high-level calculation a computational bottleneck.

In the current study, we present the first application of the MIM method to the computation of interaction energies associated with protein–ligand binding. Our approach is to develop rigorous, computationally efficient protocols that can be applied in a systematic manner to a broad variety of P–L systems. The performance of our approach will be assessed using diverse sets of ligands bound to different protein receptors with known experimental binding affinities.

The calculation of protein–ligand binding affinities can be highly challenging, and it is important to define the scope of the applications that are practical and can be carried out in a cost-effective manner. Our goal is not only to reduce the scaling of quantum methods but to explore potential approximations to the expensive problems plaguing the calculation of P–L complexes. The main purpose of this first study is to present a systematic and computationally efficient drug design protocol for estimating the relative strength of P–L interactions for a set of structurally similar ligands using MIM. In addition, our proposed protocol is a full QM-based methodology that can be rigorously improved to accurately calculate binding energies. To make a broadly applicable protocol, a few careful simplifications have been made in P–L structure preparation.

First, the interaction energy is calculated using a single, representative P–L complex obtained from the experimental crystal structure, docking process, or minimization (vide infra). Since we aim to estimate the relative trend in interaction energies of a set of similar ligands with small modifications rather than the absolute binding energies, this approach avoids the high computational cost associated with protein conformational sampling. In this context, to minimize the risk of the structure being trapped in a local minimum, ligand analogues are generated from a structurally similar ligand that is cocrystallized with the receptor, assuming their binding modes to be similar (vide infra). We do realize, however, that such a model may not be adequate when the ligand structure changes substantially or when the flexibility of the protein plays an important role in the binding motif.

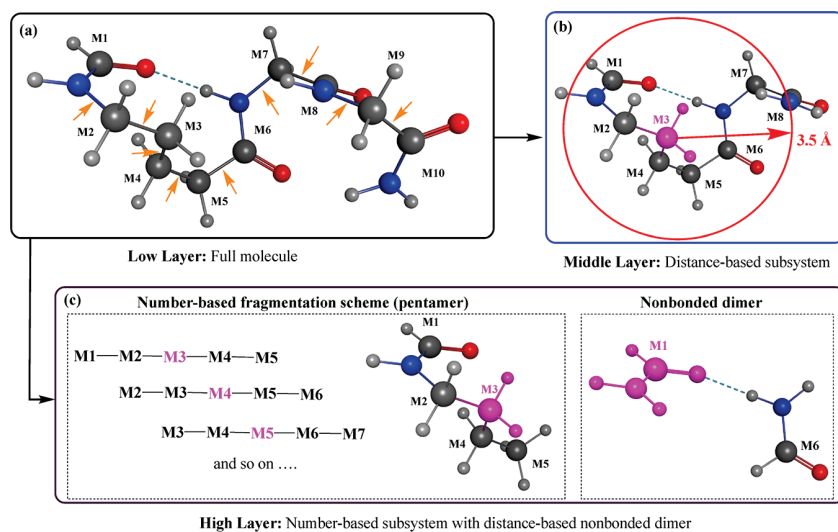
Second, we calculate only the gas-phase binding energies in this work and do not include solvation effects on the protein–ligand binding. This model is based on the following assumption, whose validity will be tested: as the modified ligands closely resemble their starting cocrystallized structure bound to the same protein, the difference in protein desolvation and entropic contributions to the total binding energy are expected to be systematic and cancel out when relative energies are considered. In such cases, the P–L binding is expected to be mainly governed by the contribution

from the total electronic energy, so calculation of the total electronic energies should be sufficient to predict the relative binding strengths of the ligands. However, for ligand modifications involving a significant polarity change or ligands that are partially solvent-exposed, a proper accounting of the ligand desolvation energy might be necessary to estimate accurate binding affinities (vide infra).

Our approach to the calculation of P–L interactions is carefully analyzed through an assessment of the correlation between theoretical interaction energies and the experimental binding potencies (which would include all of the aforementioned effects, however minor). Since the calculations are performed without incorporating the change in entropy and the solvation energy of the complex, the total interaction energies are expected to be significantly overestimated compared to the experimental values (by as much as an order of magnitude). Nevertheless, this study also serves as a good benchmark on how far we can go with a single P–L structure and accurate QM calculations. Though the current study is restricted to QM calculations on the protein and P–L complex in the gas phase to minimize the computational cost, other aspects such as conformational sampling, QM geometry optimization, and vibrational frequency analyses of those systems in an aqueous medium are proactively under development in our group and will be considered in future studies using new protocols.

2. COMPUTATIONAL METHODS

2.1. Structure Preparation. To prepare our test sets, published crystal structures of proteins and cocrystallized ligands were used as starting points. Each ligand analogue was generated from a closely related cocrystallized ligand and docked into the published crystal structure under the assumption that similar ligands bind to the receptor in similar binding modes. The flexible alignment module, implemented in the Molecular Operating Environment (MOE) program (version 2016.08),⁶⁶ was used to align the modified ligands with the cocrystallized starting point to ensure that key interactions, such as classical hydrogen bonds, were retained during modification. In cases where the added moieties were larger than those on the original ligand, a conformational search was performed (within a 5.0 kcal/mol energy window) using the LowModeMD⁶⁷ module implemented in MOE while keeping the backbone as close as possible to the orientation of the parent ligand. Among the lowest-energy ligand conformers, structures having any unphysical overlap with the receptor when placed in the binding cavity were discarded. The remaining conformers were further refined by performing single-point QM calculations at the B97-D3BJ/6-311++G(d,p) level, and the lowest-energy structure was used in calculating the interaction energy. In doing so, a careful analysis of the ligand structure and nearby residues was carried out to ensure the possibility of forming favorable interactions (e.g., hydrogen bonding, π – π stacking) by the modified part of the ligand with the receptor. It is important to note that the modifications of the ligand were typically small compared to the cocrystallized ligand and that not many conformations were possible due to the geometrical constraints in the binding cavity. Missing hydrogen atoms in the crystal structures were added at pH 7.0 with the Protonate3D⁶⁸ tool as implemented in MOE, which recursively samples different orientations of added hydrogen atoms to attain the lowest-energy structure while maximizing the possibility of hydrogen bonding. The protonation of

Scheme 1. Illustration of Fragmentation Schemes Used in the MIM3 Calculations^a

^aIn (a), the orange arrows pointing toward bonds designate bonds broken during MIM fragmentation. The full molecule is used in the low-layer calculation. In (b) and (c), the monomer starting a subsystem is colored in pink. In (b), the red circle represents the cutoff radius of 3.5 Å specified for distance-based fragmentation in the middle layer. In (c), the box on the left represents the number-based fragmentation scheme (pentamers) and the box on the right represents a distance-based nonbonded dimer formed in the high layer.

histidine residues present within 5 Å of the ligand was further determined by analyzing the possibility of hydrogen-bond formation with nearby residues. The assigned states are given in Table S1. No unusual protonation states of other charged residues were observed. To temper electrostatic interactions during minimization, the generalized Born/volume integral implicit solvent model was used with an external solvent dielectric constant of 80 and an internal dielectric constant of 2 for the binding pocket. Finally, each protein–ligand complex was minimized in MOE with the AMBER10:EHT force field^{69,70} (the Amber10 force field to describe the protein and extended Hückel theory for small molecules in order to include electronic effects) under a 0.5 Å restraint for every atom with respect to the starting structure.

Seven test sets totaling 89 P–L complexes with experimentally known binding affinities were used in this study (Tables S2–S8). The tested proteins include well-known receptors and drug targets: interleukin-2-inducible T-cell kinase (ITK) (set-I, set-II, and set-VII), cyclin-dependent kinase 2 (CDK2) (set-III), avidin (set-IV), and a G-protein-coupled receptor (P2Y12) (set-V and set-VI). In set-I, a benzothiazole (BZT)-based ligand (11) cocrystallized with ITK was taken from the RCSB Protein Data Bank (PDB ID 4MF0) and modified to obtain 14 other complexes (Table S2) following the minimization protocol discussed above. Set-II includes 11 indazole (IND)-based inhibitors complexed with ITK; the cocrystal of 2g with ITK (PDB ID 4PP9) was processed to obtain protein–ligand complexes for 2a–2h and 2j. In the case of 2i and 2k, the cocrystal of 2k and ITK (PDB ID 4PPA) was used as the starting structure because of a difference in ligand–protein hydrogen bonds compared with the rest of the set. Set-III comprises a series of 13 CDK2 complexes with experimentally reported crystal structures (see Table S4 for the PDB IDs). A cocrystallized structure of an avidin–biotin complex isolated from egg white (4n, PDB ID 1AVD) was used to obtain the 14 protein–ligand complexes of set-IV. Two sets of ligand variants, namely, the 3-azetidinyl

series (set-V, 11 ligands) and 4-piperidinyl series (set-VI, seven ligands), complexed with a human G-protein-coupled receptor, P2Y12, were obtained by modifying complex 6f (PDB ID 4NTJ). Lastly, a set of 18 sulfonylpyridine (SAP)-based ligands complexed with the ITK receptor (set-VII) were obtained by modifying the cocrystallized ligand 7i (PDB ID 4QD6).

All residues and water molecules within 5.0 Å of the ligand (a total of 250–600 atoms) were included in the QM calculations. Dangling bonds were saturated with hydrogen atoms using MOE. All of the calculations were performed in the gas phase, and solvation effects were not included in this work. Due to the overestimation of electrostatic interactions in the gas phase, charged ligands and all charged residues (Lys, Arg, His, Asp, and Glu) were neutralized to better match the stabilization seen in solution. This is an appropriate approximation for the calculation of interaction energies in the gas phase, as validated by our findings in this work (vide infra). We do realize, however, that this is a limitation of our approach, and future work in obtaining solution Gibbs free energies will deal with the necessity of proper treatment of charged residues along with implicit solvation.

2.2. MIM Protocol. The inner workings of the MIM fragmentation method have been detailed in our previous publications.^{65,71} Here we present only a general outline describing the details necessary to understand our chosen protocol. In MIM, proteins are divided into small, primitive fragments (monomers) by cutting single heavy atom–heavy atom bonds, as shown in Scheme 1. Peptide bonds are left unbroken because of the partial double-bond character of these moieties, as detailed in our previous work.^{65,72} Each monomer initiates an overlapping primary subsystem based on a selected fragmentation scheme, and all of the cut bonds are saturated with hydrogen link atoms. In multilayer MIM, cheaper QM methods are used to describe environmental polarization effects as well as long-range interactions not included in each specific subsystem. In this study, we employ three-layer MIM (denoted as MIM3), in which a number-based fragmentation

augmented with nonbonded dimers based on a cutoff distance is used in the high layer, a distance-based fragmentation is used in the middle layer, and the full molecule is calculated using a semiempirical method in the low layer. This fragmentation scheme is the prescribed fragmentation scheme for large proteins from our previous study.⁶⁵ In a number-based scheme, primary subsystems are generated by combining η covalently bonded monomers. In this case, we employ a pentamer scheme ($\eta = 5$), in which the pentamers are formed by combining two covalently bonded monomers on either side of the starting monomer, preserving the local bonded environment. Additionally, some of the nonbonded interactions are included in primary subsystems as distance-based dimers within a cutoff distance d ; here we chose d to be 3.5 Å in accordance with our previous work (denoted as NSD). The nonbonded dimers are formed in order to incorporate some of the short-range nonbonded interactions (e.g., hydrogen bonding, π - π stacking) that would normally be absent from subsystems in the high layer. To capture some of the important long-range effects that may be neglected by the NSD scheme in the high layer, we employ a distance-based middle layer with a cutoff distance of 3.5 Å. As in the high layer, each monomer starts a subsystem in the middle layer. All of the non-overlapping monomers within the specified distance parameter of the starting monomer are combined to form a subsystem. This results in larger subsystems to be treated with an intermediate level of theory. In the low layer, the full molecule is used to include the interactions not captured by the middle layer. This fragmentation scheme is illustrated in Scheme 1.

Once the primary subsystems are established, derivative subsystems are obtained from the overlap between the primary subsystems, and the energies summed according to the inclusion-exclusion principle. Three levels of theory ("high", "medium", and "low") are used in MIM3, and the full fragmentation energies are summed as in eq 1:

$$E^{\text{MIM3}} = E_{\text{high}}^r - (E_{\text{med}}^r - E_{\text{med}}^{r'}) - (E_{\text{low}}^{r'} - E_{\text{low}}^R) \quad (1)$$

where r and r' (with $r < r'$) symbolically represent the relative sizes of the systems; E_{high}^r and E_{med}^r are the total energies obtained with the NSD scheme calculated at the high and medium levels of theory, respectively; $E_{\text{med}}^{r'}$ and $E_{\text{low}}^{r'}$ are the total energies obtained with distance-based fragmentation at the medium and low levels of theory, respectively; and E_{low}^R is the total energy of the full molecule ($R = \infty$) calculated at the low level of theory. Since the sizes of both the high- and medium-level subsystems are independent of the size of the full system, using an inexpensive semiempirical calculation (vide infra) for the full molecule makes the overall scaling of the MIM3 model with the size of the full system effectively linear.

2.3. Interaction Energy Calculation. The interaction energy between a receptor and a ligand in a complex is calculated as

$$\Delta E_{\text{interaction}} = E^{\text{complex}} - E^{\text{protein}} - E^{\text{ligand}} \quad (2)$$

where E^{complex} , E^{protein} , and E^{ligand} are the dispersion-corrected gas-phase electronic energies of the protein-ligand complex, protein, and ligand, respectively. The outlined fragmentation protocol was used for the protein with and without the bound ligand.

The total interaction energy including the contribution of the ligand desolvation energy (the penalty of abstracting the ligand from the solvent) is calculated as

$$\Delta E'_{\text{interaction}} = \Delta E_{\text{interaction}} + \Delta E_{\text{(desolv)}}^{\text{ligand}} \quad (3)$$

where $\Delta E_{\text{interaction}}$ is the interaction energy calculated in the gas phase and $\Delta E_{\text{(desolv)}}^{\text{ligand}}$ is the desolvation energy of the ligand in the formation of the P-L complex. $\Delta E_{\text{(desolv)}}^{\text{ligand}}$ is obtained with eq 4:

$$\Delta E_{\text{(desolv)}}^{\text{ligand}} = -\gamma \Delta E_{\text{(solv)}}^{\text{ligand}} \quad (4)$$

where $\Delta E_{\text{(solv)}}^{\text{ligand}}$ is the solvation energy of the unbound ligand and γ is the fraction of ligand surface area in contact with the receptor in the P-L complex. γ is calculated as

$$\gamma = \frac{\text{CA}_{\text{ligand-receptor}}}{\text{SASA}_{\text{ligand}}} \quad (5)$$

in which $\text{CA}_{\text{ligand-receptor}}$ is the contact area of the ligand with the receptor, given by

$$\text{CA}_{\text{ligand-receptor}} = \frac{\text{SASA}_{\text{ligand}} + \text{SASA}_{\text{protein}} - \text{SASA}_{\text{complex}}}{2} \quad (6)$$

where $\text{SASA}_{\text{ligand}}$, $\text{SASA}_{\text{protein}}$, and $\text{SASA}_{\text{complex}}$ are the solvent-accessible surface areas of the free ligand, free protein, and P-L complex, respectively. The SASA is calculated using the VMD package⁷³ with default atomic radii and a probe radius of 1.40 Å, corresponding to water. Thus, the desolvation energy of the ligand is approximately taken into account by this simple scaling procedure.

The dispersion-corrected B97-D3BJ density functional (the B97 functional^{74,75} with Grimme's D3 dispersion correction⁷⁶ and Becke-Johnson damping⁷⁷) with the 6-311++G(d,p) basis set was used as the high level of theory. To avoid discrepancies between different density functionals, B97-D3BJ with the smaller 6-31+G(d) basis set was used as the intermediate level of theory.⁷⁸⁻⁸² PM6-D3⁸³ was used as the low level of theory in the MIM3 calculations. The solvation energy of the ligand in aqueous solution was calculated using the SMD⁸⁴ implicit polarizable continuum solvation model at the B97-D3BJ/6-311++G(d,p) level. All of the DFT calculations were performed using the Gaussian 16 program suite,⁸⁵ and our MIM external program was used to generate and sum the fragmented systems.

3. RESULTS AND DISCUSSION

3.1. Fragmentation versus High Level. In our most recent publication with MIM, we performed a careful assessment of the various fragmentation strategies applicable for large proteins. MIM routinely achieved an accuracy of <2 kcal/mol using MIM2 and MIM3 compared to the total energy of the full, unfragmented molecule at the DFT/6-311++G(d,p) level of theory. In this study, we used the MIM3 fragmentation method with the best-performing fragmentation scheme in our previous study to assess the consistency of MIM3 in predicting the interaction energies as well as the total energies.

The binding mode of ITK and the set of 15 benzothiazole ligands used for this assessment are shown in Figure 1 and Table 1, respectively. Residues and water molecules within 5.0 Å of the ligand were included in the QM calculation. The total number of atoms in the protein-ligand complex ranges from 400 to 500. Table 1 presents the total interaction energies calculated with MIM3 (B97-D3BJ/6-311++G(d,p):B97-D3BJ/6-31+G(d):PM6-D3) and their deviations from the full, unfragmented calculation at the B97-D3BJ/6-311++G-

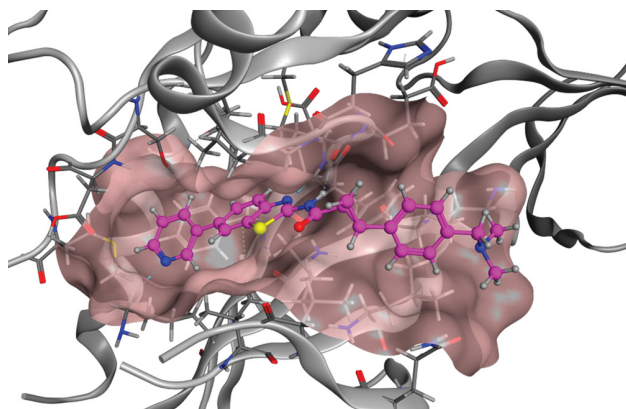


Figure 1. Active site of the interleukin-2-inducible T-cell kinase (ITK) receptor with a cocrystallized benzothiazole-based ligand (11). The active-site cavity is shown as a light-pink surface, and the ligand carbons are shown in dark pink.

(d,p) level of theory. The mean absolute error (MAE) in the MIM3 total interaction energy was calculated to be only 1.62 kcal/mol with a 0.84 kcal/mol population standard deviation over the test set, showing a good agreement with the full calculations. The lower value of the standard deviation stems from the observation that MIM3 seems to consistently overestimate the interaction energy (mean signed error (MSE) = -1.62 kcal/mol), likely because of a systematic mismatch between the high and low levels of theory in terms of the treatment of nonbonded interactions. The mean absolute deviation (MAD) after subtraction of the mean signed error is significantly reduced to 0.67 kcal/mol. Figure 2 shows a linear correlation of the MIM3 total interaction energies with the full calculations. A correlation coefficient (R^2) of 0.91 with respect to the $y = x$ line and an R^2 of 0.98 with respect to the line of best fit further assures that the MIM3 method fully reproduces the relative trend of protein–ligand interaction energies. This observation affirms that the MIM3 fragmentation method can be used to calculate the interaction energies of large protein–ligand complexes and provides a good alternative to the necessity of performing an expensive QM calculation on the full molecule. Thus, only MIM3 calculations were performed for all of the other data sets investigated in this study.

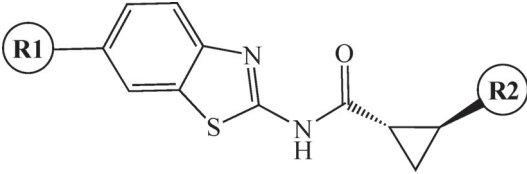
3.2. Prediction of Protein–Ligand Interaction Energies. Using the MIM3 protocol discussed above, we first calculated the gas-phase total interaction energies of 89 protein–ligand complexes at the B97-D3BJ/6-311++G-(d,p):B97-D3BJ/6-31+G(d):PM6-D3 combination of levels of theory. The MIM3-calculated interaction energies and the literature experimental binding potencies are given in Tables S2–S8. The calculated gas-phase interaction energies ($\Delta E_{\text{interaction}}$) range from -40 to -110 kcal/mol. As the calculations were performed in the gas phase without any correction for the solvation free energy or associated entropic penalties, the computed values were expected to be substantially larger than the normally observed binding energies in solution. However, the contributions from these solvation effects are understood to be systematic, and we expect to see some correlation between the measured binding potencies and the calculated interaction energies. The correlation plots comparing the experimental binding energies (ΔG_{bind}) or the negative logarithms of the binding affinities ($\text{p}K_i$) and the calculated interaction energies for the first six

data sets (set-I to set-VI) are shown in Figure 3. For the six data sets, we obtained a significant linear correlation ($R^2 = 0.79\text{--}0.94$, Spearman (ρ) = $0.87\text{--}0.94$) between the experimental binding affinities and the calculated interaction energies. This suggests that the calculated gas-phase interaction energies correctly reflect the trend of the experimentally observed binding potencies.

A closer analysis of the ligand structures with respect to the observed correlation reveals some interesting results. The highest correlation between the calculated interaction energies and the experimental binding affinities ($\text{p}K_i$) was obtained in the case of the indazole-based ITK inhibitors (set-II), with an impressive R^2 of 0.94. For this data set, the modified ligands closely resemble the cocrystallized ligand, **2g** (**2k** in the cases of **2k** and **2i**), used as the starting structure. This observation also holds true for set-VI, where modification of the ligand mainly involves substitutions on the phenyl ring of the cocrystallized ligand, **6f**. On the other hand, the lowest correlation ($R^2 = 0.79$, $\rho = 0.87$) was obtained in the case of the 3-azetidynyl set of inhibitors with P2Y12. In this case, the 4-piperidynyl group of the starting compound, **6f** (PDB ID 4NTJ), was replaced by the slightly smaller 3-azetidynyl ring and more drastic substitutions were made on the terminal phenyl group to obtain the desired set of ligands. These results indicate that the correlation between the calculated gas-phase and experimental binding affinities to some extent may be dependent on the similarities between the cocrystallized ligand and the modified ones in addition to the actual performance of the method, which was expected. Nevertheless, it is motivating to see that the correlation coefficients for the data sets obtained via modification of a cocrystallized ligand (set-I, -II, -IV, -V, and -VI; $R^2 = 0.79\text{--}0.94$, $\rho = 0.87\text{--}0.94$) are as good as or better than those for the set employing only published crystal structures (set-III; $R^2 = 0.84$, $\rho = 0.88$). This further validates that a carefully modeled protein–ligand complex can be a reliable prototype in the design of a potent drug molecule.

Set-VII, a set of SAP-based ligands complexed with the ITK receptor, is a particularly interesting case. With the computational protocol discussed above, the gas-phase interaction energies showed only a modest correlation ($R^2 = 0.49$) with the experimental binding potencies ($\text{p}K_i$'s) (Figure 4 and Table S8). The calculated correlation is significantly weaker than the correlation observed in the other six data sets. It is assumed that the observed moderate correlation of the calculated gas-phase interaction energies in our study is due to either the neglect of conformational changes in the ligand/protein structure, the lack of solvation effects, or a combination of these two factors. First, there could be a significant change in the protein and/or ligand conformation for some of the modified P–L complexes compared with the cocrystallized complex. This situation is more likely to be present when the sizes of the modified and cocrystallized ligands are significantly different, as noted earlier in the case of set-V. In such a case, the protein must rearrange to appropriately accommodate the ligand and maximize the interactions. The contribution from this type of protein conformational change is difficult to estimate without structural sampling. This is a well-known limitation of the “single-structure approach” used in this study.⁶ Nevertheless, a closer inspection of the interaction energy trend with respect to the experimental $\text{p}K_i$ reveals ligand **7q** to have a significantly lower interaction energy than the expected value (Figure 4 and Table S8). Consistent with this observation, for the same P–L complex, a significant

Table 1. Full and MIM3-Calculated Interaction Energies (in kcal/mol) for Various Benzothiazole ITK Inhibitors (Set-I)



Compound	R1	R2	$\Delta E_{\text{interaction}}$ (MIM3) ^a	$\Delta E_{\text{interaction}}$ (Full) ^b	Error ^c	Error after removing MSE ^d
1a	MeO----		-57.46	-57.29	-0.17	1.45
1b	MeO----		-59.38	-58.24	-1.14	0.48
1c			-64.36	-63.16	-1.21	0.42
1d	MeO----		-60.41	-58.63	-1.78	-0.16
1e			-68.11	-67.00	-1.11	0.51
1f			-68.22	-66.23	-1.99	-0.37
1g			-65.11	-63.39	-1.72	-0.10
1h			-72.62	-70.33	-2.29	-0.67
1i			-72.07	-70.65	-1.42	0.20
1j			-67.30	-66.68	-0.62	1.00
1k			-73.45	-71.20	-2.25	-0.63
1l			-71.27	-70.55	-0.72	0.90
1m			-75.02	-72.00	-3.03	-1.40
1n			-75.85	-72.53	-3.31	-1.70
1o			-74.02	-72.44	-1.59	0.04
			mean absolute error (MAE)	1.62	mean absolute	
			mean signed error (MSE)	-1.62	deviation	
			standard deviation (SD)	0.84	(MAD) 0.67	

^a $\Delta E_{\text{interaction}}$ calculated using MIM3 at the B97-D3BJ/6-311++G(d,p):B97-D3BJ/6-31+G(d):PM6-D3 combination of levels of theory. ^b $\Delta E_{\text{interaction}}$ of the full, unfragmented molecule calculated at the B97-D3BJ/6-311++G(d,p) level. ^cErrors in the MIM3 total interaction energies compared with the full calculation. ^dErrors in the MIM3 total interaction energies after removal of the average systematic error of -1.62 kcal/mol.

change in protein and ligand structure was observed by Ortwine and co-workers.⁵⁶ When **7q** was excluded from the data set as a significant outlier, a much better correlation coefficient was obtained ($R^2 = 0.68$).

The second possible reason for the observed modest correlation would be the neglect of the solvation energy contribution. For a set of ligands bound to the same protein active site, the desolvation energy of the protein during formation of the P-L complex can be expected to be similar. However, the desolvation energies of the ligands, depending on

their polarity and the fraction of ligand still exposed to the solvent in the P-L complex, can change significantly. In polar solvents like water, the ligand desolvation energy ($\Delta E_{\text{ligand}}^{\text{desolv}}$) for ligands that can strongly polarize the solvent and form hydrogen bonds will be larger than that for less-polar ligands. To assess the contribution of the ligand desolvation energy to the total interaction energy, we computed the solvation energies of the ligands in aqueous solution using implicit solvation, as described in section 2. This approach, however, assumes that the entire ligand is surrounded by the protein, i.e.,

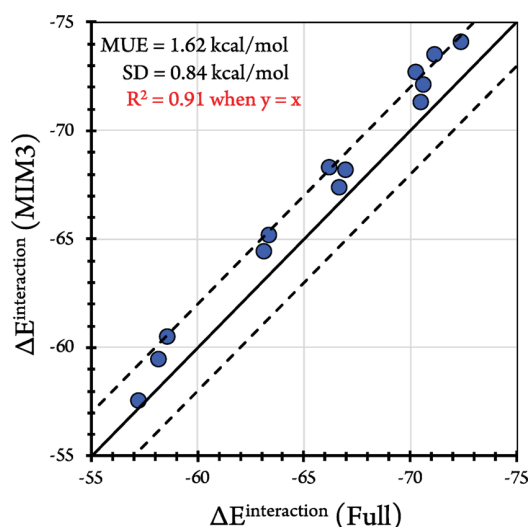


Figure 2. Correlation between MIM3-calculated interaction energies at the B97-D3BJ/6-311++G(d,p):B97-D3BJ/6-31+G(d):PM6-D3 combination of levels and the interaction energies of full, unfragmented molecules at the B97-D3BJ/6-311++G(d,p) level for benzothiazole-based ITK inhibitors (set-I) (in kcal/mol). Dashed lines represent an error range of ± 2.0 kcal/mol. R^2 with respect to the line of best fit is 0.98.

that it is *fully* desolvated in all directions. However, this is typically not the case, and some part of the ligand is still in contact with solvent. The effective ligand desolvation energy during the P–L complex formation was obtained by scaling the ligand solvation energy by the fraction of contact surface area with the receptor in P–L complex (eqs 4–6; vide supra). Interestingly, for set-VII the contribution of the ligand desolvation energy is quite significant, ranging from 12 to 21 kcal/mol (Table S8, last column). When the ligand desolvation energies are added to the total gas-phase interaction energies, R^2 dramatically improves from 0.49 to 0.75, and ρ increases from 0.63 to 0.86 (Figure 4 and Table 2). For the other test sets (set-I to set-VI), wherein the gas-phase interaction energies correlated significantly well with experiment ($R^2 = 0.79$ –0.94; $\rho = 0.87$ –0.94), addition of the ligand desolvation energy caused only a small change and maintained the good correlation with experiment ($R^2 = 0.74$ –0.93, $\rho = 0.83$ –0.93; Table 2 and Figure S2). The inclusion of the ligand desolvation energy in the total interaction energy remarkably improves the correlation coefficients of set-V (R^2 from 0.79 to 0.88, ρ from 0.87 to 0.94). For set-I and set-IV, a small lowering of the correlation coefficient (by ~ 0.05) is observed, while the others remain very close to the gas-phase results. As expected, the inclusion of ligand desolvation energy also significantly lowers the gas-phase interaction energy (by up to 22 kcal/mol; Tables S9–S15). Overall, the inclusion of the ligand desolvation energy using our simple scaling model significantly improves the correlation with the experimental binding potencies while decreasing the binding energy as expected.

The data sets used in this study overlap with those employed in other fragmentation-based theoretical studies. For the ITK inhibitors (set-I and set-II), Ortwin and co-workers calculated the gas-phase pairwise interaction energies using the FMO method at the MP2/6-31G* level of theory and obtained impressive correlations to the experimental binding potencies

($R^2 = 0.84$ and 0.91, respectively).⁵⁶ Söderhjelm et al.⁵⁸ computed the binding affinities of seven of the 14 investigated biotin-analogue avidin inhibitors (set-IV) using the polarizable multipole interaction with supermolecular pairs (PMISP/MM) approach at the MP2/cc-pVTZ level and found a modest correlation with the experimental binding potencies ($R^2 = 0.55$). When nonpolar solvation and entropy components were included, the correlation between the theoretical binding affinities and experiment worsened by a large degree ($R^2 = 0.27$). He and co-workers also computed the binding affinities of the set-IV complexes using their EE-GMFCC method combined with molecular dynamics (MD) simulations and the MM (GAFF parameters with Amber03 charges and PBSA solvation), HF-D3/6-31G* (with CPCM implicit solvation), and B3LYP-D3/6-31G* (with CPCM) levels of theory.⁵⁵ The He study obtained a modest correlation ($R^2 = 0.56$) for the HF-D3/6-31G*-calculated binding affinities using single snapshots and the MM/PBSA results using 300 snapshots taken from MD simulations. For the single-snapshot approach, a better correlation ($R^2 = 0.72$) was obtained with B3LYP-D3/6-31G*, which was further improved ($R^2 = 0.77$) when the binding affinities were averaged over three snapshots. Interestingly, the best results ($R^2 = 0.79$) were obtained when the contribution from CPCM solvation energy was removed from the total interaction energy.

For the 3-azetidinyl-based GPCR (P2Y12) inhibitors (set-VI), Heifetz et al.⁵⁴ calculated the pairwise interaction energies using FMO at the MP2/6-31G* level of theory and obtained a good correlation with the experimental binding potencies ($R^2 = 0.76$). For set-VII, Ortwin and co-workers calculated the gas-phase total pairwise interaction energies using FMO at the MP2/6-31G* level of theory and obtained an impressive R^2 of 0.87.⁵⁶ Comparing the correlation coefficients calculated for seven data sets in this study ($R^2 = 0.74$ –0.94, $\rho = 0.83$ –0.94) with those from other similar theoretical studies, it is clear that our MIM method performs as well as or better than previous methods discussed in the literature. Furthermore, since the high-layer subsystems in our study are only pentamers, the accuracy can be further improved by using more accurate methods, such as MP2 with sufficiently large basis sets or coupled-cluster theory, without making the high-level calculation a computational bottleneck. Additionally, the ligand desolvation energy is approximated by calculating the solvation energy of the ligand using implicit solvation and solvent-accessible surface area; this procedure requires negligible computational cost compared with the P–L complex.

4. CONCLUSIONS

In this study, we have presented a simple, efficient, and reliable protocol for the accurate prediction of interaction energies of large protein–ligand complexes using MIM. The performance of three-layer MIM (MIM3) in predicting the total interaction energies of 15 protein–ligand complexes at the B97-D3BJ/6-311++G(d,p) level of theory has been assessed. MIM3 energies were calculated using a prescribed fragmentation scheme (i.e., connectivity-based pentamers with distance-based dimers (N5D) in the high layer; distance-based (3.5 Å) in the middle layer) and combinations of high, medium, and low basis sets (high, 6-311++G(d,p); medium, 6-31+G(d); low, PM6-D3) from our recent work calculating the absolute energies of large proteins. The high-layer subsystem size was purposefully kept small to ensure that the calculations remain tractable even for more accurate correlated methods if needed.

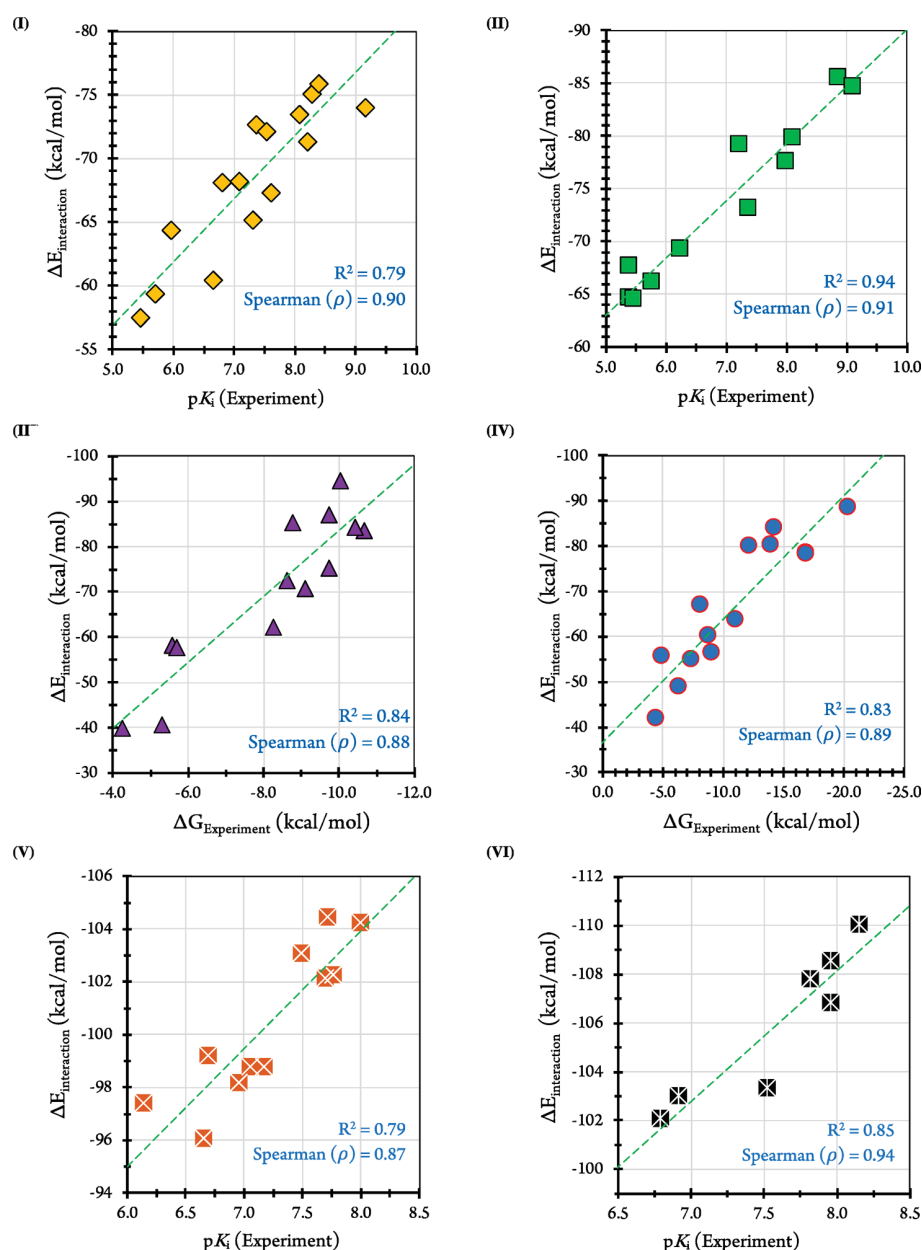


Figure 3. Correlations between experimentally measured binding affinities and the gas-phase interaction energies calculated using MIM3 (B97-D3BJ/6-311++G(d,p):B97-D3BJ/6-31+G(d):PM6-D3) for various protein–ligand complexes: (I) benzothiazole-based ITK inhibitors; (II) indazole-based ITK inhibitors; (III) CDK2 inhibitors; (IV) biotin-analogue avidin inhibitors; (V) 3-azetidiny-based P2Y12 inhibitors; (VI) 4-piperidinyl-based P2Y12 inhibitors. Dashed lines are lines of best fit.

Impressive accuracy (set-I, MAE = 1.62, MAD = 0.67, standard deviation = 0.84 kcal/mol) in MIM3 interaction energies compared to the full unfragmented calculations at the B97-D3BJ/6-311++G(d,p) level of theory was achieved, establishing the MIM3 method as a valid alternative for the calculation of accurate interaction energies.

The correlation between the gas-phase interaction energies and the experimentally measured binding potencies was determined by calculating the interaction energies of seven sets of protein–ligand complexes containing a total of 89 ligands complexed with four different proteins at the B97-D3BJ/6-311++G(d,p) level of theory using MIM3 as described above. For six data sets (set-I to set-VI), the results show a remarkable correlation of the gas-phase interaction energies to

the experimental protein–ligand binding potencies ($R^2 = 0.79$ – 0.94 , $\rho = 0.87$ – 0.94). However, a mediocre correlation was obtained for set-VII ($R^2 = 0.49$, $\rho = 0.63$), suggesting that the energy contributions from protein/ligand structure changes and solvation energy could play an important role. The contribution from ligand desolvation energy was included in the calculated interaction energies to assess the effect of solvation on the correlation between the calculated interaction energies with the experiment. The results show an impressive correlation of the calculated interaction energies to the experimental protein–ligand binding potencies ($R^2 = 0.74$ – 0.93 , $\rho = 0.83$ – 0.93).

While very good correlations have been obtained between the calculated interaction energies and experimental binding

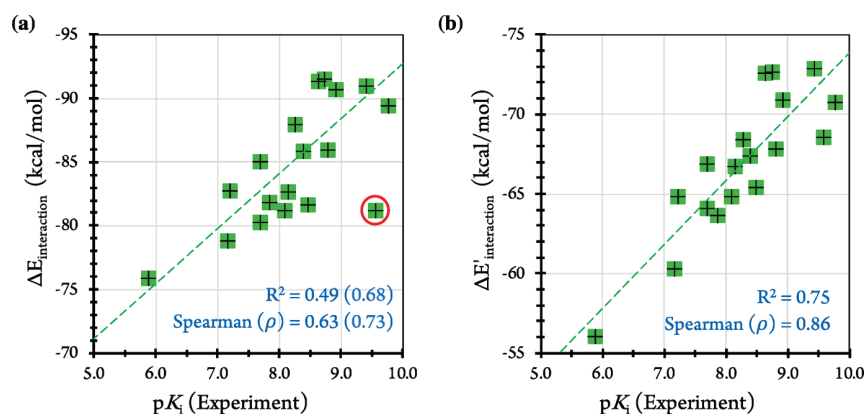


Figure 4. Correlation between experimentally measured binding affinities and calculated interaction energies at MIM3(B97-D3BJ/6-311++G(d,p):B97-D3BJ/6-31+G(d):PM6-D3) for sulphonylpyridine-based ITK inhibitors: (a) only gas phase interaction energies, and (b) gas phase interaction energy plus ligand desolvation energy. In (a) the R^2 and Spearman (ρ) correlation coefficient values in parentheses are calculated without the red-circled point in the plot, complex 7q.

Table 2. Correlation between MIM3-Calculated Interaction Energies and Experimental Binding Potencies (pK_i or ΔG_{bind})

data set	$\Delta E_{\text{interaction}}^a$		$\Delta E'_{\text{interaction}}^b$	
	R^2	ρ	R^2	ρ
set-I	0.79	0.90	0.74	0.83
set-II	0.94	0.91	0.93	0.92
set-III	0.84	0.88	0.85	0.93
set-IV	0.83	0.89	0.78	0.85
set-V	0.79	0.87	0.88	0.94
set-VI	0.85	0.94	0.83	0.90
set-VII	0.49	0.63	0.75	0.86

^aGas-phase total interaction energy calculated using MIM3 (B97-D3BJ/6-311++G(d,p):B97-D3BJ/6-31+G(d):PM6-D3). ^bGas-phase total interaction energy plus ligand desolvation energy calculated using eq 4.

affinities for the considered data sets, we point out that the current approach has some limitations that will be systematically addressed in the future. The first significant limitation is the use of the single-structure approach. For a related set of ligands with relatively small structural modifications, the contributions from protein conformational entropy are expected to cancel out, as observed in this study. However, it may be less likely for cases involving a significant change in the ligand structure and/or when the flexibility of the protein plays a significant role in binding. In such cases, accounting for the protein conformational change during binding may be critical. The second limitation of our current approach is the solvation energy calculation. While our results showed that approximating the differential solvation effects using the ligand desolvation energy in conjunction with the solvent-accessible surface area is a fairly accurate and cost-effective approach, it is necessary to include the solvation energy of the protein and the protein–ligand complex in the total interaction energy for charged residues or charged ligands. The inclusion of solvation effects will substantially decrease the overestimation in the ligand binding affinities in the gas-phase approach used in this paper, and we plan to explore this more carefully in future work on obtaining absolute binding free energies in solution.

In conclusion, the main purpose of this study was to present a computationally efficient protocol for estimating the relative

strength of P–L interactions for a set of structurally similar ligands, and our results appear highly promising in this regard. Overall, the impressive correlation between the calculated interaction energies and the experimental binding affinities suggests that QM-calculated interaction energies can provide a reliable estimate for the trend of binding potencies with substituted ligands. These results also suggest that the MIM protocol presented in this study offers a reliable, computationally affordable prediction of protein–ligand binding potencies that can be an effective tool for structure-based drug design in the future.

■ ASSOCIATED CONTENT

● Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00531.

Tables with protonation states of histidine residues, structures of ligands, experimental pK_i values and MIM3-calculated interaction energies; correlation plots between the experimentally measured binding affinities and MIM3-calculated (gas + ligand desolvation) interaction energies; and tables with solvation energies of ligands and SASAs of ligands, proteins, and protein–ligand complexes (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kraghava@indiana.edu.

ORCID

Bishnu Thapa: 0000-0003-3521-1062

Daniel Beckett: 0000-0003-4833-2269

Krishnan Raghavachari: 0000-0003-3275-1426

Funding

This work was supported by Eli Lilly and Company through the Lilly Research Award Program at Indiana University. The methods used in this project were developed through support from National Science Foundation Grant CHE-1665427 at Indiana University.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Big Red II supercomputing facility at Indiana University for the computer time.

■ REFERENCES

- (1) Hubbard, R. E. *Structure-Based Drug Discovery: An Overview*, 1st ed.; Royal Society of Chemistry: Cambridge, U.K., 2006; Vol. 3.
- (2) Blass, B. *Basic Principles of Drug Discovery and Development*; Elsevier: Philadelphia, PA, 2015.
- (3) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439–446.
- (4) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.
- (5) Gohlke, H.; Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.
- (6) Ryde, U.; Söderhjelm, P. Ligand-binding affinity estimates supported by quantum-mechanical methods. *Chem. Rev.* **2016**, *116*, 5520–5566.
- (7) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.
- (8) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
- (9) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
- (10) Beveridge, D. L.; DiCapua, F. Free energy via molecular simulation: Applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
- (11) Michel, J.; Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: Assumptions, pitfalls and expectations. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 639–658.
- (12) Chipot, C.; Rozanska, X.; Dixit, S. B. Can free energy calculations be fast and accurate at the same time? Binding of low-affinity, non-peptide inhibitors to the sh2 domain of the src protein. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 765–770.
- (13) Mobley, D. L.; Klimovich, P. V. Perspective: Alchemical free energy calculations for drug discovery. *J. Chem. Phys.* **2012**, *137*, 230901.
- (14) Steinbrecher, T.; Labahn, A. Towards accurate free energy calculations in ligand protein-binding studies. *Curr. Med. Chem.* **2010**, *17*, 767–785.
- (15) Hansson, T.; Marelus, J.; Åqvist, J. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27–35.
- (16) Sham, Y. Y.; Chu, Z. T.; Tao, H.; Warshel, A. Examining methods for calculations of binding free energies: Lra, lie, pdld-lra, and pdld/s-lra calculations of ligands binding to an HIV protease. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 393–407.
- (17) Foloppe, N.; Hubbard, R. Towards predictive ligand design with free-energy based computational methods? *Curr. Med. Chem.* **2006**, *13*, 3583–3608.
- (18) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., III. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- (19) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate–DNA helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.
- (20) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discovery* **2015**, *10*, 449–461.
- (21) Bissantz, C.; Kuhn, B.; Stahl, M. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.* **2010**, *53*, S061–S084.
- (22) Aleksandrov, A.; Thompson, D.; Simonson, T. Alchemical free energy simulations for biological complexes: Powerful but temperamental. *J. Mol. Recognit.* **2010**, *23*, 117–127.
- (23) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. The role of quantum mechanics in structure-based drug design. *Drug Discovery Today* **2007**, *12*, 725–731.
- (24) Cavalli, A.; Carloni, P.; Recanatini, M. Target-related applications of first principles quantum chemical methods in drug design. *Chem. Rev.* **2006**, *106*, 3497–3519.
- (25) Zhou, T.; Huang, D.; Cafisch, A. Quantum mechanical methods for drug design. *Curr. Top. Med. Chem.* **2010**, *10*, 33–45.
- (26) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Courier Corporation: North Chelmsford, MA, 2012.
- (27) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.
- (28) *Semiempirical Methods of Electronic Structure Calculation, Part B: Applications*; Segal, G., Ed.; Modern Theoretical Chemistry, Vol. 8; Plenum Press: New York, 1977.
- (29) Raghavachari, K.; Saha, A. Accurate composite and fragment-based quantum chemical models for large molecules. *Chem. Rev.* **2015**, *115*, 5643–5677.
- (30) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.* **2016**, *116*, 5301–5337.
- (31) Raha, K.; Merz, K. M. A quantum mechanics-based scoring function: Study of zinc ion-mediated ligand binding. *J. Am. Chem. Soc.* **2004**, *126*, 1020–1021.
- (32) Li, J.; Reynolds, C. H. A quantum mechanical approach to ligand binding—calculation of ligand–protein binding affinities for stromelysin-1 (MMP-3) inhibitors. *Can. J. Chem.* **2009**, *87*, 1480–1484.
- (33) Fanfrlík, J.; Bronowska, A. K.; Řezáč, J.; Přenosil, O.; Konvalinka, J.; Hobza, P. A reliable docking/scoring scheme based on the semiempirical quantum mechanical PM6-DH2 method accurately covering dispersion and h-bonding: HIV-1 protease with 22 ligands. *J. Phys. Chem. B* **2010**, *114*, 12666–12678.
- (34) Raha, K.; Merz, K. M. Large-scale validation of a quantum mechanics based scoring function: Predicting the binding affinity and the binding mode of a diverse set of protein–ligand complexes. *J. Med. Chem.* **2005**, *48*, 4558–4575.
- (35) Fanfrlík, J. i.; Brahmshatriya, P. S.; Řezáč, J.; Jílková, A.; Horn, M.; Mareš, M.; Hobza, P.; Lepšík, M. Quantum mechanics-based scoring rationalizes the irreversible inactivation of parasitic schistosoma mansoni cysteine peptidase by vinyl sulfone inhibitors. *J. Phys. Chem. B* **2013**, *117*, 14973–14982.
- (36) Ryde, U. QM/MM calculations on proteins. *Methods Enzymol.* **2016**, *577*, 119–158.
- (37) Mucs, D.; Bryce, R. A. The application of quantum mechanics in structure-based drug design. *Expert Opin. Drug Discovery* **2013**, *8*, 263–276.
- (38) Lin, H.; Truhlar, D. G. QM/MM: What have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2007**, *117*, 185.
- (39) Fanfrlík, J.; Kolář, M.; Kamler, M.; Hurný, D.; Ruiz, F. X.; Cousido-Siah, A.; Mitschler, A.; Řezáč, J.; Munusamy, E.; Lepšík, M.; Matějček, P.; Veselý, J.; Podjarný, A. D.; Hobza, P. Modulation of aldose reductase inhibition by halogen bond tuning. *ACS Chem. Biol.* **2013**, *8*, 2484–2492.
- (40) Brahmshatriya, P. S.; Dobes, P.; Fanfrlík, J.; Řezáč, J.; Paruch, K.; Bronowska, A.; Lepšík, M.; Hobza, P. Quantum mechanical scoring: Structural and energetic insights into cyclin-dependent kinase

2 inhibition by pyrazolo[1,5-*a*] pyrimidines. *Curr. Comput.-Aided Drug Des.* **2013**, *9*, 118–129.

(41) Rao, L.; Zhang, I. Y.; Guo, W.; Feng, L.; Meggers, E.; Xu, X. Nonfitting protein–ligand interaction scoring function based on first-principles theoretical chemistry methods: Development and application on kinase inhibitors. *J. Comput. Chem.* **2013**, *34*, 1636–1646.

(42) Fanfrlík, J.; Ruiz, F. X.; Kadlčíková, A.; Rezáč, J.; Cousido-Siah, A.; Mitschler, A.; Haldar, S.; Lepšík, M.; Kolář, M. H.; Majer, P.; Podjarný, A. D.; Hobza, P. The effect of halogen-to-hydrogen bond substitution on human aldose reductase inhibition. *ACS Chem. Biol.* **2015**, *10*, 1637–1642.

(43) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation methods: A route to accurate calculations on large systems. *Chem. Rev.* **2012**, *112*, 632–672.

(44) Fukuzawa, K.; Kitaura, K.; Uebayasi, M.; Nakata, K.; Kaminuma, T.; Nakano, T. Ab initio quantum mechanical study of the binding energies of human estrogen receptor α with its ligands: An application of fragment molecular orbital method. *J. Comput. Chem.* **2005**, *26*, 1–10.

(45) Ohno, K.; Wada, M.; Saito, S.; Inoue, Y.; Sakurai, M. Quantum chemical study on the affinity maturation of 48G7 antibody. *J. Mol. Struct.: THEOCHEM* **2005**, *722*, 203–211.

(46) Dedachi, K.; Hirakawa, T.; Fujita, S.; Khan, M. T. H.; Sylte, I.; Kurita, N. Specific interactions and binding free energies between thermolysin and dipeptides: Molecular simulations combined with ab initio molecular orbital and classical vibrational analysis. *J. Comput. Chem.* **2011**, *32*, 3047–3057.

(47) Bettens, R. P.; Lee, A. M. On the accurate reproduction of ab initio interaction energies between an enzyme and substrate. *Chem. Phys. Lett.* **2007**, *449*, 341–346.

(48) Mazanetz, M. P.; Ichihara, O.; Law, R. J.; Whittaker, M. Prediction of cyclin-dependent kinase 2 inhibitor potency using the fragment molecular orbital method. *J. Cheminf.* **2011**, *3*, 2.

(49) Antony, J.; Grimme, S.; Liakos, D. G.; Neese, F. Protein–ligand interaction energies with dispersion corrected density functional theory and high-level wave function based methods. *J. Phys. Chem. A* **2011**, *115*, 11210–11220.

(50) Hirakawa, T.; Fujita, S.; Ohshima, T.; Dedachi, K.; Khan, M. T. H.; Sylte, I.; Kurita, N. Specific interactions and binding energies between thermolysin and potent inhibitors: Molecular simulations based on ab initio molecular orbital method. *J. Mol. Graphics Modell.* **2012**, *33*, 1–11.

(51) Nowosielski, M.; Hoffmann, M.; Kuron, A.; Korycka-Machala, M.; Dziadek, J. The MM2QM tool for combining docking, molecular dynamics, molecular mechanics, and quantum mechanics. *J. Comput. Chem.* **2013**, *34*, 750–756.

(52) Andrejić, M.; Ryde, U.; Mata, R. A.; Söderhjelm, P. Coupled-cluster interaction energies for 200-atom host–guest systems. *ChemPhysChem* **2014**, *15*, 3270–3281.

(53) Poongavanam, V.; Steinmann, C.; Kongsted, J. Inhibitor ranking through QM based chelation calculations for virtual screening of HIV-1 RNase H inhibition. *PLoS One* **2014**, *9*, e98659.

(54) Heifetz, A.; Chudyk, E. I.; Gleave, L.; Aldeghi, M.; Cherezov, V.; Fedorov, D. G.; Biggin, P. C.; Bodkin, M. J. The fragment molecular orbital method reveals new insight into the chemical nature of GPCR–ligand interactions. *J. Chem. Inf. Model.* **2016**, *56*, 159–172.

(55) Liu, J.; Wang, X.; Zhang, J. Z.; He, X. Calculation of protein–ligand binding affinities based on a fragment quantum mechanical method. *RSC Adv.* **2015**, *5*, 107020–107030.

(56) 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.; Ortwin, D. F. Fragment molecular orbital method applied to lead optimization of novel interleukin-2 inducible T-cell kinase (ITK) inhibitors. *J. Med. Chem.* **2016**, *59*, 4352–4363.

(57) Söderhjelm, P. r.; Ryde, U. How accurate can a force field become? A polarizable multipole model combined with fragment-wise quantum-mechanical calculations. *J. Phys. Chem. A* **2009**, *113*, 617–627.

(58) Söderhjelm, P.; Kongsted, J.; Ryde, U. Ligand affinities estimated by quantum chemical calculations. *J. Chem. Theory Comput.* **2010**, *6*, 1726–1737.

(59) Koyama, Y.; Ueno-Noto, K.; Takano, K. Affinity of HIV-1 antibody 2G12 with monosaccharides: A theoretical study based on explicit and implicit water models. *Comput. Biol. Chem.* **2014**, *49*, 36–44.

(60) He, X.; Zhu, T.; Wang, X.; Liu, J.; Zhang, J. Z. H. Fragment quantum mechanical calculation of proteins and its applications. *Acc. Chem. Res.* **2014**, *47*, 2748–2757.

(61) Antony, J.; Grimme, S. Fully ab initio protein–ligand interaction energies with dispersion corrected density functional theory. *J. Comput. Chem.* **2012**, *33*, 1730–1739.

(62) Fedorov, D. G.; Kitaura, K. Pair interaction energy decomposition analysis. *J. Comput. Chem.* **2007**, *28*, 222–237.

(63) Söderhjelm, P.; Öhrn, A.; Ryde, U.; Karlström, G. Accuracy of typical approximations in classical models of intermolecular polarization. *J. Chem. Phys.* **2008**, *128*, 014102.

(64) Phipps, M. J.; Fox, T.; Tautermann, C. S.; Skylaris, C.-K. Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. *Chem. Soc. Rev.* **2015**, *44*, 3177–3211.

(65) Thapa, B.; Beckett, D.; Jovan Jose, K.; Raghavachari, K. Assessment of fragmentation strategies for large proteins using the multilayer molecules-in-molecules approach. *J. Chem. Theory Comput.* **2018**, *14*, 1383–1394.

(66) *Molecular Operating Environment (MOE)*; Chemical Computing Group: Montreal, QC, 2018.

(67) Labute, P. LowModeMD implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. *J. Chem. Inf. Model.* **2010**, *50*, 792–800.

(68) Labute, P. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins: Struct., Funct., Genet.* **2009**, *75*, 187–205.

(69) Gerber, P. R.; Müller, K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 251–268.

(70) Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. Ffl4ipq: A self-consistent force field for condensed-phase simulations of proteins. *J. Chem. Theory Comput.* **2014**, *10*, 4515–4534.

(71) 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.

(72) 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*, 2012–2023.

(73) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(74) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange–correlation functionals. *J. Chem. Phys.* **1997**, *107*, 8554–8560.

(75) Schmider, H. L.; Becke, A. D. Optimized density functionals from the extended G2 test set. *J. Chem. Phys.* **1998**, *108*, 9624–9631.

(76) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(77) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(78) Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54*, 724–728.

(79) Hariharan, P. C.; Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoret. Chim. Acta* **1973**, *28*, 213–222.

(80) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972**, *56*, 2257–2261.

(81) Francel, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements. *J. Chem. Phys.* **1982**, *77*, 3654–3665.

(82) Clark, T.; Chandrasekhar, J.; Spitznagel, G. W.; Schleyer, P. V. R. Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li-F. *J. Comput. Chem.* **1983**, *4*, 294–301.

(83) Stewart, J. J. Optimization of parameters for semiempirical methods V: Modification of nndo approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(84) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(85) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, revision A.03; Gaussian, Inc.: Wallingford, CT, 2016.