UNIVERSITY OF CALIFORNIA, SAN DIEGO

Noncovalent Interactions: Evaluation of Computational Methods and Characterization of Molecular Binding

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioengineering with a Specialization in Multi-Scale Biology

by

Amanda Li

Committee in charge:

Professor Michael K. Gilson, Chair
Professor Xiaohua Huang, Co-Chair
Professor J. Andrew McCammon
Professor Andrew McCulloch
Professor Ross Walker

2016

The Dissertation of Amanda Li is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in the Journal of Chemical Theory and Computation 2014, Li, Amanda; Muddana, Hari S.; Gilson, Michael K. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in The Journal of Physical Chemistry B 2016, Li, Amanda; Voronin, Alexey; Fenley, Andrew; Gilson, Michael K. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material, Reany, Ofer; Li, Amanda; Yefet, Maayan; Gilson, Michael K.; Keinan, Ehud. The dissertation author was the secondary investigator and author of this paper.

VITA

2007	Bachelor of Science, Columbia University, New York, NY
2007	Research Associate, Intelligent Bio-Systems, Waltham, MA
2008	Research Associate II, Regeneron Pharmaceuticals, Tarrytown, NY
2016	Doctor of Philosophy, University of California, San Diego, CA

ABSTRACT OF THE DISSERTATION

Noncovalent Interactions: Evaluation of Computational Methods and Characterization of Molecular Binding

by

Amanda Li

Doctor of Philosophy in Bioengineering with a Specialization in Multi-Scale Biology

University of California, San Diego, 2016

Professor Michael K. Gilson, Chair
Professor Xiaohua Huang, Co-Chair

Noncovalent interactions are of central importance to biochemical phenomena. This dissertation includes both evaluations of the methods used to compute noncovalent interactions and analyses of their role in binding. First, various QM approaches for calculating noncovalent interaction energies are compared in over 1,200 gas-phase dimers. In particular, we study semiempirical PMx methods, density functional theory (DFT) approaches, and symmetry-adapted perturbation theory (SAPT). Linearly scaled SAPT0 (fSAPT0) methods are fitted and shown to yield high accuracy, at particularly low computational cost.

Additionally, various models of polarization are examined for their ability to reproduce perturbed electrostatic potentials (ESPs). Polarization models are broken down into two main components: the *representation* of electronic polarization, and the *response model* used to map from an inducing field to the polarization within the chosen representation. The results reveal that the inducible dipole models used in many current polarizable force fields fall far short of the optimal results in principle achievable by the atom-centered point dipole representation. Lastly, binding interactions are examined between heteroallene-containing guests and cucurbituril host systems using quantum calculations and in Grb2 SH2 complexes using molecular dynamics simulations. For the host-guest systems, the heteroallenes are shown to exhibit attractive interactions with the carbonyl oxygens of the host, and these interactions are found to be primarily electrostatic and dispersive in nature. For the Grb2 SH2 domain, the thermodynamics of ligand preorganization are studied by computing relative binding enthalpies for flexible and constrained ligands.

Chapter 1

Introduction

Noncovalent interactions occur when atoms or molecules interact with each other without forming bonds or sharing electrons. They are ubiquitous in nature and integral driving forces in biology, mediating biomolecular structure as well as molecular recognition. The list of biochemical phenomena that involve noncovalent interactions is numerous, including protein folding, DNA structure, drug binding and metabolic processes. Thus, accurate theoretical and computational modeling of noncovalent interactions has broad consequences and applications, from providing insight into how biological assemblages form and how molecular motors function, to drug design and protein engineering.

Compared to covalent interactions, which result from electron sharing through a covalent bond, noncovalent interactions tend to be weaker, having energies up to a few kilocalories per mole, but occur across longer interatomic separations, typically greater than 2 Å[20]. While the following is not a comprehensive description of all noncovalent interactions, some of the most common types are described here.

Many noncovalent interactions may described in terms of multipole moments[120]. Permanent electrostatic interactions involve permanent charge distributions, and may be

either repulsive or attractive. Induction is the attraction that occurs when a charge distribution polarizes, or induces, another charge distribution. Dispersion is also attractive and involves correlated fluctuations of atomic electron clouds. An additional short-range interaction is Pauli repulsion, or exchange repulsion, which is distinct in that it does not result from electrostatic interactions between their charge distributions, but from the Pauli exclusion principles, which leads to short-ranged steric repulsion. Collectively, the Pauli repulsion and dispersion are known as van der Waals (VDW) forces. These interactions are most relevant to neutral and nonpolar charge distributions, as stronger electrostatic interactions are more dominant in charged and polar systems[86].

Hydrogen bonds originate from electrostatic attractions and are categorized separately from VDW interactions. A hydrogen bond is formed between a polar hydrogen, covalently bonded to an electronegative atom, acting as a proton donor, and a nearby electronegative atom acting as the proton acceptor[105]. Hydrogen bonds are considered to have covalent character, and their interaction distances can be less than the sum of nominal VWD contact distances. Ionic interactions are attractive interactions that involve charged ions. Ionic bonding arises from the transfer of electrons between oppositely charged ions. Additionally, ions may also interact with polar groups of neutral charge.

Some noncovalent interactions stem not merely from electronic structure, but also from molecular motions. These can vary with temperature and have an entropic component, as with the hydrophobic effect, which describes the association of nonpolar groups due to low solubility in polar solvents. Hydrophobic bonds are those that form between the aggregated nonpolar groups[105].

In computational chemistry, noncovalent interactions are typically represented by quantum mechanical (QM), molecular mechanical (MM) models, or a combination of

both. Since most noncovalent interactions trace to interactions between electrons, the most fundamental methods are those that treat electronic structure explicitly, as in QM. On the other hand, MM models of noncovalent interactions are more approximate and, rather than treat electrons explicitly, typically rely on atom-centered parameters in a potential energy function, also known as a force field. The primary advantage of MM over QM is the computational speed it affords, but the aforementioned interaction components (e.g., electrostatics, induction, dispersion and repulsion) are approximated by terms that cannot fully reproduce their complexity.

The strength of noncovalent interactions is typically measured by an interaction, or stabilization, energy. In QM, there are varied methods for estimating this quantity. The supermolecular (or variation) approach[68, 21] computes the interaction energy as the energy of a system of molecules less the individual energies of the subsystems. In the perturbation approach, the interaction energy is calculated directly by treating interactions as perturbations.

Essential to all QM calculations is the Schrødinger equation, which describes the particle-wave behavior of electrons:

$$H\Psi = E\Psi \tag{1.1}$$

where H is the Hamiltonian operator, Ψ (an eigenfunction) is the wavefunction of a system, describing its electronic motions, and E (an eigenvalue) is the energy of the system. For systems containing more than two particles, such as most atoms or molecules, there are no analytical solutions to Schrødinger equations. Thus, all electronic structure methods discussed here serve to estimate the wavefunction and energy, and are distinguished by the approximations employed. An overview of key methods will be shared here, but the

technical details are more thoroughly discussed in literature[99, 134, 37].

The Hartree-Fock (HF) solution to 1.1 is approximated using one and two-electron operators and wave functions, or orbitals, and assumes that each electron only interacts with an average of the effects, or a mean field, of other electrons. In the interest of computational speed and facilitating calculations on larger systems, the HF solution may be further approximated by semiempirical methods, which only consider the valence electrons of a system and typically employ parameters fitted to experimental data[99]. The HF solution may also be enhanced by accounting for electron correlation, which is not accounted for when the electronic effects are averaged. Such "post-HF" methods include coupled cluster (CC)[32] and Møller-Plesset (MP)[115] perturbation theory. Another class of methods that account for electron correlation are density functional theory (DFT) methods, which estimate the total electronic energy by the electron density[99]. Various DFT methods are differentiated by their treatment of exchange, correlation and dispersion. Symmetry-adapted perturbation theory (SAPT)[79] is a unique method that treats the interaction energy as a perturbation series and accounts for the energies resulting from electrostatic, induction, dispersion and repulsion separately.

The approximations used in the QM methods for calculating interaction energies are varied, and have differing effects on computational speed and accuracy. Thus, it is important to determine the relative accuracy of QM methods, and several are assessed by comparison to high-level reference energies in Chapter 3. Furthermore, QM calculations are useful for determining the driving forces of molecular phenomena such as binding, and Chapter 4 details QM studies to characterize an experimentally observed noncovalent interaction in a host-guest system.

Molecular mechanical models are especially useful for computing thermodynamic

quantities for complex systems, as such calculations require computing the energies of many configurations of the system, and QM methods are typically too slow to be tractable in such applications. In a typical MM force field, a separate set of parameters is used to model a specific category of noncovalent interactions, such as point charges for permanent electrostatics or Lennard-Jones (L-J) parameters for van der Waals' interactions. The parameters are typically derived from experimental data, QM calculations, or both, and Chapter 2 advises specific methods (such as MO62X[195] and linearly-scaled SAPT0) for accurate and fast calculations for parameterization purposes.

A problem for traditional force fields is the accurate representation of the polarization response in a molecule, or specifically the response of a charge distribution to changing external electric fields. In recent years, polarizable force fields have been developed to explicitly account for polarization using varied representations, including fluctuating charges[135, 149], Drude oscillators[43, 98] (as in CHARMM[9]), inducible dipoles[38] (as in AMBER ff02[30]) and atomic multipoles (as in AMOEBA[156]). Both fluctuating charges and Drude oscillators model polarization using point charges: fluctuating charges are atom-centered point charges whose magnitudes are changed to equalized electronegativities[116] while Drude oscillators include a massless charge that moves about an atom center in addition to each atom-centered point charge. In comparison, inducible dipole models use atom-centered point dipoles with polarization responses calculated by atom-centered polarizabilities, and point multipoles models include higher-order multipole moments in addition to point charges and induced point dipoles. Chapter 3 determines the maximal accuracy achievable by the atom-centered point charge and point dipole representations for modeling polarization, and compares these with the accuracy of different response models for inducibles dipoles.

A key application of force fields is molecular simulations of time-varying biomolecular processes, such as protein folding and binding, which often cannot be practically modeled by QM calculations (although we note that there exist hybrid QM/MM methods[153, 23]). The potential energy function of a MM force field can be used to generate simulations of molecular dynamics (MD), which are the time-varying motions of atoms and molecules. Such simulations may be used to compute interesting thermodynamic quantities, such as binding enthalpies and free energies, which are relevant to rational drug design and virtual screening. Progress in developing simulation methods for modeling the driving forces of binding has tended to focus more on calculations of binding free energies[27]. However, in recent work [47, 66], it has been shown that host-guest binding enthalpies can be computed to good precision by a simple direct method, using just the mean potential energies of simulations. Although proteins are significantly more complex, recent advances in simulation technologies, such as programs that can utilize multiple GPUs in parallel, suggest that proteins may be tractable by the same approach. In Chapter 5, long-trajectory MD simulations are used to estimate protein-ligand relative binding enthalpies using the direct approach.

This dissertation first considers approaches for examining noncovalent interactions, and then details selected applications of QM and MM methods. Quantum mechanical methods for computing interaction energies are surveyed in Chapter 2, while representations and response models commonly found in polarizable force fields are assessed in Chapter 3. In Chapter 4, QM studies are used to characterize attractive interactions between heteroallene-containing guests and a cucurbitiril host, and in Chapter 5, relative binding enthalpies are estimated using MD simulations for flexible and constrained ligands of the Grb2 SH2 domain.

Chapter 2

Quantum Mechanical Calculation of Noncovalent Interactions: A Large-Scale Evaluation of PMx, DFT and SAPT Approaches

2.1 Introduction

Noncovalent interactions are of fundamental importance to biomolecular systems, as they help determine the structures and functions of protein and nucleic acids, and play a central role in molecular recognition. A reliable representation of noncovalent interactions therefore is critically important to computational modeling of biomolecules, with applications that include rational drug design and protein engineering[53, 184]. In molecular simulations, noncovalent interactions are typically modeled by the nonbonded terms in an empirical force field[35, 63, 82, 102, 104, 106, 132]. These account for electrostatic

and van der Waals interactions, and may also include terms to account for time-varying changes in electronic polarization during the simulation[156]. Although the parameters in an empirical force field are typically adjusted to optimize agreement with experimental data, growing computer power and a shortage of suitable experimental data are also driving increased use of quantum mechanical (QM) calculations to parameterize and test force fields[75, 110, 128, 155, 191]. In addition, concerns regarding the accuracy of empirical force fields[118, 119] are motivating the direct application of QM methods to the study of noncovalent binding in host-guest[60, 117] and protein-ligand[46, 76, 133] systems.

It would be ideal if such applications could take advantage of the highly accurate QM approach often viewed as the gold standard for computing noncovalent interactions; i.e., counterpoise-corrected couple-cluster theory, with single, double and perturbative triple excitations, extrapolated to the completed basis set limit[83]. However, the computational demands of such CCSD(T)/CBS CP calculations make them too time-consuming for routine use in force field parameterization and prohibit direct application to biomolecular systems. As a consequence, a range of other QM methods have been developed. Because all of these methods make approximations for the sake of computational efficiency, it becomes essential to evaluate their accuracy. While there are many studies which rely on high accuracy reference results for relevant molecular systems[24, 41, 70, 74, 94], there is still need for broader, comparative validation studies, which will provide users and developers with a perspective of the strengths, weaknesses and tradeoffs among the various QM approaches, and their applicability to specific classes of noncovalent interactions.

Here, therefore, we contribute a systematic assessment of accuracy and speed for a range of QM methods, using a reference dataset of over 1,200 gas-phase dimers, for which CCSD(T)/CBS CP reference energies are publicly available in the Benchmark En-

ergy and Geometry Database (BEGDB)[145]. The categories of QM method examined are: semiempirical; density functional theory (DFT), with and without dispersion corrections; and symmetry-adapted perturbation theory (SAPT). The semiempirical PM6[164] and PM7[166] methods, the most computationally efficient methods tested here, rely on empirically adjustable parameters and are often combined with additional interaction terms. We examine the PM6 method, with post hoc corrections for dispersion and hydrogen bonding interactions (PM6-DH2[95, 140], PM6-DH+[95], and halogen interactions (PM6-DH2X[141]). The PM7 method, which is based on PM6, is also included without additional corrections, as its parameterization strategy accounts for such interactions. We test the widely used DFT functionals B3LYP[15, 100], B97-D[59] and M062X[195], with and without added dispersion corrections[61], as well as the ωB97X-D[26] functional, which includes its own correction for dispersion, is also tested. Finally, we test SAPT[79], which is distinct from the PMx and DFT approaches in that it is applicable only to the calculation of noncovalent interactions (e.g., it cannot be applied to geometry optimizations), and that it provides an informative decomposition of the overall interaction energy into electrostatic, induction, exchange and dispersion components. In SAPT, the interaction energy is computed as an expansion of perturbative terms, and we examine the SAPT0, SAPT2, SAPT2+, SAPT2+(3) and SAPT2+3 truncations[70]. We also explore the potential for the fast SAPT0 (fSAPT0) method to afford accurate results through empirical scaling of its energy terms, much as done previously in a smaller study[142], and make available the detailed energy decompositions afforded by SAPT across all of the test systems, as these can be useful to guide force field parameterization[111]. The present study provides a unique perspective of the reliability and efficiency of a broad range of QM methods, and should be a useful guide to their selection and further improvement.

2.2 Methods

2.2.1 Benchmark Datasets

A growing collection of benchmark datasets provides high quality geometries and interaction energies for noncovalent complexes[70]. Here, we use several datasets (Table 2.1) from the BEGDB to explore the accuracy of various QM methods for noncovalent interactions spanning a range of system types and sizes. We study a total of 1,266 dimers, ranging in size from 20 electrons in 4 atoms to 478 electrons in 101 atoms (Figure 2.1). These various BEGDB datasets probe different classes of noncovalent interactions. In particular, the S22x5[57] and S66x8[146] datasets both contain noncovalent complexes categorized as hydrogen bonded (electrostatics dominated), dispersion dominated, or mixed electrostatic and dispersive; X40x10[147] focuses on complexes with halogen interactions; Ionic[143] contains systems with charged hydrogen bonds; SCAI[17] and JSCH[83] contain amino acid and nucleic acid complexes; and the L7[152] dataset contains even larger extended complexes, all containing greater than 200 electrons. Several of these datasets, S22x5, S66x8[146], X40x10[147] and Ionic, contain geometries generated along a dissociation path relative to the equilibrium geometry, and thus include many nonequilibrium conformations. Because the aug-cc-pVTZ basis set[44, 87, 187, 189] used in the present study is not applicable to iodine, we omit the nine iodine-containing complexes in X40x10, and term the reduced dataset X31x10. Lastly, we also include the A24[144] dataset, which contains small complexes sized to enable comparison of more accurate approaches that would otherwise be unfeasible for larger complexes. BEGDB provides counterpoise-corrected CCSD(T)/CBS interaction energies for all of these datasets, except for JSCH, which contains energies evaluated using CCSD(T)/CBS and MP2/CBS without counterpoise (CP) correction, and

Table 2.1: BEGDB datasets used in the present study. Note that the names match those on the BEGDB website, which are not necessarily consistent with the corresponding publications. For example, the X40, X40x10 and Ionic datasets have also been referred to as 'Halogens,' 'Halogensx10,' and 'Charged HB.'[74]

Equilibrium Datasets

Dataset	Description	No. of Structures	Geometry Optimization Method	Reference Energy Method
A24[144]	Small complexes of 7-11 atoms	24	CCSD(T)/CBS CP or noCP	CCSD(T)/CBS CP
S22[83]	Small complexes of 8-26 atoms	22	MP2/cc-pVTZ CP or CCSD(T)/cc-pV(T/Q)Z noCP	CCSD(T)/CBS CP
S66[146]	Small complexes of 6-18 atoms	66	MP2/cc-pVTZ CP	CCSD(T)/CBS CP
X40[147]	Complexes with halogenated molecules	40	MP2/cc-pVTZ CP	CCSD(T)/CBS CP
SCAI[17]	Amino acid side chain complexes of 22-32 atoms	24	DFT TPSS/TZVP noCP	CCSD(T)/CBS CP(D->T)
JSCH[83]	124 nucleic base complexes and 19 amino acid complexes of 29-41 atoms	143	Artificial geometries, NMR structures, crystal structures, X-ray structures, MP2/cc-pVTZ noCP or MP2/TZVPP noCP	CCSD(T)/CBS noCP or MP2/CBS noCP
L7[152]	Large complexes of 48-112 atoms	7	DFT-D TPSS-D/TZVP or other	QCISD(T)/CBS CP or CCSD(T)/CBS CP

Nonequilibrium Datasets

Dataset	Relative Displacements	No. of Structures	Geometry Optimization Method	Reference Energy Method
S22x5[57]	0.9, 1.0, 1.2, 1.5, 2.0	110	MP2/cc-pVTZ CP or CCSD(T)/cc-pV(T/Q)Z noCP	CCSD(T)/CBS CP
S66x8[146]	0.90, 0.95, 1.00, 1.05, 1.10, 1.25, 1.50, 2.00	528	MP2/cc-pVTZ CP	CCSD(T)/CBS CP
X40x10[147]	0.80, 0.85, 0.90, 0.95, 1.00, 1.05, 1.10, 1.25, 1.50, 2.00	400	MP2/cc-pVTZ CP	CCSD(T)/CBS CP
Ionic[143]	0.90, 0.95, 1.00, 1.05, 1.10, 1.25, 1.50, 2.00	120	MP2/cc-pVTZ CP	CCSD(T)/CBS CP

L7, which uses QCISD(T)/CBS CP. It is also worth noting that there are variations within and across these BEGDB datasets in both the basis sets and extrapolation schemes employed to obtain the CBS results. Such details are not trivial and can produce discrepancies as large

as 0.7 kcal/mol, as elaborated in the Results section.

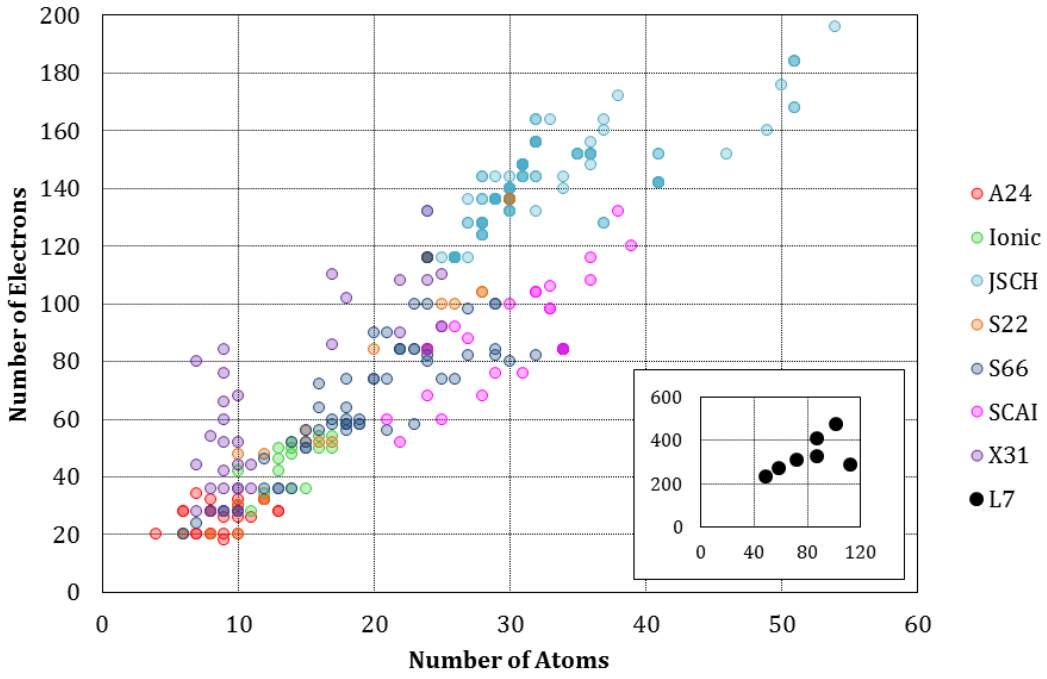


Figure 2.1: Sizes of dimers studied. For nonequilibrium datasets, only one point is shown per dimer system. The larger L7 dataset is included in the inset.

2.2.2 Computational Methods

Semiempirical PMx energy calculations were carried out with MOPAC2012[165]. The PM6[164] methods were examined with corrections for dispersion, hydrogen bonding and halogen interactions. PM6-DH2[95] and PM6-DH+[93] differ in the hydrogen bonding correction used, with the latter having improved long-range behavior. PM6-DH2X[141] adds an empirical repulsive correction for halogen interactions to the same dispersion and hydrogen-bonding corrections implemented in PM6-DH2. The more recent PM7[166] method is parameterized against a larger reference dataset than that used for PM6, and includes its own terms to account for dispersion and hydrogen bonding.

The DFT calculations with CP correction were carried out with the aug-cc-pVTZ

basis set, in revision C.01 of Gaussian 09[51]. Where SCF calculations failed to converge using default run parameters, the keyword `Integral=(Acc2E=12)` was used to increase the two-electron integral accuracy. B3LYP[15, 100], B97-D[59], M062X[195], and ωB97X-D[26] functionals were selected based on previously assessed performance for noncovalent interactions[41]. The D3 dispersion correction was applied using DFT-D3, version 3.0[61] to B3LYP, B97-D and M062X using the default parameters. These were optimized using a different basis set, (aug-)def2-QZVP. However, we have found that, for the S22 dataset, aug-cc-pVTZ and (aug-)def2-QZVP, without any dispersion correction, give results that are within 0.18 kcal/mol RMSE of each other, across all of the methods examined in the present study. We also observed that using the DFT-D3 parameters optimized for (aug-)def2-QZVP reduced the RMSE across all dataset by up to 0.12 kcal/mol, compared with using those optimized for (aug-)def2-TZVPP, the only other basis set option currently available. Becke-Johnson damping for the D3 correction (D3BJ)[62] was also tested for B3LYP and B97-D using parameters optimized for (aug-)def2-QZVP; there is no such correction available for M062X. The original 'zero-damping' D3 corrections are so-called because they employ a damping function for which the dispersion energy approaches zero with small internuclear separations. We note that, with the exception of B3LYP, all the functionals already contain some treatment of dispersion: B97-D is the B97 functional with the D2 dispersion correction; M062X is already parameterized to account for dispersion; and ωB97X-D utilizes its own specialized empirical dispersion correction.

The SAPT[79] energy calculations at varying orders (SAPT0, SAPT2, SAPT2+, SAPT2+(3), SAPT2+3) were carried out with PSI4[174]. In SAPT, the total interaction is computed as a sum of energy terms which are each classified as resulting from electrostatic, exchange, induction, or dispersion effects. The specific truncations of the SAPT expansion

are detailed in Table A.1 (Appendix A). Due to memory limitations, only lower order SAPT calculations were completed for larger systems. Thus, L7 was evaluated only with SAPT0; SCAI was evaluated at orders through SAPT2+; JSCH was evaluated through SAPT2 with the exception of 9 amino acid pair geometries (F30-K46, F30-L33, F30-Y13, F30-F49, F30-Y4, F49-K46, F49-V5, F49-W37 and F49-Y4) taken from a rubredoxin crystal structure[179] (PDB 1RB9), for which only SAPT0 calculations were completed. On the opposite end of the system size spectrum, SAPT orders up to SAPT2+3 were calculated for A24. All other datasets (S22x5, S66x8, Ionic and X31x10) were evaluated through SAPT2+(3).

2.2.3 Evaluation of Accuracy and Computational Speed

We use the root mean squared error (RMSE) as the primary metric of error in comparing the various computational methods. However, the mean signed error (MSE) is also provided to further characterize the performance of each method–a negative MSE indicates that a method overestimates the attractive interactions of a noncovalent dimer. Relative error is often reported in the literature, presumably because errors are thought to increase with interaction energy. Here, however, we saw no correlation between error and interaction ($R^2 < 0.2$ for all methods), so relative errors are not reported.

Timing studies were carried out on 8 CPUs of a 16-CPU node which was dedicated entirely to the calculation being timed. The timings of DFT and SAPT methods were examined by applying them in triplicate to each system in the A24 dataset and noting the shortest of the three wall-clock times reported as elapsed "real" time by the Unix time command. This timing approach accounts for the efficiency with which each method uses the 8 available CPUs. The timings DFT with D3 dispersion correction are recorded without

the add-on correction, as it requires negligible resources compared to the main calculation.

2.2.4 Linear Scaling of SAPT0 Energy Terms

The SAPT0 interaction energy is the sum of 6 energy terms, as detailed in Table A.1 of Appendix A, and in the fSAPT0 schemes, a separate scaling factor is applied to some or all of these terms. The linear scaling factors for the SAPT0 energy terms were determined by randomly splitting the systems in all combined datasets, except L7, into two equal subsets. One subset was used for training, and the other for testing. Thus, we applied multiple linear regression of the SAP0 energy terms to the corresponding CCSD(T)/CBS CP reference energies for the training set, to obtain a set of fitted coefficients; used these coefficients to compute interaction energies for the test set; and computed correlation coefficients and RMSE for the test-set results. This procedure was repeated 1,000 times, with different random selections of the training and testing subsets. Three different fitting schemes were tested: fSAPT0(1) scales all SAPT0 energy terms; fSAPT0(2) scales only the two dispersion terms, $E_{disp}^{(20)}$ and $E_{exch-disp}^{(20)}$, treated independently; and fSAPT0(3) scales only the sum of the two dispersion terms, $E_{disp}^{(20)}$ and $E_{exch-disp}^{(20)}$. We also tried applying scaling factors to SAPT2 thru SAPT2+(3), but this did not lead to significant improvements in accuracy.

2.3 Results

We tested a spectrum of quantum mechanical methods, spanning semiempirical (PMx), DFT, and SAPT, by comparing their results with reference interaction energies for a collection of noncovalent complexes in the gas phase. The reference collection, which comprised the A24, Ionic, JSCH, L7, SCAI, S22x5, S66x8 and X31x10 datasets from

BEGDB, totals 1,266 entries and includes a variety of molecules—nonpolar, polar, ionized and halogenated—in equilibrium and nonequilibrium geometries. Appendix A provides the interaction energies, calculated using the various QM methods for all the systems studied, along with the corresponding BEGDB reference energies.

The present quantum mechanical results were compared with the highest-accuracy reference energies available in the BEGDB for the datasets used. These were generated with CCSD(T)/CBS calculations including counterpoise corrections, except as noted in Methods. It is worth noting that the reference energies in the S22[83] and S66[146] datasets, which are more limited versions of the S22x5 and S66x8 datasets used here, were recently revised, based on larger basis sets, additional points for the CBS extrapolation, or both. The more rigorous results differ from those originally publisrrhed by 0.2 kcal/mol and 0.1 kcal/mol RMSE, respectively, with maximum unsigned errors of up to 0.7 kcal/mol. Given that the reference energies were not computed at such a high level, they also presumably have errors similar in magnitude. This uncertainty in the reference energies implies that the present study cannot meaningfully resolve errors less than 0.2 kcal/mol.

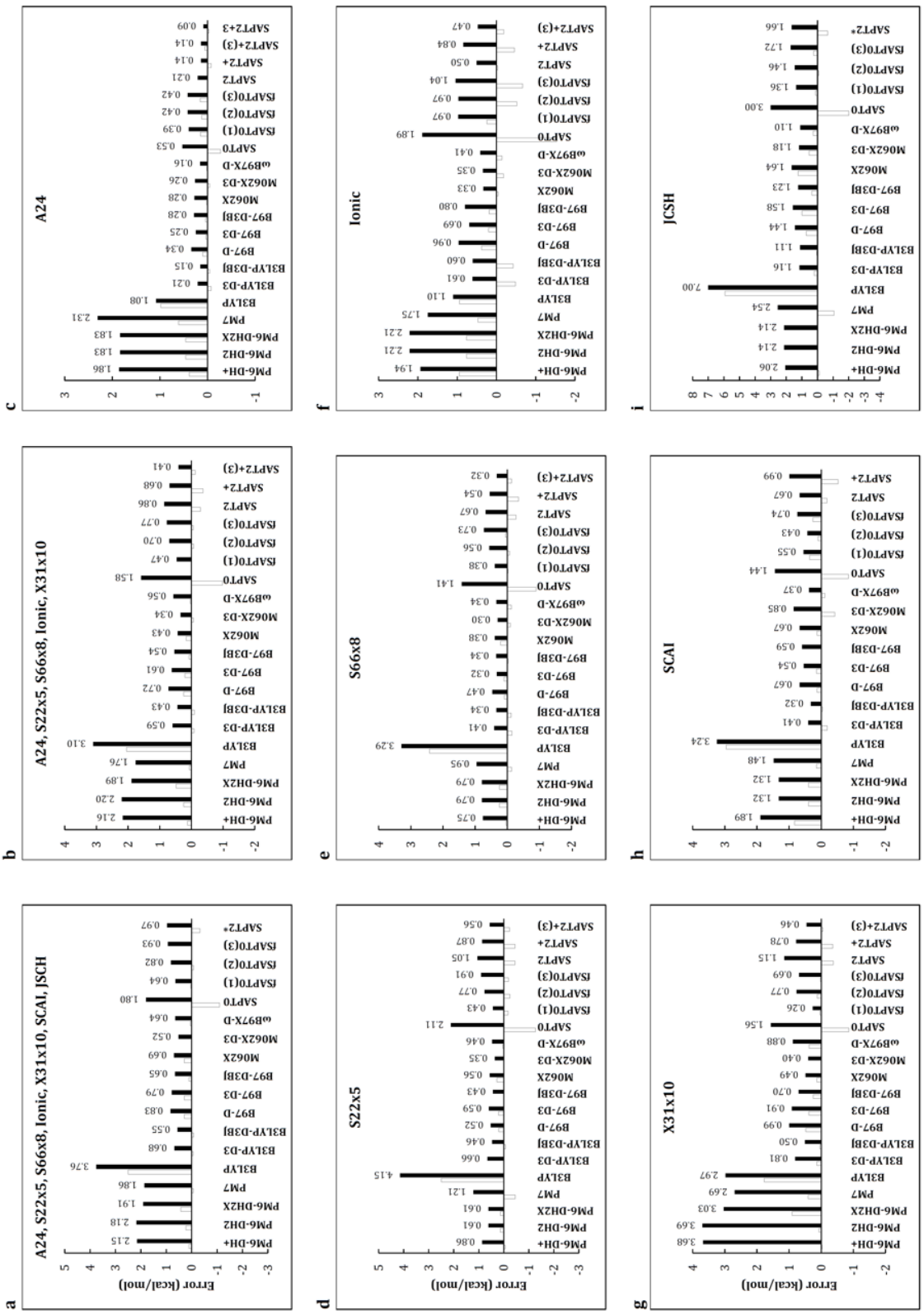
The following subsections provide an overview of the results, followed by more detailed discussions of the PMx, DFT and SAPT approaches. Timing comparisons are then given for the DFT and SAPT methods. Finally, we show that a simple scaling of SAPT0 energy components offers substantially improved accuracy at minimal computational cost.

2.3.1 Overview

As shown in Figure 2.2 and tabulated in Table 2.2, the RMSE values of the various quantum methods averaged across all datasets range from 0.52 to 3.76 kcal/mol. The methods which yield the lowest overall errors are SAPT2+(3) and M062X, both with and

without its dispersion correction, but a number of methods also yield overall RMSE values within 1 kcal/mol. The methods which yield highest overall errors are the semiempirical (PMx) methods, B3LYP without dispersion correction, and SAPT0. The MSE values are more informative; they show that both the PMx and DFT methods without dispersion corrections tend to provide interaction energies more positive (less favorable) than the reference results, while the SAPT methods tend to provide interaction energies more negative (more favorable) than the reference results. As expected, supplementing the DFT methods with negative dispersion energy terms reduces the tendency to overestimate the interaction energy; the resulting improvement is particularly striking on going from B3LYP to B3LYP-D3.

Figure 2.2: Evaluation of QM methods for combined and individual benchmark datasets. Errors evaluated relative to CCSD(T)/CBS CP energies. *SAPT2 calculations were only evaluated for 134 out of 143 of the JCSH systems.


a
A24, S22x5, S66x8, Ionic, X31x10, SCAI, JSCH
Error (kcal/mol)
PM6-DH+ 2.15
PM6-DH2 2.18
PM6-DHZX 1.91
PM7 1.86
B3LYP 3.76
B3LYP-D3 0.68
B3LYP-D3BJ 0.55
B97-D 0.83
B97-D3 0.79
B97-D3BJ 0.65
M06ZX 0.69
M06ZX-D3 0.52
ωB97X-D 0.64
SAPT0 1.80
fSAPT0(1) 0.64
fSAPT0(2) 0.82
fSAPT0(3) 0.93
SAPT2* 0.97
b
A24, S22x5, S66x8, Ionic, X31x10
PM6-DH+ 2.16
PM6-DH2 2.20
PM6-DHZX 1.89
PM7 1.76
B3LYP 3.10
B3LYP-D3 0.59
B3LYP-D3BJ 0.43
B97-D 0.72
B97-D3 0.61
B97-D3BJ 0.54
M06ZX 0.43
M06ZX-D3 0.34
ωB97X-D 0.56
SAPT0 1.58
fSAPT0(1) 0.47
fSAPT0(2) 0.70
fSAPT0(3) 0.77
SAPT2 0.86
SAPT2+ 0.68
SAPT2+(3) 0.41
c
A24
PM6-DH+ 1.86
PM6-DH2 1.83
PM6-DHZX 1.83
PM7 2.31
B3LYP 1.08
B3LYP-D3 0.21
B3LYP-D3BJ 0.15
B97-D 0.34
B97-D3 0.25
B97-D3BJ 0.28
M06ZX 0.28
M06ZX-D3 0.26
ωB97X-D 0.16
SAPT0 0.53
fSAPT0(1) 0.39
fSAPT0(2) 0.42
fSAPT0(3) 0.42
SAPT2 0.21
SAPT2+ 0.14
SAPT2+(3) 0.14
SAPT2+3 0.09
d
S22x5
Error (kcal/mol)
PM6-DH+ 0.86
PM6-DH2 0.61
PM6-DHZX 0.61
PM7 1.21
B3LYP 4.15
B3LYP-D3 0.66
B3LYP-D3BJ 0.46
B97-D 0.52
B97-D3 0.59
B97-D3BJ 0.43
M06ZX 0.56
M06ZX-D3 0.35
ωB97X-D 0.46
SAPT0 2.11
fSAPT0(1) 0.43
fSAPT0(2) 0.77
fSAPT0(3) 0.91
SAPT2 1.05
SAPT2+ 0.87
SAPT2+(3) 0.56
e
S66x8
PM6-DH+ 0.75
PM6-DH2 0.79
PM6-DHZX 0.79
PM7 0.95
B3LYP 3.29
B3LYP-D3 0.41
B3LYP-D3BJ 0.34
B97-D 0.47
B97-D3 0.32
B97-D3BJ 0.34
M06ZX 0.38
M06ZX-D3 0.30
ωB97X-D 0.34
SAPT0 1.41
fSAPT0(1) 0.38
fSAPT0(2) 0.56
fSAPT0(3) 0.73
SAPT2 0.67
SAPT2+ 0.54
SAPT2+(3) 0.32
f
Ionic
PM6-DH+ 1.94
PM6-DH2 2.21
PM6-DHZX 2.21
PM7 1.75
B3LYP 1.10
B3LYP-D3 0.61
B3LYP-D3BJ 0.60
B97-D 0.96
B97-D3 0.69
B97-D3BJ 0.80
M06ZX 0.33
M06ZX-D3 0.35
ωB97X-D 0.41
SAPT0 1.89
fSAPT0(1) 0.97
fSAPT0(2) 0.97
fSAPT0(3) 1.04
SAPT2 0.50
SAPT2+ 0.84
SAPT2+(3) 0.47
g
X31x10
Error (kcal/mol)
PM6-DH+ 3.68
PM6-DH2 3.69
PM6-DHZX 3.03
PM7 2.69
B3LYP 2.97
B3LYP-D3 0.81
B3LYP-D3BJ 0.50
B97-D 0.99
B97-D3 0.91
B97-D3BJ 0.70
M06ZX 0.49
M06ZX-D3 0.40
ωB97X-D 0.88
SAPT0 1.56
fSAPT0(1) 0.26
fSAPT0(2) 0.77
fSAPT0(3) 0.69
SAPT2 1.15
SAPT2+ 0.78
SAPT2+(3) 0.46
h
SCAI
PM6-DH+ 1.89
PM6-DH2 1.32
PM6-DHZX 1.32
PM7 1.48
B3LYP 3.24
B3LYP-D3 0.41
B3LYP-D3BJ 0.32
B97-D 0.67
B97-D3 0.54
B97-D3BJ 0.59
M06ZX 0.67
M06ZX-D3 0.85
ωB97X-D 0.37
SAPT0 1.44
fSAPT0(1) 0.55
fSAPT0(2) 0.43
fSAPT0(3) 0.74
SAPT2 0.67
SAPT2+ 0.99
i
JCSH
PM6-DH+ 2.06
PM6-DH2 2.14
PM6-DHZX 2.14
PM7 2.54
B3LYP 7.00
B3LYP-D3 1.16
B3LYP-D3BJ 1.11
B97-D 1.44
B97-D3 1.58
B97-D3BJ 1.23
M06ZX 1.64
M06ZX-D3 1.18
ωB97X-D 1.10
SAPT0 3.00
fSAPT0(1) 1.36
fSAPT0(2) 1.46
fSAPT0(3) 1.72
SAPT2* 1.66
MSE
RMSE

The performance of the quantum approaches varies significantly across the datasets, as shown in Figure 2.2. Perhaps most striking is that the PMx methods provide substantially better relative results for the S22x5, S66x8, SCAI, and JCSH datasets, and worse results for A24, Ionic and X31x10. The problems for A24 and X31x10 appear to arise largely from errors associated specifically with halogenated molecules. The Ionic dataset includes no halogens, however, and we speculate that the errors here may trace to the lack of ionized hydrogen-bonded complexes in the datasets used to parameterize the PMx methods. In addition, the minimal basis sets used in the PMx methods may have difficulty accounting for the strong polarization effects present in such ionized complexes. Other than B3LYP, lower-order SAPT and PMx, all methods are within 1 kcal/mol RMSE of the reference energies for all sets except JSCH. For JSCH, all approaches yield larger errors (note the scale of the vertical axis in Figure 2.2c). This is perhaps not surprising, because the JSCH dataset contains the largest dimer systems, and one may expect larger systems to effectively include more interactions, each potentially associated with some level of errorFigure 2.1. Ranking the methods according to their overall accuracy and their accuracy on each data set, Table 2.2 shows that, although certain methods remain near the top of the rankings across the board, the detailed ordering of the methods varies across datasets.

The scatter plots in Figure 2.3 provide further insight into the performance of the various approaches. All the methods tested provide excellent correlation with the reference energies (R^2>0.86), and, not surprisingly, those with the largest RMSE values Figure 2.2 also yield the lowest R^2 values (Figure 2.3). This analysis also allows further characterization of the errors associated with some of the methods. First, the PMx scatter plots include outliers arranged in smooth arcs. Further analysis indicates that each arc corresponds to the dissociation curve of one dimer system, and the dimer systems which generate these

Table 2.2: Ranking of QM methods by RMSE for combined and individual benchmark datasets (kcal/mol). Dashed lines mark RMSE levels of 0.50 and 1.00 kcal/mol. Higher orders of SAPT calculations were not completed for some datasets, due to memory limitations: A24 was evaluated at orders through SAPT2+3, SCAI through SAPT2+, JSCH through SAPT2, and all other dataset through SAPT2+(3). *SAPT2 calculations were only evaluated for 134 out of 143 of the JCSH systems

	A24, Ionic, S22x5, S66x8, X31x10, SCAI, JSCH		A24, Ionic, S22x5, S66x8, X31x10		A24		Ionic		S22x5		S66x8		X31x10		SCAI		JSCH	
1	M062X-D3	*0.52*	M062X-D3	*0.34*	SAPT2+3	*0.09*	M062X	*0.33*	M062X-D3	*0.35*	M062X-D3	*0.30*	fSAPT0(1)	*0.26*	B3LYP-D3BJ	*0.32*	ωB97X-D	*1.10*
2	B3LYP-D3BJ	*0.55*	SAPT2+(3)	*0.41*	SAPT2+(3)	*0.14*	M062X-D3	*0.35*	B97-D3BJ	*0.43*	B97-D3	*0.32*	M062X-D3	*0.40*	ωB97X-D	*0.37*	B3LYP-D3BJ	*1.11*
3	fSAPT0(1)	*0.64*	M062X	*0.43*	SAPT2+	*0.14*	ωB97X-D	*0.41*	fSAPT0(1)	*0.43*	SAPT2+(3)	*0.32*	SAPT2+(3)	*0.46*	B3LYP-D3	*0.41*	B3LYP-D3	*1.16*
4	ωB97X-D	*0.64*	B3LYP-D3BJ	*0.43*	B3LYP-D3BJ	*0.15*	SAPT2+(3)	*0.47*	B3LYP-D3BJ	*0.46*	B3LYP-D3BJ	*0.34*	M062X	*0.49*	fSAPT0(2)	*0.43*	M062X-D3	*1.18*
5	B97-D3BJ	*0.65*	fSAPT0(1)	*0.47*	ωB97X-D	*0.16*	SAPT2	*0.50*	ωB97X-D	*0.46*	ωB97X-D	*0.34*	B3LYP-D3BJ	*0.50*	B97-D3	*0.54*	B97-D3BJ	*1.23*
6	B3LYP-D3	*0.68*	B97-D3BJ	*0.54*	B3LYP-D3	*0.21*	B3LYP-D3BJ	*0.60*	B97-D	*0.52*	B97-D3BJ	*0.34*	fSAPT0(3)	*0.69*	fSAPT0(1)	*0.55*	fSAPT0(1)	*1.36*
7	M062X	*0.69*	ωB97X-D	*0.56*	SAPT2	*0.21*	B3LYP-D3	*0.61*	SAPT2+(3)	*0.56*	fSAPT0(1)	*0.38*	B97-D3BJ	*0.70*	B97-D3BJ	*0.59*	B97-D	*1.44*
8	B97-D3	*0.79*	B3LYP-D3	*0.59*	B97-D3	*0.25*	B97-D3	*0.69*	M062X	*0.56*	M062X	*0.38*	fSAPT0(2)	*0.77*	SAPT2	*0.67*	fSAPT0(2)	*1.46*
9	fSAPT0(2)	*0.82*	B97-D3	*0.61*	M062X-D3	*0.26*	B97-D3BJ	*0.80*	B97-D3	*0.59*	B3LYP-D3	*0.41*	SAPT2+	*0.78*	B97-D	*0.67*	B97-D3	*1.58*
10	B97-D	*0.83*	SAPT2+	*0.68*	M062X	*0.28*	SAPT2+	*0.84*	PM6-DH2	*0.61*	B97-D	*0.47*	B3LYP-D3	*0.81*	M062X	*0.67*	M062X	*1.64*
11	fSAPT0(3)	*0.93*	fSAPT0(2)	*0.70*	B97-D3BJ	*0.28*	B97-D	*0.96*	PM6-DH2X	*0.61*	SAPT2+	*0.54*	ωB97X-D	*0.88*	fSAPT0(3)	*0.74*	SAPT2*	*1.66*
12	SAPT2*	*0.97*	B97-D	*0.72*	B97-D	*0.34*	fSAPT0(2)	*0.97*	B3LYP-D3	*0.66*	fSAPT0(2)	*0.56*	B97-D3	*0.91*	M062X-D3	*0.85*	fSAPT0(3)	*1.72*
13	SAPT0	*1.80*	fSAPT0(3)	*0.77*	fSAPT0(1)	*0.39*	fSAPT0(1)	*0.97*	fSAPT0(2)	*0.77*	SAPT2	*0.67*	B97-D	*0.99*	SAPT2+	*0.99*	PM6-DH+	*2.06*
14	PM7	*1.86*	SAPT2	*0.86*	fSAPT0(3)	*0.42*	fSAPT0(3)	*1.04*	PM6-DH+	*0.86*	fSAPT0(3)	*0.73*	SAPT2	*1.15*	PM6-DH2	*1.32*	PM6-DH2	*2.14*
15	PM6-DH2X	*1.91*	SAPT0	*1.58*	fSAPT0(2)	*0.42*	B3LYP	*1.10*	SAPT2+	*0.87*	PM6-DH+	*0.75*	SAPT0	*1.56*	PM6-DH2X	*1.32*	PM6-DH2X	*2.14*
16	PM6-DH+	*2.15*	PM7	*1.76*	SAPT0	*0.53*	PM7	*1.75*	fSAPT0(3)	*0.91*	PM6-DH2	*0.79*	PM7	*2.69*	SAPT0	*1.44*	PM7	*2.54*
17	PM6-DH2	*2.18*	PM6-DH2X	*1.89*	B3LYP	*1.08*	SAPT0	*1.89*	SAPT2	*1.05*	PM6-DH2X	*0.79*	B3LYP	*2.97*	PM7	*1.48*	SAPT0	*3.00*
18	B3LYP	*3.76*	PM6-DH+	*2.16*	PM6-DH2	*1.83*	PM6-DH+	*1.94*	PM7	*1.21*	PM7	*0.95*	PM6-DH2X	*3.03*	PM6-DH+	*1.89*	B3LYP	*7.00*
19			PM6-DH2	*2.20*	PM6-DH2X	*1.83*	PM6-DH2	*2.21*	SAPT0	*2.11*	SAPT0	*1.41*	PM6-DH+	*3.68*	B3LYP	*3.24*		
20			B3LYP	*3.10*	PM6-DH+	*1.86*	PM6-DH2X	*2.21*	B3LYP	*4.15*	B3LYP	*3.29*	PM6-DH2	*3.69*				
21					PM7	*2.31*												

arcs are ones for which the PMx method yields idiosyncratically high errors, as discussed below. Second, most of the errors of the B3LYP method are associated with dimer systems whose interactions are primarily dispersive, as indicated by the red cluster of off-diagonal points. These errors are largely corrected by addition of the D3 dispersion correction. In contrast, the tendency of the SAPT methods to overestimate dimer affinities is largely independent of interaction type, as points of all colors are found below the lines of identity in the SAPT scatter plots. Third, adding a dispersion correction to the DFT methods uniformly improves the correlation, and the D3 correction performs somewhat better than D2, where the comparison is made (B97-D versus B97-D3). Finally, the DFT methods have a weak tendency to overestimate the unfavorable energy of the most repulsive interactions, while the PMx and SAPT methods tend to assign overly favorable energies in these cases. These repulsive interactions tend to have intermediate electrostatic-dispersive character, as indicated by the cyan color of these points.

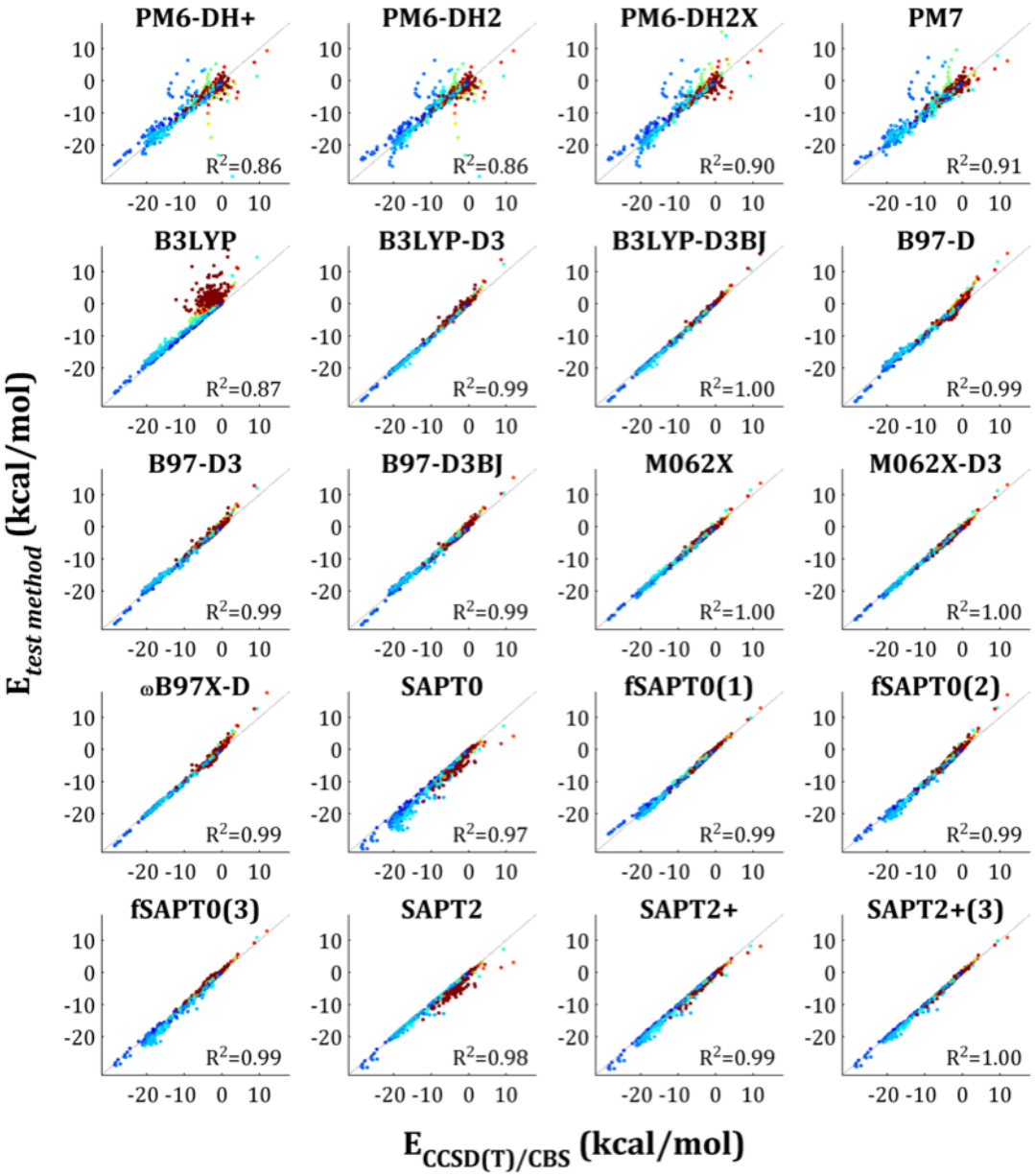


Figure 2.3: Correlation of QM methods with CCSD(T)/CBS CP. We present only those datasets to which all methods could be applied; i.e., A24, Ionic, S22x5, S66x8, X31x10. Each entry is colored by interaction type character (see spectrum below), which is defined as $|E_{disp}/E_{elst}|$, where E_{disp} and E_{elst} are the total electrostatic and dispersion energy components taken from the SAPT2+(3) calculations.

It is of interest to examine how the performance of the various methods depends on whether they are applied to equilibrium versus nonequilibrium geometries. Datasets S22x5,

S66x8, X31x10 and Ionic make such comparisons possible, as they contain dissociation curves for a total of 134 dimer systems (see Methods). Figure 2.4 compares the MSE and RMSE for each method at close, equilibrium and far separations, defined here as 90%, 100% and 200% of the equilibrium separations, respectively. The rankings of the methods for equilibrium geometries correlate well with the rankings for the close geometries, but poorly with rankings a far separations (Table 2.3).

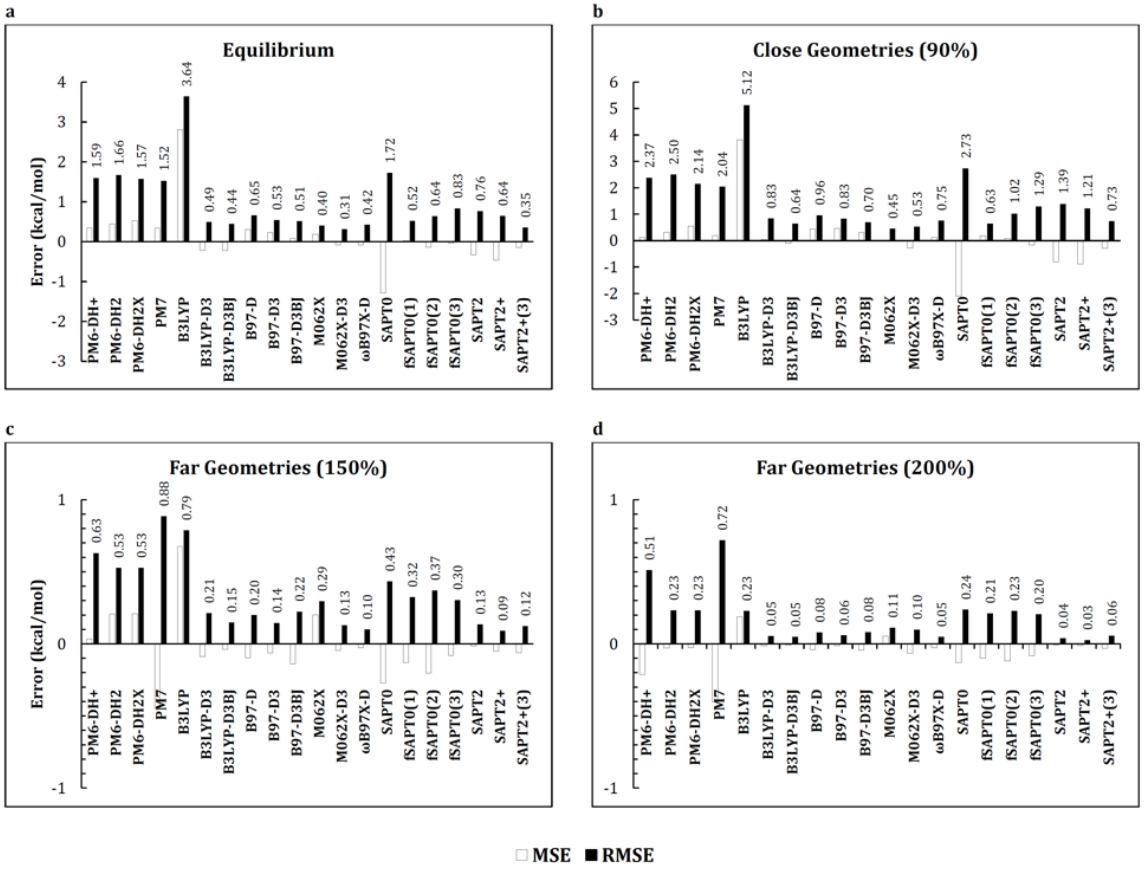


Figure 2.4: Evaluation of QM methods for equilibrium and nonequilibrium geometries. Geometries from the S22x5, S66x8, X31x10 and Ionic datasets. Errors evaluated relative to CCSD(T)/CBS CP energies.

Given that the long-range interactions are smaller in absolute terms, this observation suggests that a study of equilibrium geometries suffices to determine which methods work best overall. On the other hand, the errors rise at short distance for all methods, so that none

Table 2.3: Ranking of QM methods by RMSE for equilibrium and nonequilibrium geometries (kcal/mol). Dashed lines mark RMSE levels of 0.5 and 1.0 kcal/mol.

	Equilibrium		**90%**		**150%**		**200%**	
1	M062X-D3	*0.31*	M062X	*0.45*	SAPT2+	*0.09*	SAPT2+	*0.03*
2	SAPT2+(3)	*0.35*	M062X-D3	*0.53*	ωB97X-D	*0.10*	SAPT2	*0.04*
3	M062X	*0.40*	fSAPT0(1)	*0.63*	SAPT2+(3)	*0.12*	ωB97X-D	*0.05*
4	ωB97X-D	*0.42*	B3LYP-D3BJ	*0.64*	M062X-D3	*0.13*	B3LYP-D3BJ	*0.05*
5	B3LYP-D3BJ	*0.44*	B97-D3BJ	*0.70*	SAPT2	*0.13*	B3LYP-D3	*0.05*
6	B3LYP-D3	*0.49*	SAPT2+(3)	*0.73*	B97-D3	*0.14*	SAPT2+(3)	*0.06*
7	B97-D3BJ	*0.51*	ωB97X-D	*0.75*	B3LYP-D3BJ	*0.15*	B97-D3	*0.06*
8	fSAPT0(1)	*0.52*	B97-D3	*0.83*	B97-D	*0.20*	B97-D	*0.08*
9	B97-D3	*0.53*	B3LYP-D3	*0.83*	B3LYP-D3	*0.21*	B97-D3BJ	*0.08*
10	fSAPT0(2)	*0.64*	B97-D	*0.96*	B97-D3BJ	*0.22*	M062X-D3	*0.10*
11	SAPT2+	*0.64*	fSAPT0(2)	*1.02*	M062X	*0.29*	M062X	*0.11*
12	B97-D	*0.65*	SAPT2+	*1.21*	fSAPT0(3)	*0.30*	fSAPT0(3)	*0.20*
13	SAPT2	*0.76*	fSAPT0(3)	*1.29*	fSAPT0(1)	*0.32*	fSAPT0(1)	*0.21*
14	fSAPT0(3)	*0.83*	SAPT2	*1.39*	fSAPT0(2)	*0.37*	B3LYP	*0.23*
15	PM7	*1.52*	PM7	*2.04*	SAPT0	*0.43*	fSAPT0(2)	*0.23*
16	PM6-DH2X	*1.57*	PM6-DH2X	*2.14*	PM6-DH2	*0.53*	PM6-DH2X	*0.23*
17	PM6-DH+	*1.59*	PM6-DH+	*2.37*	PM6-DH2X	*0.53*	PM6-DH2	*0.23*
18	PM6-DH2	*1.66*	PM6-DH2	*2.50*	PM6-DH+	*0.63*	SAPT0	*0.24*
19	SAPT0	*1.72*	SAPT0	*2.73*	B3LYP	*0.79*	PM6-DH+	*0.51*
20	B3LYP	*3.64*	B3LYP	*5.12*	PM7	*0.88*	PM7	*0.72*

provide excellent accuracies for the close geometries, and only M062X has an RMSE below 0.5 kcal/mol. At far separations, all methods are within 1.0 kcal/mol RMSE, and several fall under 0.2 kcal/mol RMSE, which is comparable to the size of errors associated with basis set choice in computing the CCSD(T) correction, as discussed above.

2.3.2 Semiempirical PMx Methods

The semiempirical PMx methods are roughly comparable in accuracy to the DFT-D3 and higher-order SAPT methods for the S22x5, S66x8 ,and JSCH datasets, but is considerably less accurate for A24, Ionic and X31x10, much as previously noted[74]. We conjecture that this difference stems in part from the fact that the PMx methods considered here, as well as their corrections (e.g. DH+), were parameterized using systems similar in character to those in the S22x5, S66x8 and JSCH datasets. In addition, as suggested by the scatter plots in Figure 2.3, some of the larger errors of the PMx methods are associated with specific systems for which they give idiosyncratically poor results. In particular, the PMx methods supply problematic results for bromobenzene$\cdots$trimethylamine and systems containing HF; i.e., HF dimer, HF$\cdots$methane, HF$\cdots$methanol, and HF$\cdots$methylamine. Accordingly, omitting these problematic cases significantly improves the RMSE of the PMx methods by 0.8-1.8 kcal/mol for the A24 and X31x10 datasets, and by 0.2-0.7 kcal/mol across the full reference collection of datasets, as shown in b. The bromine-nitrogen problem, as found here in the bromobenzene$\cdots$trimethylamine system, is a known issue for the PM6 method, and is improved by the halogen ('X') correction for PM6, or by going to the PM7 method. However, we have not found previous comments on the issue for HF, and we are not aware of a correction for it. The fact that HF is problematic for all of the PMx methods is evident from the fact that the corresponding RMSE values for the A24 dataset, which lacks

the bromobenzene$\cdots$trimethylamine system, improve by 0.5-0.9 kcal/mol when only the HF systems are omitted. When the bromobenzene$\cdots$trimethylamine and HF-containing systems are omitted from the PMx results, their RMSE values across all systems fall to 1.4-1.6 kcal/mol (Table 2.4, bottom row, right). However, this improvement does not significantly change their position in the rankings in Table 2.2.

Table 2.4: Evaluation of PMx methods with and without problematic dimer cases. Errors are presented as RMSE values, in kcal/mol.

	Original				No bromobenzene trimethylamine or HF			
	PM6-DH+	PM6-DH2	PM6-DH2X	PM7	PM6-DH+	PM6-DH2	PM6-DH2X	PM7
A24	1.86	1.83	1.83	2.31	0.88	0.82	0.82	0.93
X31x10	3.68	3.69	3.03	2.69	1.88	1.92	2.26	1.72
All	2.16	2.20	1.89	1.76	1.44	1.48	1.59	1.52

No single PMx method emerges as the most reliable from these data. For example, although PM7 has a slightly better RMSE across the entire data collection than the corrected PM6 methods, it is not clear how significant this difference is, as its relative performance is quite inconsistent across the separate datasets (Figure 2.2). The PM6-DH2 and PM6-DH2X methods are equivalent for all systems except those containing halogens, and thus produce identical results for S22x5, S66x8, Ionic, SCAI and JSCH datasets. As expected, using PM7 or applying the "X" correction provides significant improvement in RMSE for the halogen-containing X31x10 dataset. However, removing the specific problem systems mentioned in the prior paragraph essentially eliminates the advantage of these more advanced methods. The utility of the "X" correction, which is specifically designed to improve the treatment of halogens, may be examined more closely by comparing the various PMx methods for the full X40x10 dataset, for which all systems contain at least one halogen atom, and which includes both equilibrium and non-equilibrium distances. PM6-DH2X is more accurate than

PM6-DH2 at all distances, except that the "X" correction generates a particularly large error (22.6 kcal/mol RMSE) for dimers at 80% of their equilibrium distances, as shown Table A.2 (Appendix A). Interestingly, when the iodine-containing systems are omitted, to create the X31x10 subset of X40x10, the "X" correction yields improved or equal results at all distances, and the anomalously high error at short range is absent. Thus, although the "X" correction yields an overall improvement, it seems problematic for the particular case of short-ranged interactions involving iodine. The PM7 method lacks this short-range anomaly, but is somewhat less accurate for the iodinated compounds at longer ranges.

2.3.3 DFT with and without Dispersion Corrections

The DFT methods which incorporate some treatment of dispersion show good overall performance, with RMSE values ranging from 0.52 to 0.83 kcal/mol. In contrast, uncorrected B3LYP yields a rather large RMSE of 3.76 kcal/mol, and its largest errors are associated chiefly with dispersive systems (red in Figure 2.2), for which the method underestimates the attractive forces. Supplementing B3LYP with attractive D3 dispersion corrections markedly improves the overall RMSE across all test systems to 0.68 kcal/mol with D3 and to 0.55 kcal/mol with D3BJ. For the B3LYP functional, the BJ-damped version of the D3 correction typically produces results closer to reference compared with zero-damped D3. Interestingly, there is no marked improvement going from the two-body D2 correction to the three-body D3 correction, in the context of the B97-D method, even for the larger systems in the JSCH and SCAI datasets, where three-body contributions are expected to be more important. Furthermore, while B97-D3BJ has a lower overall RMSE across all systems compared with B97-D3 (0.65 kcal/mol compared to 0.79 kcal/mol), the former produces higher RMSE values for the Ionic and SCAI datasets. Thus, it is difficult to gauge

the benefit of applying BJ-damping over zero-damping for B97-D. The uncorrected M062X method performs equally well for electrostatic and dispersive systems (Figure 2.3), but its accuracy appears to be slightly improved by supplementing it with the D3 dispersion term (Figure 2.2). The ωB97X-D method includes its own dispersion correction distinct from D2 or D3, and this method ranks well across all the datasets. It is perhaps worth noting that the counterpoise corrections for all for DFT methods (B3LYP, B97-D, M062X, and ωB97X-D) are small, averaging 0.15 kcal/mol across all methods and systems. The mean correction rises only slightly, to 0.21 kcal/mol for nonequilibrium systems at close range (90% of equilibrium separation). These corrections are small, in the sense that they are similar in magnitude to the uncertainty in the reference energies used here, as discussed above. Finally, it is worth noting that the low errors observed for D3-corrected DFT functionals in S22x5 are, perhaps, unsurprising, since the same molecules in similar geometries were included in the dataset used to parameterize DFT-D3.

2.3.4 SAPT

The accuracy of the SAPT approach tends to increase with order, as expected, and the higher orders are comparable in accuracy to the best DFT methods (Figure 2.2). The trend of increased accuracy with increased order holds for all individual datasets, except Ionic, for which SAPT2 yields a lower RMSE than SAPT2+. Since SAPT2+ differs from SAPT2 by only two dispersion terms, it is interesting that the inclusion of these terms seemingly degrades accuracy here. Perhaps the excellent performance of SAPT2 for this particular dataset results from a fortuitous cancellation of errors. It is worth noting that all orders of SAPT tend to overestimate attractive forces, regardless of system character, as evident from the negative MSE values in Figure 2.2 and by inspection of the scatter plots in Figure

2.3. This overestimation is particularly marked for SAPT0, suggesting the presence of a systematic error that might be mitigated by a post-calculation correction. Finally, because the SAPT energy components can be useful for tuning individual force field terms[111], we provide, in the supplementary information, the detailed SAPT2+(3) decompositions for all the dimer systems studied here.

2.3.5 Timing Analysis

We used the A24 dataset to compare the computational speeds of the various methods. The PMx methods all finished in less than 0.02s real (wall clock) time on a single CPU, making them over 1000 times faster than the DFT or SAPT methods. The latter were timed for all A24 systems on 8 dedicated CPUs, and Figure 2.5 plots real time against system size, as measured by the number of atoms, while Figure 2.6 plots the tradeoff between accuracy and computer time. Overall, SAPT0 is clearly the fastest approach, SAPT2+(3) is the slowest, and the DFT timings are rather similar to each other and to SAPT2+. The level of accuracy broadly correlates with computer time, except in the case of uncorrected B3LYP.

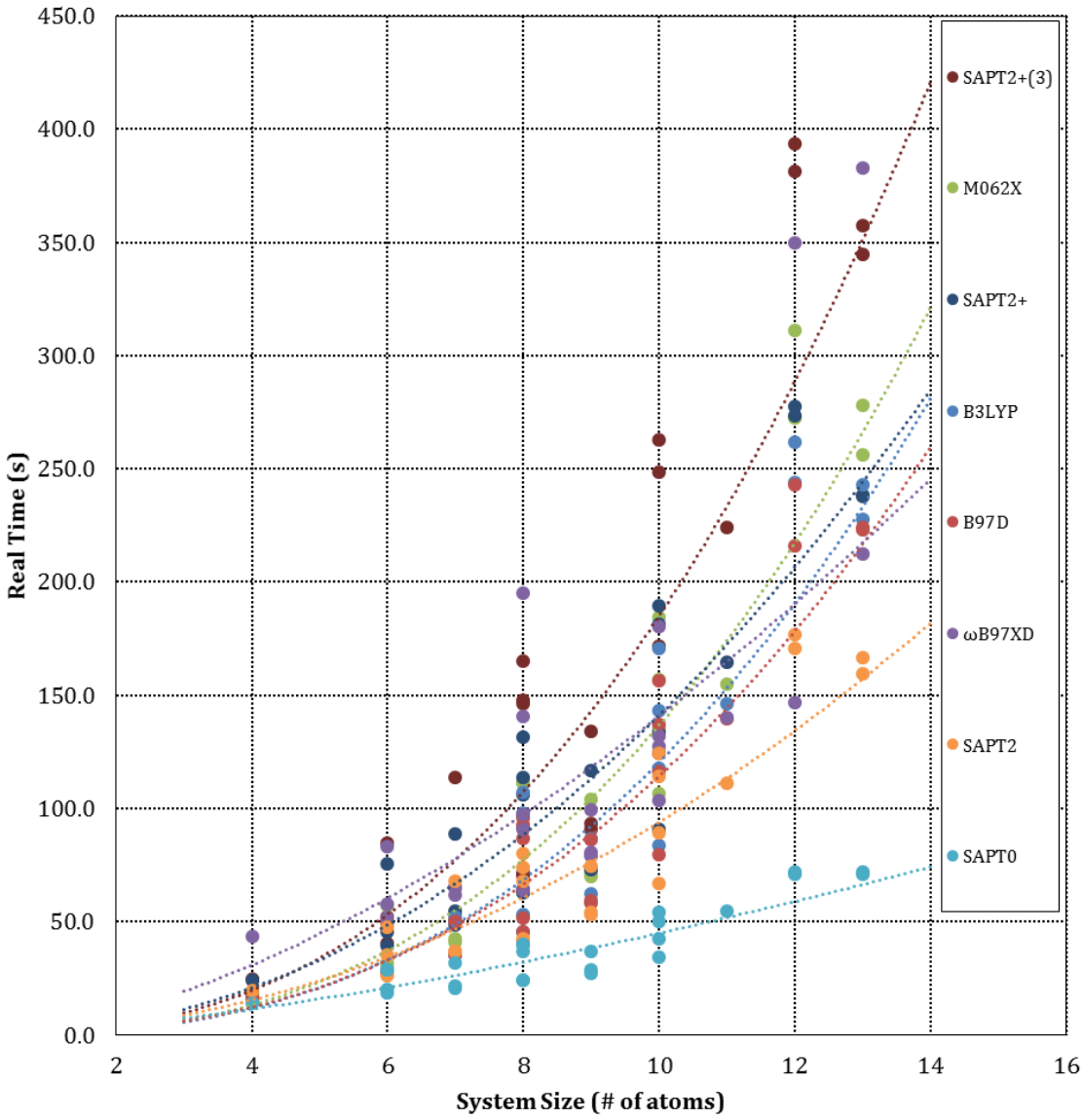


Figure 2.5: Scaling of calculation time with system size. Results are presented for the A24 dataset, where system size is measured by the number of atoms in each dimer.

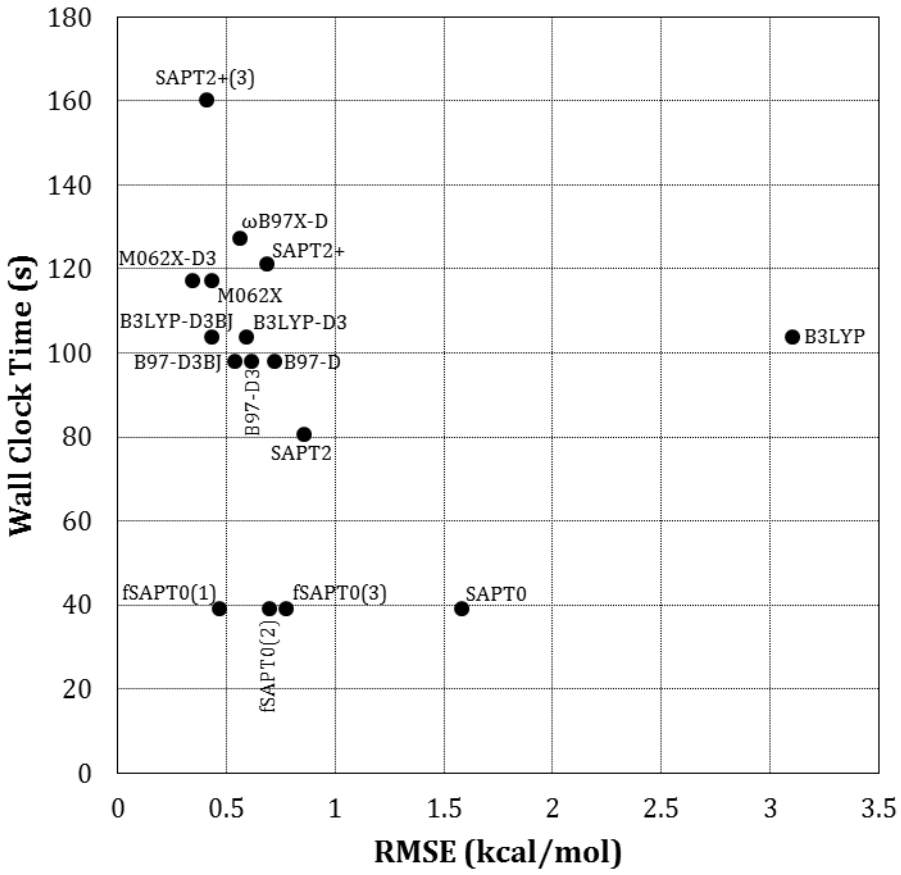


Figure 2.6: Tradeoff between accuracy and calculation time. Accuracies are presented as RMSE across the A24, Ionic, S22x5, S66x8, and X31x10 datasets, while calculation times are averaged for the A24 dataset alone. Note that the *post hoc* D3 dispersion corrections and the SAPT0 fitting require negligible calculation time.

The scaling of computer time with system size was examined by fitting the timings for each method to a power model of the form t=anb, where t is real time and n is the number of atoms or electrons. The curve fits are detailed in Table A.3 (Appendix A). As shown in Figure 2.5, all of the DFT methods except ωB97X-D have exponents of about 2.5 and prefactors of about 0.4. The ωB97X-D DFT method appears to scale rather differently, as its exponent and prefactor are 1.65 and 3.13, respectively. On the other hand, the R^2 value of its fit to the power model is only 0.67, so its scaling behavior is not clearly defined by

these data. The exponents for the SAPT methods increase with order: SAPT0 scales as the number of atoms to the power 1.49, while the SAPT2+(3) time varies as the 2.44 power. Analogous trends across the methods are observed when one fits the timings to the number of electrons in the system, rather than the number of atoms. Perhaps surprisingly, however, the R^2 values of the fits are much lower, as evident in Table A.3.

2.3.6 Linear Scaling of SAPT0 Energy Terms

Of the methods tested here, SAPT0 is faster than all but the semiempirical PMx methods, as detailed above. The fact that it decomposes the total dimer interaction energy into seven contributions, which capture aspects of electrostatics, exchange, induction and dispersion, provides an opportunity to try generating a fast method with improved accuracy by scaling these contributions, as detailed in Methods. Table 2.5 lists the means and standard deviations of the resulting scaling coefficients for the SAPT0 terms, across the 1000 different training sets, and the mean and standard deviations of the RMSE and R^2 values when the trained coefficients are applied to the respective test sets. Most of the scaling coefficients are near unity; the term which requires the most scaling is the $E^{(20)}_{exch-disp}$ term. Scaling all terms, in fSAPT0(1), produces the lowest test-set RMSE, followed by scaling the dispersion terms individually, in fSAPT0(2), and then by scaling the summed dispersion terms, in fSAPT0(3). The fact that these results are obtained on test-sets not used to set the coefficients means that the improvement in performance for the more highly parameterized models do not reflect overfitting. The accuracy of the three scaling schemes is also compared with the various QM methods in Figures 2.2 and 2.3 and Tables 2.2 and 2.3. Across all datasets, except L7, applying scaling factors to the SAPT0 terms reduces the RMSE from 1.58 kcal/mol to as low as 0.47 kcal/mol, and corrects the tendency of SAPT0 to overestimate the attractive

nature of the dimer interactions. Indeed, the fitted SAPT0 results approach the accuracy of the DFT methods, with the differences within the estimates of CCSD(T)/CBS basis set choice errors (above). Note that this improvement in SAPT0, through the application of simple scaling factors, incurs negligible additional computational cost, so that the fSAPT0 scaling methods provide a particularly favorable combination of efficiency and accuracy, as shown in Figure 2.6. Figure A.1 (Appendix A) furthermore examines the accuracy of fSAPT0, as well as the other QM methods, for the large noncovalent complexes of the L7 dataset; the results are generally consistent with those obtained for the other datasets. The energy components of the fitted SAPT0 method still correlate well with the corresponding energy components calculated at the SAPT2+(3) level, as detailed in Table A.4 (Appendix A). The good agreement suggests that the energy decomposition derived using the scaled terms is still physically meaningful.

2.4 Discussion

The present study systematically evaluates the accuracy and speed of a broad range of electronic structure methods for estimating noncovalent interaction energies. Methods spanning PMx, DFT and SAPT were applied to over 1,200 geometries of gas-phase dimers drawn from the BEGDB resource, which is tailored to probe a variety of interaction motifs relevant to biomolecules and drug-like compounds. These results offer useful guidance regarding which methods are most suitable for various types of applications where "gold-standard" CCSD(T)/CBS CP calculations are too time-consuming or impractical, as now discussed. Key findings and implications are now discussed

The PMx methods studied here are dramatically faster than both the DFT and SAPT

Table 2.5: Linear scaling factors for SAPT0/aug-cc-pVTZ energy terms. Three different fitting schemes were tested: fSAPT0(1) scales all terms; fSAPT0(2) scales only the two dispersion terms, $E^{(20)}_{disp}$ and $E^{(20)}_{exch-disp}$, treated independently; and fSAPT0(3) scales only the sum of the two dispersion terms, $E^{(20)}_{disp}$ and $E^{(20)}_{exch-disp}$. The scaling factors were determined over 1,000 iterations of multiple linear regression on randomly selected training subsets of the dimer systems, while RMSE and R^2 were evaluated over the same iterations using test subsets comprising all dimer systems not included in the training subset. Training and test subsets were equal in size.

	fSAPT0(1)	fSAPT0(2)	fSAPT0(3)
$E^{(10)}_{elst,r}$	1.01±0.02	1.00*	1.00*
$E^{(10)}_{exch}$	1.02±0.02	1.00*	1.00*
$E^{(20)}_{ind,r}$	0.76±0.08	1.00*	1.00*
$E^{(20)}_{exch-ind,r}$	0.70±0.08	1.00*	1.00*
$\delta E^{(2)}_{HF,r}$	1.06±0.08	1.00*	1.00*
$E^{(20)}_{disp}$	0.93±0.01	0.96±0.02	0.76†±0.01
$E^{(20)}_{exch-disp}$	1.7±0.2	2.1±0.2	0.76†±0.01
Test RMSE	0.66±0.06	0.82±0.05	0.93±0.04
Test R^2	0.995±0.009	0.993±0.001	0.992±0.001

*Not fitted. †Both dispersion terms share a single fitted coefficient.

approaches, and they are more readily applied to larger molecules. However, they are in general less accurate, particularly for halogenated and ionic molecules, as well as for a few types of systems with idiosyncratic results[1]. Perhaps surprisingly, none of the various PMx methods tested here is clearly superior to the others, in terms of overall accuracy. The DFT methods are slower and more difficult to apply to large systems, but they can achieve high accuracy, so long as dispersion is accounted for, either implicitly, as in M062X, or via an add-on term, as in B97-D3. The performance of the SAPT approach depends strongly on the order of the SAPT expansion. The SAPT2+ and SAPT2+(3) orders span the range of accuracy seen for the dispersion-corrected DFT methods. However, while the speed of the SAPT2+ method is comparable to that of the DFT methods, the more accurate SAPT2+(3)

is considerably slower. It is also worth noting that, at least in the current PSI4 software, the memory requirements of SAPT at orders higher than SAPT0 can become problematic for the larger systems examined here. The lowest order of SAPT, SAPT0, is similar in accuracy to the PMx methods, but significantly slower. However, we find that a simple empirical scaling of one or more SAPT0 energy terms leads to accuracy approaching that of the best DFT methods, at far less computational cost. With further development, an empirically adjusted SAPT0 approach might provide a powerful alternative to DFT methods for the study of noncovalent interactions in larger systems.

The results of this study have implications for improving the treatment of noncovalent interactions in molecular modeling, as QM calculations are used to guide the development of force fields for simulation, and may even replace force fields in some applications. The more accurate DFT and DFT-D3 methods maybe most suitable for force field parameterization, given their reliability and consistency across many types of molecular systems, and the fact that their moderate computational cost is not a major liability for this application. Despite the high speed of the PMx methods, their lower accuracy, especially for ionic systems and halogens, along with occasional idiosyncratic performance, makes them less suitable for parameterization of force fields. However, continued development of such semiempirical methods, including training on broader datasets, remains promising. In addition, these methods may already be more accurate than typical simulation force fields, so their high speed makes them a reasonable choice for direct modeling of biomolecular systems. The higher order SAPT methods are about as accurate as DFT, but are relatively slow, while SAPT0 is fast but inaccurate. Interestingly, the scaled SAPT0 method offers a promising blend of accuracy and computational speed, especially for larger molecular systems. In addition, the present scaling approach is relatively simple, and more sophisticated schemes which

account for geometry and chemistry might be even more accurate at minimal computational cost.

2.5 Acknowledgements

This work was supported in part by Grants R01GM61300 and T32EB93803 from the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Computing resources were supported by the NSF XSEDE program (Gordon and Trestles systems at the SDSC) and by the NIGMS (P41GM103426) of the NIH (computing cluster hosted by NBCR). The authors would like to thank Dr. David Sherrill and his group for their help with PSI4, the Gaussian support team for their help with Gaussian 09, and Dr. Jan Jensen for valuable comments and suggestions.

Chapter 2, in full, is a reprint of the material as it appears in the Journal of Chemical Theory and Computation 2014, Li, Amanda; Muddana, Hari S.; Gilson, Michael K. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Evaluation of Representations and Response Models for Polarizable Force Fields

3.1 Introduction

Classical simulations of condensed phase molecular systems rely on potential functions, or force fields, to map the spatial coordinates of the atomic nuclei to the potential energy of the system and the force on each atom. Commonly used force fields model electrostatic interactions in terms of Coulombic interactions among atom-centered point charges with fixed values[82, 22, 36, 182, 124, 107, 108, 130]. This functional form strikes a practical balance between computational efficiency and accuracy, and thus has found wide application. Indeed, we have argued that, given the modest scope of experimental data used so far to adjust force field parameters, force fields using this functional form can become even more accurate[66]. On the other hand, their accuracy must ultimately be

limited by their neglect of configuration-dependent changes in electronic polarization; i.e., shifts in molecular electron densities induced by variations in the electrical field felt by a molecule, as the system evolves over time. This limitation of the fixed-charge model has motivated development of force fields that incorporate configuration-dependent electronic polarization[132, 31, 125, 176, 12, 148].

One functional form used to model electronic polarization keeps the partial charges constant and adds atom-centered dipoles, whose moments vary with the local electric field. A well-regarded version of this functional form uses atom-centered point polarizabilities, where the field felt by each point polarizability is that generated by the point charges and the other induced dipoles in the system. This implementation is computationally burdensome, however, because it requires solving a matrix equation for the self-consistent set of induced dipoles. This problem may be moderated by improved mathematical approaches [159], or removed entirely by variants where the field felt by each point polarizability omits any contribution from the other induced dipoles[167, 151, 89, 183]. These direct, or first-order, methods are fast, because there is no self-consistent matrix problem to solve, but the lack of physical consistency between fields and dipoles might lead to reduced accuracy. On the other hand, even the full, self-consistent point-polarizability model is itself a simplified representation of relatively complex electron population shifts, and going from self-consistency to the direct approximation may not add much more error. Another way to avoid solving the self-consistent induced dipoles problem is to use the Drude oscillator[43, 98, 9], or charge-on-spring[192], model. This approximates atom-centered point polarizabilities by attaching an artificial particle, with a small mass and a point charge, to each atom treated as polarizable, and including the motions of these charges as part of the dynamical system. It is also important to mention more detailed models, such as

SIBFA[58], which place anisotropic polarizabilities off atom centers; and treatments of electronic polarization based on a continuum dielectric representation [154, 168, 173].

Another functional form used to model electronic polarization does not use point dipoles, but instead allows the partial atomic charges to vary in response to time-varying electric fields. For example, the fluctuating charge implementation of this approach [135, 149] uses an electronegativity equalization [116] ansatz to define how charges vary with field and treats the changes in charge as additional dynamical variables via an extended Lagrangian method. However, other polarization implementations based on variable partial charges would also possible[48].

The applicability of these two basic models, inducible point dipoles and variable point charges, raises the question whether one is fundamentally better suited than the other to model shifts in the electron density of a molecule induced by external fields. That is, setting aside how these quantities are assigned in any given implementation, is there a difference in the ability of atom-centered point charges versus point dipoles to model the changes in electrical field due to induced polarization? This question originally arose in discussions about aqueous nitrobenzene: perhaps a configuration with a water molecule hydrogen-bonded to only one nitro oxygen would lead to a redistribution of electrons between the two oxygens that would not be readily captured by atom-centered dipoles (William Swope, personal discussion). On the other hand, it is well known that atom-centered point charges are not well suited to capture out-of-plane polarization of a planar molecule, such as nitrobenzene. Additionally, given the greater computational cost of the self-consistent induced dipole model versus the direct approximation, it is worth asking how much accuracy is lost in going to the direct model. Finally, when formulating a polarization model based on atom-centered polarizabilities, one must choose between simply overlaying

the new polarizabilities on an existing set of fixed partial atomic charges (e.g., RESP charges[14]), or optimizing a new set of charges for use with the polarizability model.

Here, we address these and related issues by computing changes in molecular electrostatic potentials induced by point charges at multiple locations around small molecules in vacuum, using electronic structure methods; and then testing the how well optimized implementations of various polarizability models replicate these changes. The results have implications for the formulation of force fields that account for configuration-dependent electronic polarization.

3.2 Methods

The basic approach taken here is to assess how well various polarization models can replicate electrostatic potentials (ESPs), computed by quantum mechanical (QM) methods, around molecules polarized by artificial inducing charges. Our use of polarized QM ESPs as a reference for polarizability models follows the work of others[85, 9]. We regard each polarization model as consisting of a *representation* of polarization in terms either of shifts in atom-centered point charges or of added atom-centered point dipoles; and a *response model*, which is a recipe for computing these charges or dipoles. For example, fluctuating charge[135, 149] is a model which represents polarization via changes in point charges, and which uses an electronegativity equalization response model to derive these charges[116]. Similarly, the self-consistent and direct polarization models both represent polarization in terms of atom-centered point dipoles, but they use somewhat different response models to compute the induced dipoles. We studied the following polarization models:

Model 0: Global optimized point charges A single charge set is optimized to best repli-

cate the full set of polarized ESPs. Chemically equivalent atoms are constrained to have equal charges. It should be emphasized that the charges in this model are constant across all locations of the inducing charge. In addition, there is no response model for computing the charges beyond fitting to QM results, so this model could not be used in an actual simulation. However, this model is informative, as it reveals the greatest accuracy attainable by using a single set of permanent point-charge to capture the various polarized states of a molecule.

Model 1: Optimal point charges Polarization is represented by changes in atom-centered point charges, and, unlike Model 0, a different charge set is optimized to best replicate each polarized ESP. There is no response model for computing the charges beyond fitting to QM results, so this model could not be used in an actual simulation. However, this model demonstrates the greatest accuracy attainable with any model using the point-charge representation of polarization.

Model 2: RESP charges and optimal point dipoles Restrained ESP (RESP) charges[14] are assigned based on the unpolarized ESP and then held constant, while a different set of atom-centered point dipoles is optimized to best replicate each polarized ESP. As for optimal point charges (above), there is no response model for computing the dipoles beyond fitting to QM results, so this model could not be used in a simulation. However, it reveals the greatest accuracy attainable by the point-dipole representation of polarizability, in the context of the baseline RESP charges.

Model 3: RESP charges and direct polarizabilities RESP charges are assigned based on the unpolarized ESP and then held constant; then a single set of atom-centered point polarizabilities, modeled as not interacting with each other, is adjusted for each

molecule, so as to allow optimal replication of the full set of polarized QM ESPs. Chemically equivalent atoms are constrained to have equal polarizabilities.

Model 4: Co-optimized charges and direct polarizabilities This is the same as Model 3, except that a single, fixed set of atomic charges is optimized, along with the point polarizabilities, to best replicate the full set of polarized ESPs. Thus, the final set of point charges differs from the RESP charges, as it is chosen to best model not only the baseline unpolarized potential, but also all induced potentials, in conjunction with the inducible dipoles. Chemically equivalent atoms are constrained to have equal polarizabilities and charges.

Model 5: RESP charges and self-consistent polarizabilities Same as Model 3, except that the induced dipoles interact with each other, and their induced dipole moments are solved self-consistently.

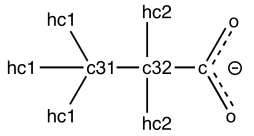
Model 6: Co-optimized charges and self-consistent polarizabilities Same as Model 5, except a single, optimal set of point charges is obtained along with the optimal point polarizabilities.

For completeness, we also examined how much the polarized ESPs deviated from ESPs computed with standard RESP charges optimized to the baseline, unpolarized potentials. On the other hand, we did not examine the Drude model, because it is essentially an implementation of the self-consistent, atom-centered, point-polarizability model, which is examined here (Models 4,6). In addition, due to its dynamical nature, the Drude model does not appear to provide an unambiguous mapping between a molecular conformation and a set of induced dipoles.

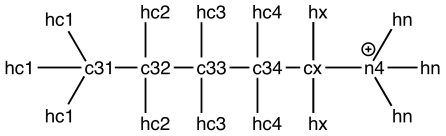
3.2.1 Molecular Structures and Atom Types

In order to examine polarization in the context of biologically relevant molecules, we studied nonpolar, neutral polar, and ionized amino acid side chain analogs (Figure 3.1). The molecule nitrobenzene was also included, to test the ability of point dipoles to model any charge redistribution between the two nitro oxygens when an inducing charge is near one of them, and also to probe how well the two polarization representations handle out-of-plane induction. In addition, gas-phase dimers of the valine analog and of methane were studied. These are informative because they increase the number of dipole-dipole interactions among point polarizabilities, relative to the small number of such interactions present in the monomers. They are therefore useful to further probe the accuracy of the direct *versus* self-consistent polarizability models. Multiple conformations were included for the aspartate, tyrosine and valine analogs. The coordinates of all molecular systems are energy-minimized structures drawn from BEGDB[17].

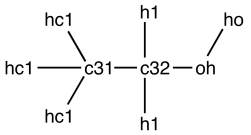
In the point-polarizability models, whose parameters are optimized across all polarized ESPs, chemically equivalent atoms are assigned identical polarizabilities and atomic charges. In particular, using the fact that the RESP software available through the Antechamber program [181] automatically forces chemically equivalent atoms to have identical RESP charges, we assigned equivalent parameters to groups of atoms with identical RESP charges. To do this, we assigned a type to each group of chemically equivalent atoms in a molecule; for example, the valine side-chain analog has two carbon types and two hydrogen types (Figure 3.1g). Note that these types do not necessarily correspond to force field atom types, and that no equivalence was enforced between molecules. For example, *c31* can have different parameters in each molecule where it occurs.



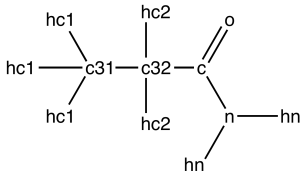
(a) Aspartate(-)



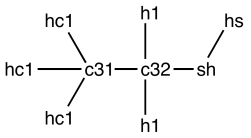
(b) Lysine(+)



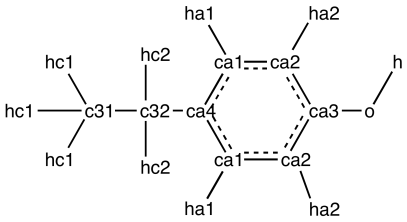
(c) Serine



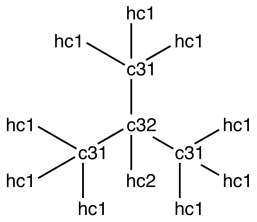
(d) Asparagine



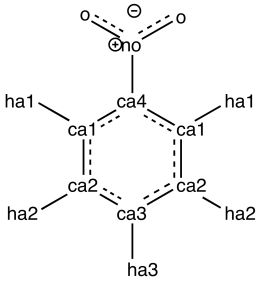
(e) Cysteine



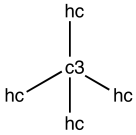
(f) Tyrosine



(g) Valine



(h) Nitrobenzene



(i) Methane

Figure 3.1: Molecular structures with atom types. Note that 3.1a-3.1g are amino acid side chain analogs rather than the full amino acids, and that atom types are molecule-specific.

3.2.2 Calculations of Reference QM Electrostatic Potentials

For each molecule or dimer, gas-phase QM ESPs were computed at the HF/6-31G* level with Gaussian 09 (RevD.01)[52]. In addition to baseline ESPs, which represent unpolarized states, polarized ESPs were generated by solving the wave equation with an inducing point charge, $\pm 1.0e$, at locations around the molecule determined with the dot molecular surface program dms included with MIDAS[49]. In order to position the inducing charges roughly one heavy-atom diameter away from the atom-centers, the surfaces were generated with all default atom radii incremented by 1.9Å. The default probe radius of 1.4Å was used, and the point density was set to 0.1 points per Å^2, leading to roughly 150 points per molecule. The resulting baseline and polarized ESPs were treated as reference data to optimize the various polarization models, as detailed below.

3.2.3 Description of Systems and Characterization of Errors

We consider a molecule with N atoms, so that, for $i = 1, \ldots, N$, q_i are atom-centered point charges, μ_i are atom-centered point dipoles, and α_i are atom-centered polarizabilities, where the atom centers are at locations r_i. The molecule has $T \leq N$ atom types, so that for $t_i \in [1, \ldots, T]$, q_{t_i} are atom-typed point charges and α_{t_i} are atom-typed polarizabilities. The ESP values are computed at M locations r_m, $m = 1, \ldots, M$, where M is determined by Gaussian 09 and varies from one molecule to another. The unpolarized reference quantum ESP values are obtained in the absence of an inducing charge, while polarized reference quantum ESP values are each computed in the presence of an inducing charge $q_k = \pm 1$ at the dms-assigned location r_k. The inducing charge locations are repeated for both positive and negative point charges so that the reference set consists of a total of $2K$ polarized QM ESPs. At an ESP point r_m, the unpolarized reference potential is ϕ^0_{m0}, while the QM

potential computed in the presence of inducing charge $q_k = 1$ is ϕ^0_{mk} for $k = 1,\ldots,K$, and the potential computed in the presence of inducing charge $q_k = -1$ is ϕ^0_{mk} for $k = K+1,\ldots,2K$. The corresponding potentials from a given polarization model are ϕ_{mk}, for $k = 0,\ldots 2K$. Note that all ESP values considered here omit the direct Coulombic contribution from the inducing charges.

Optimization of a polarization model means minimizing errors computed in terms of sums of the squared potential differences $(\phi_{mk} - \phi^0_{mk})^2$. We report the overall error of any model, when applied to a given molecule in the presence of inducing charges in positions corresponding to $k = 1,\ldots,K$ as the root-mean-square error of the potential (kcal/mol-e) across all ESP points for both the positive ($k \in [1,K]$) and negative ($k \in [K+1,2K]$) inducing charges located at r_k:

$$R_k = \left(\frac{\chi^2_k + \chi^2_{k+K}}{2M}\right)^{\frac{1}{2}} \tag{3.1}$$

$$\chi^2_k = \sum_{m=1}^{M} (\phi_{mk} - \phi^0_{mk})^2 \tag{3.2}$$

The error across all inducing charges for a given molecule can then be written as

$$R = \left(\frac{1}{K}\sum_{k=1}^{K} R^2_k\right)^{\frac{1}{2}} \tag{3.3}$$

3.2.4 Implementation of Models

The following subsections detail the calculation of RESP charges for the baseline, unpolarized potential, and the implementation of each polarization model listed above.

Baseline RESP charges

Atom-centered point charges were fitted to best replicate the baseline, unpolarized QM ESP of each molecule, ϕ^0_{0m}, using RESP as implemented in the Antechamber program [181]. The two-stage fitting process uses the default AMBER force field charge restraint weights of 0.0005 for the first stage and 0.001 for the second.

Model 0: Global Optimal Point Charges

For each molecule, atom-centered point charges were optimized for the best simultaneous fit to all polarized electrostatic potentials. The electrostatic potential at ESP site r_m, due to the full set of atomic charges, $q_i, i = 1,\ldots,N$, is calculated by

$$\phi_{mk} = \sum_i^N \frac{q_i}{|r_{im}|} \tag{3.4}$$

where well-known physical constants are omitted for simplicity, and

$$r_{ij} \equiv r_i - r_j. \tag{3.5}$$

The indices i and j are used for atomic centers, while m refers to the ESP locations. Note that the same set of atomic charges is used for every inducing charge k, so that the potential ϕ_{mk} is actually independent of k. The optimization procedure, which minimizes the quantity R in Equation 3.3, is described in the Optimization of Parameters section.

Model 1: Optimal Point Charges

Atom-centered point charges were fitted separately to each polarized QM ESP using the RESP software and the same set of ESP points as for the baseline, unpolarized, potential,

but with the restraint weights set to zero and with atom equivalency disabled to allow free optimization of point charges. Thus, this model reports on the maximal accuracy attainable by any polarization model that uses the atom-centered point-charge representation of polarization. However, it cannot be employed in simulations, because it requires a new quantum calculation for each molecular configuration.

Model 2: RESP Charges and Optimal Point Dipoles

Atom-centered point charges were first fitted to the baseline, unpolarized ESP using standard RESP, then atom-centered point dipoles, superimposed on the point charges, were optimized separately for each polarized QM ESP. This model reports on the maximal accuracy attainable by any polarization model that uses the atom-centered point dipole representation of polarization along with RESP baseline point charges. However, like the optimal point charge model above, it cannot be employed in simulations, because it requires a new quantum calculation for each molecular configuration. The procedure for computing optimal point dipoles is now described.

The electrostatic potential at r_m due to the baseline RESP partial charges, q_i, and the atom-centered point dipoles, μ_i, is given by

$$\phi_m = \sum_i \frac{q_i}{|r_{im}|} + \sum_i \mu_i A_{im} \tag{3.6}$$

where well-known physical constants are omitted for simplicity, r_{ij} is as in Equation 3.5 and

$$A_{ij} \equiv \frac{r_{ij}}{|r_{ij}^3|} \tag{3.7}$$

For each external charge position r_k, where index k refers to the inducing charge, the error

metric χ_k^2 (Equation 3.2) may be written as

$$\chi_k^2 = \sum_m \left(\phi'_{mk} - \sum_i \mu_i A_{im} \right)^2 \tag{3.8}$$
$$\phi'_{mk} = \phi^0_{mk} - \sum_i \frac{q_i}{|r_{im}|}$$

where ϕ'_{mk} contains all quantities independent of the dipoles.

Optimal dipoles are arrived at by setting $\frac{\partial(\chi_k^2)}{\partial \mu_i} = 0$ for all μ_i. This yields the following system of linear equations:

$$\begin{bmatrix} \sum_m A_{1m}\phi'_m \\ \vdots \\ \sum_m A_{Nm}\phi'_m \end{bmatrix}_k = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix}_k \begin{bmatrix} \sum_m A_{1m}^2 & \cdots & \sum_m A_{1m}A_{Nm} \\ \vdots & \ddots & \vdots \\ \sum_m A_{Nm}A_{1m} & \cdots & \sum_m A_{Nm}^2 \end{bmatrix} \tag{3.9}$$

Solving this matrix equation yields the desired atom-centered point dipoles optimized for external charge position r_k. The NumPy *linalg.norm*[175] function was used to find the solution to the matrix equation.

Model 3: RESP Charges and Direct Polarizabilities

First, atom-centered point charges were fitted to the baseline, unpolarized ESP, using standard RESP. Then a single set of atom-typed, atom-centered, isotropic polarizabilities was optimized for best simultaneous fit to all polarized ESPs ($k = 1 \ldots 2K$), where the model ESPs are computed as sums of the potentials from the baseline RESP charges and the potentials from the induced dipoles. In this first-order, or direct, polarizability model, the field at each atom i, E_i, is that due to the inducing charge, $E_i^{q_k}$, plus that due to the fixed

RESP charges, E_i^q; fields due to induced dipoles are ignored. Thus

$$E_i = E_i^{q_k} + E_i^q = q_k A_{ik} + \sum_{j \neq i}^{N} S_{ij} q_j A_{ij} \tag{3.10}$$

Here S_{ij} is a screening function which excludes 1-2 and 1-3 interactions and scales 1-4 interactions by a factor of 0.5, in keeping with common practice in computing Coulombic interactions with empirical force fields. The induced dipoles are proportional to the inducing field E_i.

$$\mu_i = \alpha_{t_i} E_i = \alpha_{t_i} [E_i^{q_k} + E_i^q] \tag{3.11}$$

the atom-typed, point polarizabilities, α_{t_i}, are parameters which must be adjusted to minimize the mean squared error, R^2, across all polarization states $k = 1, \ldots, 2K$; see Equation 3.3. The optimization procedure is described in the Optimization of Parameters section. Note that the set of charges and point polarizabilities derived in this way is applicable to all inducing charges and charge positions, as the inducible dipoles respond to the inducing field according to Equation 3.11. Thus, no additional QM calculations are needed, and this model could be used in a simulation.

Model 4: Co-optimized Charges and Direct Polarizabilities

This model is identical to Model 3, except that a single set of atom-centered, atom-typed, point charges is co-optimized with the point polarizabilities across all polarization states, to minimize the global mean squared error R^2. Again, this yields a model which could be used in a simulation, because no additional QM calculations are needed.

Model 5: RESP Charges and Self-consistent Polarizabilities

This model is the same as Model 3, except that the field at each atom i now includes a contributions from other induced dipoles, E_i^{μ}, and the system of induced dipoles is solved self-consistently. For a given inducing charge k, the total electric field at atom i becomes

$$E_i = E_i^{q_k} + E_i^q + E_i^{\mu} = q_k A_{ik} + \sum_{j \neq i}^{N} S_{ij} q_j A_{ij} + \sum_{j \neq i}^{N} S_{ij} \mu_j B_{ij} \tag{3.12}$$

(Because the RESP charges are atom-typed, one could properly replace q_j by q_{t_j}.) We use the Applequist dipole interaction model [10], for which B_{ij} depends only on separation r_{ij}

$$B_{ij}^{\beta\gamma} = \frac{3 r_i^{\beta} r_j^{\gamma}}{r_{ij}^5} - \frac{\delta_{\beta\gamma}}{r_{ij}^3} \tag{3.13}$$

where $\beta, \gamma \in \{x, y, z\}$.

The induced dipoles are, again, computed from these fields and the atom-typed point polarizabilities:

$$\mu_i = \alpha_i E_i = \alpha_{t_i} [E_i^{q_k} + E_i^q + E_i^{\mu}] \tag{3.14}$$

These equations can be rewritten in matrix form as

$$\begin{bmatrix} E_1^{q_k} + E_1^q \\ \vdots \\ \vdots \\ E_N^{q_k} + E_N^q \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \vdots \\ \mu_N \end{bmatrix} \begin{bmatrix} \alpha'_{t_1} & -B_{12} & \dots & -B_{1N} \\ -B_{12} & \alpha'_{t_2} & \dots & -B_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ -B_{N1} & -B_{N2} & \dots & \alpha'_{t_N} \end{bmatrix} \tag{3.15}$$

where $\alpha'_{t_i} = \alpha_{t_i}^{-1} I$ and B_{ij} are 3x3 matrices. Solving this matrix equation yields a set of self-consistent induced dipoles μ_i, particular to a set of point charges q_i, polarizabilities α_{t_i},

and inducing charge q_k at r_k.

Using baseline, atom-typed RESP charges for q_{t_i}, the atom-typed, isotropic polarizabilities α_{t_i} are solved for by the same optimization process as used the direct polarization models (see Optimization of Parameters section). The resulting polarization model is particular to the set of baseline RESP charges, but applicable to all external charge locations r_k. Again, this is a model which could be used in a simulation, because no further QM calculations are required once the polarizabilities are established.

The treatment of close-ranged dipole-dipole interactions adopted here deserves comment. As noted above, we exclude 1-2 and 1-3 interactions, and scale 1-4 interactions by 0.5. This approach is consistent with many existing polarizable force fields, especially those also utilizing the Applequist model, such as AMBER ff02[30, 185]. Some dipole-dipole screening models, such as that of Thole [171] and its variations [139], also modify the B_{ij} term (Equation 3.13) to prevent the polarization catastrophe to which the Applequist model is susceptible. However, excluding 1,2 and 1,3 short-range intramolecular interactions and halving 1,4 interactions, as done here, prevents this problem, even with the Applequist model, and our test calculations demonstrated that adding Thole or Thole-like screening terms to the present models produced negligible changes in the results (data not shown). Indeed, Cieplak et al.[31] note that other screening models deviate significantly from the Applequist model only at ranges shorter than about 3.0Å; but the 1-2 and 1-3 exclusions already eliminate many interactions within this range so the screening models become virtually equivalent. This observation is corroborated by the similarity of fitting results and parameters in Wang et al[180] for the CL, CE, CT and CA models, where C refers to the same short-range scaling as detailed above and the second letter refers to added screening functions tested.

Model 6: Co-optimized Charges and Self-Consistent Polarizabilities

This model is the same as Model 5, except that a fixed set of atom-typed partial charges is co-optimized with the polarizabilities against the full set of polarized ESPs. Thus, the atom-typed parameters q_{t_i} and α_{t_i} are simultaneously adjusted, using the same optimization process and error function as used for the other inducible dipole models. Again, this model could be used in a simulation, as no additional QM calculations are needed.

3.2.5 Optimization of Parameters

All models require global parameter optimization. For Model 0 and Models 3-6, parameters were optimized to minimize R^2, the mean of the squared potential deviations across all inducing charge sites k, for each molecule of interest (Equation 3.3). For Models 1 and 2, parameters were optimized to minimize R_k^2, the mean squared potential deviations for each separate inducing charge position (Equation 3.1). All optimizations were performed with a SciPy implementation of L-BFGS-B[197, 81], a gradient-based constrained minimization method. Charges are left unconstrained, while polarizabilities are restricted to positive values. For the inducible dipole models with fixed RESP charges (Models 3 and 5), only the atom-typed polarizabilities,α_{t_i}, require adjustment; for those with co-optimized point charges, the atom-typed charges, q_{t_i}, are adjusted along with the polarizabilities.

For each molecule, multiple optimizations were run with initial parameter values drawn from a uniform distribution using *numpy.random.rand()*[175]. For Model 0, 5 optimizations were run using initial charges randomly drawn from the range -1.0e to 1.0e. For Models 3-6, 50 optimizations were run using initial polarizabilities randomly drawn from between 0 to 10 bohrs^3 (0 to 1.482 Å^3). The parameter set with the lowest value of R^2 was selected as the optimum. When charges were co-optimized (Models 4 and 6), the

baseline RESP charges were used as their starting values. Only 10 optimizations were run for tyrosine, as the calculations became time-consuming for this relatively large molecule.

3.2.6 Calculation of Isotropic Molecular Polarizabilities

To check the plausibility of the optimized atomic polarizabilities for the inducible dipole models, we computed the corresponding molecular polarizabilities and compared them with molecular polarizabilities for the same compounds computed with the Gaussian 09 software. The isotropic molecular polarizability of each molecule was calculated from the optimized isotropic atomic polarizabilities as

$$\alpha_{mol} = \sum_i \alpha_{t_i} \tag{3.16}$$

The molecular polarizabilities from Gaussian 09 were calculated at the HF/6-31G* level, for consistency with the ESP calculations used to derive the atomic polarizabilities.

3.3 Results and Discussion

This section reports on the accuracy of the various representations and models of polarization, and then appraises the charges and polarizabilities obtained by optimizing these parameters for the four inducible dipole models.

3.3.1 Accuracy of Polarization Models

Overall Accuracy

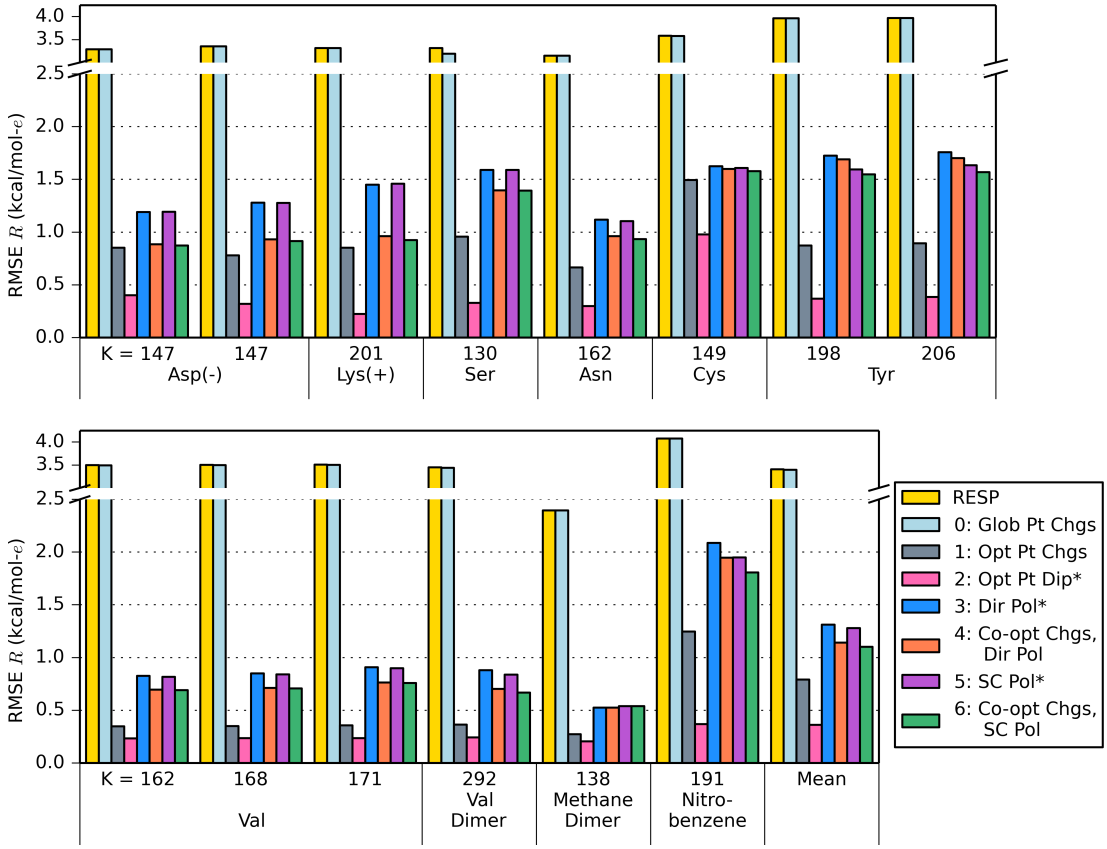


Figure 3.2: RMSE (R) of polarization models (kcal/mol-e). An asterisk in the legend indicates that RESP charges were used.

The errors of the models are reported as R (Equation 3.3), which is the RMSE of the ESPs approximated by each model for all inducing charges. Note that these errors will vary with the magnitudes of the inducing fields, with larger fields generating larger errors. Here, the inducing fields were generated by unit monopoles placed roughly one atomic diameter away from the nearest atom-center of the molecule, *in vacuo*. Thus, roughly similar errors might be anticipated in simulations of systems with univalent ions near organic compounds in a low dielectric environment, like a lipid membrane; the overall error in such a system

will then accumulate and/or cancel across a multitude of complex interatomic interactions.

In order to establish a baseline for comparing the various models, it is of interest to summarize the errors incurred by *not* using any polarizable model at all. This baseline R measures the RMSE of the approximate ESPs computed using standard RESP point charges, which were fit to the unpolarized QM ESP, for all the inducing charges. As shown in Figure 3.2 (yellow columns), these errors range from about 3-4 kcal/mol-e. No significant reduction in error is obtained by using a single set of point charges simultaneously optimized to the ESPs associated with all inducing charges (Model 0), as evident in Figure 3.2. In fact, the values of R from Model 0 are the same as those from plain RESP to within 0.01 kcal/mol-e. The validity of this result was confirmed by running five different optimizations of the Model 0 charges from different randomized starting values; the standard deviation of the optimized atomic charges for each atom across the five runs was within 0.0026e. Interestingly, the Model 0 charges strongly resemble the RESP charges: a linear regression of Model 0 versus RESP charges across all molecules gives a slope of 0.99, y-intercept of 0.0017 and Pearson correlation coefficient of 0.997. Due to the similarity of the Model 0 and RESP charges and errors, subsequent references to the results when polarization is neglected may be considered to reference either the baseline RESP results or Model 0.

As detailed in Methods, we tested two representations of polarization tuned to fit each individual polarized QM ESP: adjustable atom-centered point charges (Model 1), and adjustable atom-centered point dipoles (Model 2). Both representations lead to errors in the induced ESPs well below the 3-4 kcal/mol-e errors obtained when polarization is neglected (prior paragraph). Optimal point charges reduce the error, on average, to about 0.75 kcal/mol-e (gray columns, Figure 3.2), while optimal point dipoles lead to even lower average errors of about 0.4 kcal/mol-e (magenta columns, Figure 3.2). The advantage

of point dipoles over point charges is consistent across all molecules studied, though the difference varies from case to case and is lowest for methane and for the simple alkane model of valine. It should be emphasized, however, that these results bear only on the ability of adjustable point charges and point dipoles to capture induced changes in molecular ESPs.

We investigated whether Model 2 provides greater accuracy than Model 1 less because it offers a better description of the polarization as that it provides a more accurate description of the baseline, unpolarized electrostatic field. We used nitrobenzene as a test case to check for this possibility, creating a model with RESP charges supplemented by point dipoles optimized to replicate the baseline, unpolarized ESP of this molecule. These baseline optimized dipoles, μ_i^0, may be viewed as optimized permanent dipoles for the unpolarized molecule, and they help correct for the errors of RESP charges in replicating the unpolarized quantum mechanical ESP. We then computed the RMSE of the ESP generated by this permanent-charge plus permanent-dipole model against the full set of induced ESPs for all 382 inducing charges ($\pm 1.0e$ at each of 191 locations). The resulting RMSE of 4.06 kcal/mol-e is essentially the same as that for baseline RESP charges alone, 4.08 kcal/mol-e. Thus, adding permanent dipoles to the baseline RESP charges produces minimal improvement in the ESP fits across the set of inducing charges. In contrast, the RMSE is 0.37 kcal/mol-e for Model 2, in which a new set of point dipoles is optimized for each inducing charge. Thus, the plower errors observed for Model 2 versus baseline RESP charges are attributable to the improved description of induced polarization, rather than to an improved description of the baseline potential.

The four polarizability models based on inducible, atom-centered point dipoles (Models 3-6), were then examined. In each case, a single set of polarizabilities and, where applicable, point charges, was optimized to best replicate the full set of QM polarized ESPs

for each molecule. Figure 3.2 summarizes the accuracy of the resulting parameterized models.

Models 3 and 4 both use the efficient direct polarization approximation, but, whereas Model 3 keeps the baseline RESP point charges (first paragraph of Results), Model 4 uses a new set of point charges optimized along with the polarizabilities. Both of these models yield errors (R) averaging about 1.25 kcal/mol-e (blue and orange columns, respectively, Figure 3.2). This is substantially better than the baseline errors of 3-4 kcal/mol-e associated with unpolarized RESP charges (above), but about threefold worse than the theoretical optimum of about 0.4 kcal/mol-e for the point dipole representation (Model 2). It is also nearly twofold worse than the theoretical optimum of about 0.75 kcal/mol-e achievable with the variable point-charge model (Model 1).

One possible source of error in Models 3 and 4 is their nonphysical neglect of interactions among the induced dipoles. However, Models 5 and 6, which are the same, respectively, except that they include these interactions and thus require solving a matrix equation, yield at best marginal improvements in accuracy across all cases tested (purple and green columns, Figure 3.2); the differences in R, are all <0.15 kcal/mol-e. A possible concern regarding this observation is that the exclusion of 1-2 and 1-3 interactions, and the scaling of 1-4 interactions, eliminates or weakens many dipole-dipole interactions that would otherwise have been fully included in the self-consistent calculations; thus, the direct model may not be very different from the self-consistent model. We addressed this concern by including two dimers (valine analog and methane), reasoning that no intermolecular dipole-dipole interactions are excluded, so these cases should better probe the differences between the direct and self-consistent approaches. Nonetheless, the self-consistent models do not particularly outperform the direct models for the dimers either (Figure 3.2); in fact,

they are marginally worse than the direct models for the methane dimer.

Note that the highest-level polarizability model tested here, Model 6, which uses self-consistent induced dipoles and co-optimized point charges, does not reach the maximal accuracy achievable for the point-charge representation of polarization (Model 1). This means that a sufficiently well-designed point-charge model of polarizability could outperform Model 6, which is the standard self-consistent induced dipole model. Given the failure of Model 6 to approach the accuracy of the theoretical optimal point-dipole model (Model 2), there is even more room to improve the accuracy of polarization models based on the point-dipole representation.

We examined whether the greater accuracy of Models 1 and 2, relative to the linear response models, Models 3-6, perhaps resulted from their ability to account for nonlinearity in the quantum mechanical polarization response. We again used nitrobenzene as a test case, rerunning quantum calculations with charges of ± 0.2, ± 0.4, ± 0.6, and ± 0.8 e at each of the 191 inducing charge locations, and testing for linearity by several criteria. First, for each inducing charge location, we ran a linear regression of the molecular dipole moment provided by Gaussian against the value of the inducing charge: the Pearson correlation coefficients were found to be at least 0.9999 for all locations of the inducing charge, confirming the linearity of the overall polarization response. Second, for each inducing charge location, we carried out linear regressions of the magnitudes of the optimized (Model 2) point dipoles at each atom against the magnitude of the inducing charges. (Since nitrobenzene has 14 atoms, there are 382 x 14 regression fits.) We averaged the Pearson correlation coefficients for each atom across charge locations to provide summary statistics, and the lowest mean correlation coefficient is found to be 0.9999, again indicating a linear response. The corresponding analysis for the optimal point charge model, Model 1, similarly

yielded correlation coefficients $\geq$0.9990. Finally, we tested how well the accuracy of a linear extrapolation of the atom-centered point dipoles fitted to inducing charges of $\pm 0.2e$ could replicate quantum ESPs induced with a charge of $\pm 1.0e$. For each inducing charge location (not indexed, for simplicity) and atom (i), the dipole moment for a unit inducing charge is extrapolated as follows $\mu_i^{1.0} \approx 5\left(\mu_i^{0.2} - \mu_i^0\right) + \mu_i^0$, where μ indicates a dipole moment, and the superscripts indicate the value of the inducing charges to which they pertain. (The expression for a negative inducing charge is analogous.) The ESPs from these extrapolated dipoles were compared with the quantum ESPs for the corresponding inducing charge locations but obtained with inducing unit charges. The RMSE across all inducing charge positions is 0.372 kcal/mol-e, which is essentially the same as the Model 2 result obtained by optimizing dipoles for inducing charges of unit magnitude, 0.369 kcal/mol-e. The analogous extrapolation from inducing charges of 0.2 for the optimal point charge model, Model 1, yields an RMSE of 1.247 kcal/mol-e, which is essentially the same as the error for Model 1 charges optimized directly against ESPs for inducing charges of unit magnitude, 1.246 kcal/mol-e. Thus, the greater accuracy of Models 1 and 2, relative to the linear polarization models, Models 3-6, is not related to their ability to capture a nonlinear polarization response.

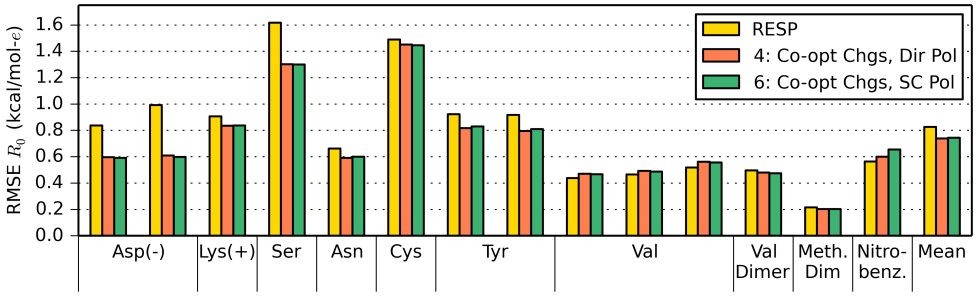


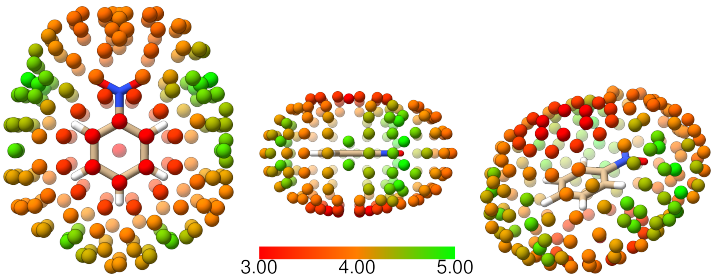
Figure 3.3: RMSE (R_0) of polarization models for unpolarized states (kcal/mol-e).

Another dimension of the inducible dipole models studied here is their use of either

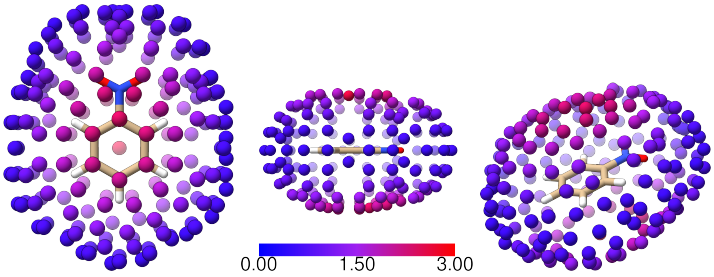
baseline unpolarized RESP charges, or a new set of point-charges optimized along with the polarizabilities to best replicate the QM polarized ESPs. The use of co-optimized charges yields somewhat greater accuracy in all cases; compare Model 4 with Model 3, and Model 6 with Model 5 (Figure 3.2). The improvements are greatest (0.3 to 0.5 kcal/mol-e) for the ionized systems, and least for the cysteine analog ($\sim$0.03 kcal/mol-e). Importantly, when Models 4 and 6, with co-optimized charges and either direct or self-consistent polarizabilities, are used to compute the baseline, unpolarized ESP, the agreement is somewhat better, on average, than that provided by standard (unpolarized) RESP charges (Figure 3.3). Thus, optimization of charges and polarizabilities against sets of polarized ESPs results in parameters that are also applicable to the unpolarized state.

Accuracy at Each Inducing Charge Location

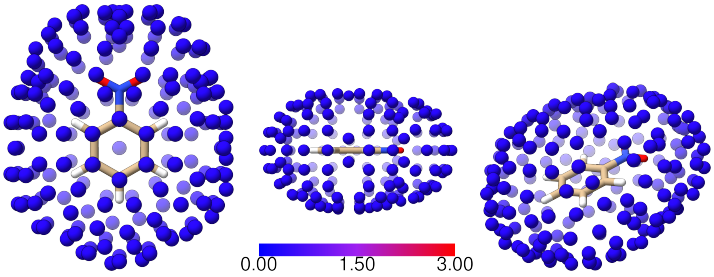
Further insight into the strengths and weaknesses of the various polarization models can be obtained by visualizing the errors associated with different positions, r_k of the inducing charge. Figures 3.4 and 3.5 depict nitrobenzene and the valine analog, respectively, each surrounded by colored spheres at representative positions, r_k of the inducing charge. Each sphere is colored according to the overall error R_k (Equation 3.1) associated with the inducing charge at the location of the sphere. Results are presented for the baseline RESP charges, optimal point charges, optimal point dipoles, and co-optimized charges and self-consistent polarization. We used different color scales for different panels, in order to bring out the geometric variations associated with each model and molecule.



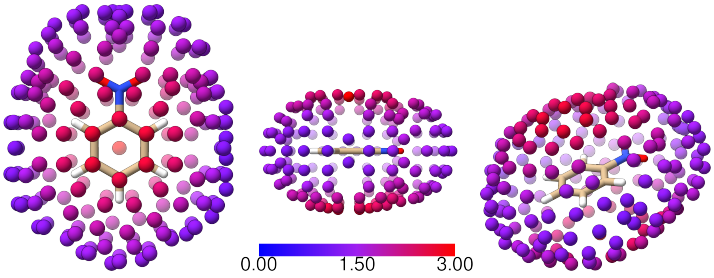
(a) Baseline RESP



(b) Optimal point charges

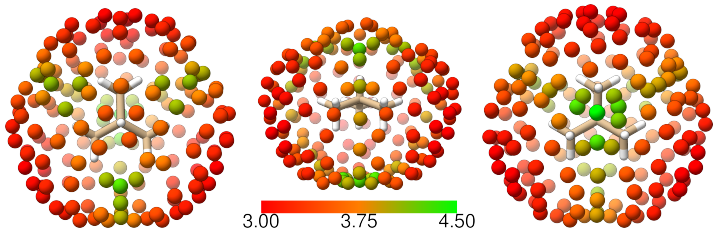


(c) Optimal point dipoles

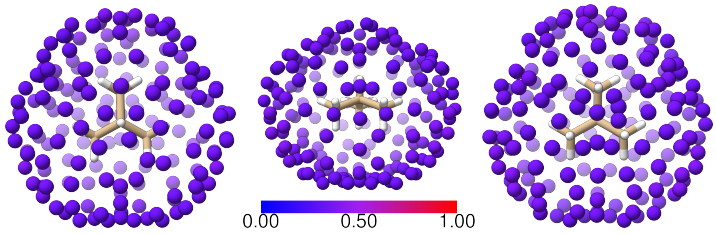


(d) Co-optimized charges and self-consistent polarizabilities

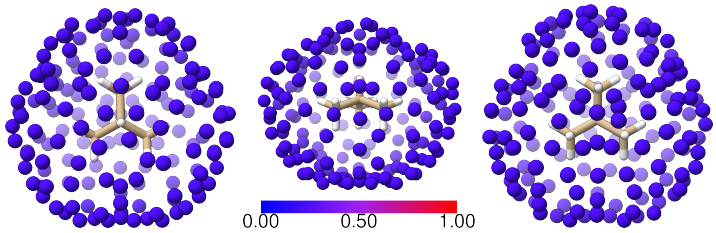
Figure 3.4: RMSE (R_k) for nitrobenzene (kcal/mol-e), as a function of inducing charge location r_k.



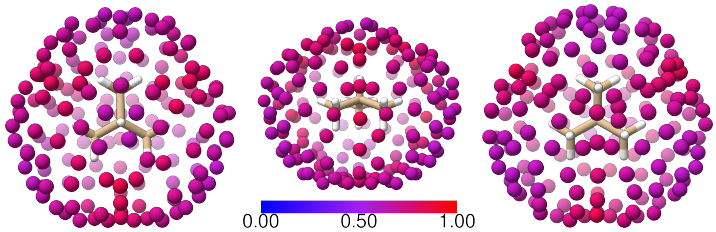
(a) Baseline RESP



(b) Optimal point charges



(c) Optimal point dipoles



(d) Co-optimized charges and self-consistent polarizabilities

Figure 3.5: RMSE (R_k) for valine analog (kcal/mol-e), as a function of inducing charge location r_k.

Not surprisingly (see Figure 3.2), baseline RESP charges do a poor job of replicating the induced potentials. For nitrobenzene, the errors are less for out-of-plane than in-plane inducing charges (Figure 3.4a), while for the valine analog, the errors are least for inducing

charges near the axes of the carbon-carbon bonds (Figure 3.5a). All of the polarization models yield lower errors than the baseline RESP at all external locations, so accounting for polarization is consistently advantageous. The relative ranking of the methods is also consistent: optimal point dipoles (Model 2) yield the lowest errors, followed by optimal point charges (Model 1), and then co-optimized charges with self-consistent polarizabilities (Model 6). In the case of valine, which is nonplanar: the errors for all three models vary little with the position of the inducing charge, as the values of R_k remain within a 0.7 kcal/mol-e range for each model. However, the pattern is more complex for the planar molecule nitrobenzene. Here, the errors are greater for out-of-plane inducing charges than for in-plane ones, for both Models 1 and 6 (Figures 3.4b,d). (A similar pattern is seen for the aromatic ring of the tyrosine analog; data not shown.) It is necessarily the case that the point-charge representation of polarization in Model 1 is unable to account for out-of-plane polarization for a planar molecule. It is harder to explain why Model 6 performs worse for out-of-plane than in-plane charges, and indeed, worse for out-of-plane charges than Model 1: compare Figures 3.4b and3.4d.

Optimal point dipoles (Model 2) can yield out-of-plane induced moments, and they afford good accuracy for inducing charges at in-plane and out-of-plane positions around nitrobenzene, as well as for all positions around the valine analog. Indeed, the errors for optimal point dipoles are uniform, to within 0.0035 kcal/mol-e, for inducing charges around both molecules. This result supports a view that the point dipole representation can provide consistent performance in modeling all polarization phenomena. However, optimal performance is clearly not attained by the customary inducible dipole models.

3.3.2 Optimized Polarizabilities and Charges

A basic check of the plausibility of the optimized polarizabilities is that they should be consistent with independently determined molecular polarizabilities. Isotropic molecular polarizabilities were computed from the atomic polarizabilities with Equation 3.16, and Table 3.1 compares these values with molecular polarizabilities computed independently with Gaussian 09 QM calculations at the same HF/6-31G* level used to compute the molecular ESPs, as well as with available experimental molecular polarizabilities. The molecular polarizabilities from the inducible dipole models agree well with the independently computed QM molecular polarizabilities, and hence with each other. However, the QM molecular polarizabilities underestimate the experimental results by on the order of 30%. This is consistent with a broader set of data showing that polarizabilities computed at the 6-31G* level underestimate experimental results, while more complete basis sets, such as aug-cc-pVTZ[87, 190], yield more accurate results[3].

All optimized parameters for the four inducible dipole models are listed in Table 3.2; the unpolarized RESP charges are also shown, as these are used in Models 3 and 5, and may be compared with the co-optimized charges of Models 4 and 6. The values of the optimized polarizabilities range from 0.000 to 3.643Å^3. The zeroes are, arguably, nonphysical, but we note that, in many of these cases, there appears to be compensation by neighboring atoms with particularly large polarizabilities, consistent with the fact that the overall molecular polarizabilities are physically reasonable (above). For example, in nitrobenzene, where the direct and self-consistent polarizabilities, optimized with RESP charges, are zero for the *no* atom (Figure 3.1), the polarizabilities of the immediately neighboring *ca4* atoms are relatively large, at 2.964 or 3.643Å^3, respectively. Similarly, while the *c31* atoms in the

Table 3.1: Isotropic molecular polarizabilities (Å^3), from Equation 3.16 using optimized parameters, QM calculations, and experiment. The experimental values listed for the dimers are merely twice the experimental monomer results.

	3. RESP, Dir Pol	4. Co-opt Chgs, Dir Pol	5. RESP, SC Pol	6. Co-opt Chgs, SC Pol	HF/6-31G*	Expt[4]
Asp(-)	4.806	5.050	4.813	5.082	5.294	
Asp(-)	4.820	5.048	4.833	5.087	5.306	
Lys(+)	7.258	8.004	7.313	8.249	8.589	
Ser	3.666	3.641	3.674	3.650	3.763	5.41
Asn	5.087	5.256	5.128	5.320	5.625	
Cys	5.073	5.113	5.128	5.178	5.414	7.41
Tyr	10.312	10.443	10.812	10.931	11.734	
Tyr	10.199	10.339	10.725	10.893	11.826	
Val	6.064	6.121	6.108	6.166	6.302	8.14
Val	6.065	6.123	6.110	6.169	6.315	"
Val	6.063	6.129	6.105	6.169	6.314	"
Val Dimer	11.966	12.034	12.241	12.285	12.829	16.28
Methane Dimer	3.571	3.571	3.609	3.609	3.618	5.186
Nitrobenzene	8.218	8.579	8.666	9.043	9.808	

valine analog are consistently assigned a polarizability of 0.000Å^3, the central *c32* atom type has a relatively large polarizability (>2Å^3). Cases like these could be avoided by adding further restraints during the optimization, such as by forcing all carbon atoms in the valine analog to have equal polarizabilities. However, adding restraints would presumably lessen the accuracy of the agreement of the models with the QM polarized ESPs, and the present minimally restrained optimizations have the merit of revealing the best possible performance of each model.

The optimized polarizabilities may also be analyzed by element. As shown in Table 3.3, carbon consistently emerges as most polarizable, followed by oxygen, and with nitrogen and hydrogen approximately tied for third place. (The solitary sulfur atom in the models studied is assigned polarizabilities of 1.7-1.9 Å^3, well above the mean of about 1.0 for carbon.) This ranking is in rough agreement with the polarizabilities assigned to these

Table 3.2: Optimized parameters for inducible dipole models; polarizabilities (α) in Å^3 and charges (q) in e.

		RESP	3. RESP, Dir Pol	4. Co-opt Chgs, Dir Pol		5. RESP, SC Pol	6. Co-opt Chgs, SC Pol	
		q	α	q	α	α	q	α
Asp(-)	c	0.819	0.127	0.864	0.304	0.000	0.868	0.251
	c31	-0.062	0.000	-0.145	0.446	0.000	-0.137	0.490
	c32	0.018	2.884	-0.267	2.058	2.938	-0.269	2.042
	hc1	-0.002	0.162	0.049	0.306	0.160	0.048	0.305
	hc2	-0.050	0.000	0.041	0.000	0.000	0.041	0.000
	o	-0.835	0.655	-0.841	0.661	0.698	-0.844	0.693
Asp(-)	c	0.808	0.186	0.884	0.240	0.116	0.889	0.094
	c31	-0.103	0.652	0.004	1.330	0.655	-0.004	1.356
	c32	0.213	2.695	-0.206	1.842	2.713	-0.203	1.854
	hc1	-0.008	0.000	0.001	0.081	0.000	0.004	0.085
	hc2	-0.110	0.000	0.007	0.000	0.000	0.007	0.000
	o	-0.837	0.643	-0.850	0.696	0.674	-0.853	0.764
Lys(+)	c31	-0.348	0.000	-0.360	0.000	0.000	-0.364	0.011
	c32	0.242	0.000	0.297	2.155	0.000	0.306	1.929
	c33	-0.102	1.574	-0.221	0.007	1.473	-0.210	0.133
	c34	-0.096	0.000	0.022	0.003	0.000	0.019	0.000
	c3x	0.283	2.920	0.441	1.990	3.007	0.432	1.972
	hc1	0.094	0.481	0.087	0.377	0.475	0.087	0.393
	hc2	-0.024	0.376	-0.047	0.044	0.411	-0.052	0.123
	hc3	0.035	0.000	0.061	0.502	0.000	0.056	0.524
	hc4	0.041	0.000	0.013	0.444	0.000	0.015	0.457
	hn	0.373	0.190	0.390	0.246	0.196	0.389	0.253
	hx	0.043	0.000	-0.000	0.000	0.000	0.003	0.000
	n4	-0.568	0.000	-0.663	0.004	0.000	-0.654	0.057
Ser	c	0.866	1.992	0.966	0.683	2.018	0.968	0.634
	c31	-0.135	0.943	-0.204	0.678	0.550	-0.199	0.597
	c32	-0.086	1.191	-0.214	1.972	1.249	-0.213	1.918
	hc1	0.045	0.106	0.065	0.162	0.221	0.064	0.192
	hc2	0.029	0.000	0.070	0.047	0.000	0.069	0.088
	n	-1.087	0.644	-1.094	0.507	0.649	-1.108	0.322
Asn	c31	-0.282	0.014	-0.255	0.000	0.019	-0.257	0.000
	c32	0.526	1.888	0.468	1.929	1.897	0.466	1.931
	h1	-0.062	0.000	-0.053	0.000	0.000	-0.053	0.000
	hc1	0.061	0.473	0.059	0.452	0.474	0.059	0.455
	hn	0.443	0.000	0.451	0.190	0.000	0.459	0.281
	ho	0.419	0.279	0.376	0.192	0.312	0.376	0.229
	o	-0.637	0.000	-0.691	0.455	0.000	-0.697	0.535
	oh	-0.722	0.066	-0.659	0.164	0.023	-0.658	0.126
Cys	c31	-0.125	0.000	-0.161	0.000	0.000	-0.178	0.000
	c32	0.050	1.211	0.101	1.174	0.973	0.114	0.820
	h1	0.038	0.271	0.026	0.250	0.334	0.022	0.348
	hc1	0.057	0.455	0.065	0.464	0.483	0.069	0.499
	hs	0.189	0.254	0.177	0.276	0.279	0.177	0.294
	sh	-0.360	1.702	-0.364	1.771	1.758	-0.365	1.870

		RESP	3. RESP, Dir Pol	4. Co-opt Chgs, Dir Pol		5. RESP, SC Pol	6. Co-opt Chgs, SC Pol	
		q	α	q	α	α	q	α
Tyr	c31	-0.128	1.358	-0.116	0.925	2.146	-0.012	2.138
	c32	0.136	0.552	0.063	1.911	0.000	0.079	0.000
	ca1	-0.047	1.724	-0.120	0.957	1.834	-0.067	1.609
	ca2	-0.451	0.777	-0.495	1.775	0.994	-0.525	1.601
	ca3	0.542	1.183	0.684	0.000	0.421	0.693	0.000
	ca4	-0.127	0.000	0.007	0.063	0.000	-0.082	0.000
	ha1	0.143	0.251	0.162	0.350	0.320	0.155	0.272
	ha2	0.208	0.034	0.214	0.000	0.097	0.215	0.000
	hc1	0.032	0.195	0.030	0.223	0.000	0.002	0.000
	hc2	-0.009	0.270	-0.007	0.000	0.462	-0.003	0.523
	ho	0.376	0.000	0.389	0.000	0.056	0.398	0.199
	oh	-0.581	0.520	-0.626	0.711	0.773	-0.632	0.585
Tyr	c31	-0.123	1.419	-0.184	0.473	2.164	-0.116	1.538
	c32	0.103	0.444	0.011	2.169	0.000	0.010	1.442
	ca1	-0.271	1.702	-0.425	0.828	1.643	-0.424	1.059
	ca2	-0.241	0.662	-0.212	1.803	0.727	-0.200	1.860
	ca3	0.335	1.407	0.400	0.072	1.219	0.373	0.000
	ca4	0.103	0.000	0.315	0.001	0.000	0.301	0.000
	ha1	0.182	0.204	0.213	0.343	0.388	0.209	0.398
	ha2	0.172	0.070	0.179	0.000	0.121	0.177	0.000
	hc1	0.030	0.177	0.048	0.316	0.000	0.030	0.087
	hc2	-0.010	0.322	-0.002	0.000	0.475	-0.001	0.134
	ho	0.372	0.088	0.389	0.015	0.139	0.401	0.399
	oh	-0.546	0.388	-0.581	0.712	0.494	-0.578	0.353
Val	c31	-0.510	0.000	-0.412	0.000	0.000	-0.416	0.000
	c32	0.572	2.858	0.557	2.291	2.809	0.563	2.279
	hc1	0.114	0.356	0.087	0.410	0.367	0.088	0.416
	hc2	-0.068	0.000	-0.107	0.138	0.000	-0.108	0.143
Val	c31	-0.514	0.000	-0.405	0.000	0.000	-0.409	0.000
	c32	0.564	2.877	0.551	2.297	2.827	0.558	2.282
	hc1	0.116	0.354	0.086	0.409	0.365	0.086	0.415
	hc2	-0.066	0.000	-0.108	0.142	0.000	-0.109	0.149
Val	c31	-0.526	0.000	-0.386	0.000	0.000	-0.392	0.000
	c32	0.555	2.931	0.522	2.312	2.874	0.529	2.284
	hc1	0.120	0.348	0.082	0.408	0.359	0.083	0.414
	hc2	-0.061	0.000	-0.103	0.145	0.000	-0.104	0.157
Val Dimer	c31	-0.523	0.000	-0.371	0.000	0.000	-0.379	0.000
	c32	0.567	3.148	0.495	2.540	3.051	0.503	2.519
	hc1	0.118	0.315	0.079	0.374	0.341	0.080	0.390
	hc2	-0.062	0.000	-0.090	0.109	0.000	-0.090	0.113
Methane Dimer	c3	-0.500	0.980	-0.496	0.983	1.020	-0.497	1.018
	hc	0.125	0.201	0.124	0.201	0.196	0.124	0.196
Nitro-benzene	ca1	-0.194	0.000	-0.286	0.598	0.000	-0.291	1.670
	ca2	-0.111	0.779	-0.047	1.867	0.498	-0.005	1.523
	ca3	-0.127	1.573	-0.133	0.000	1.914	-0.156	0.000
	ca4	0.092	2.964	0.170	1.099	3.643	0.158	0.000
	ha1	0.183	0.000	0.196	0.145	0.000	0.188	0.062
	ha2	0.146	0.443	0.149	0.041	0.505	0.145	0.093
	ha3	0.151	0.000	0.152	0.328	0.000	0.151	0.436
	no	0.751	0.000	0.715	0.862	0.000	0.707	0.062
	o	-0.458	0.619	-0.465	0.495	0.552	-0.467	0.925

elements in a prior study, which used only experimental molecular polarizabilities as target data [180]. Perhaps the largest discrepancy is for nitrogen, whose optimized polarizabilities are significantly lower here. However, when appropriately atom-typed polarizabilities from the prior study are substituted for the optimized ones developed here, they yield polarized ESPs that deviate about twice as much from the reference polarized QM ESPs, regardless of the point-charge model used (data not shown). About 30% of this increased deviation

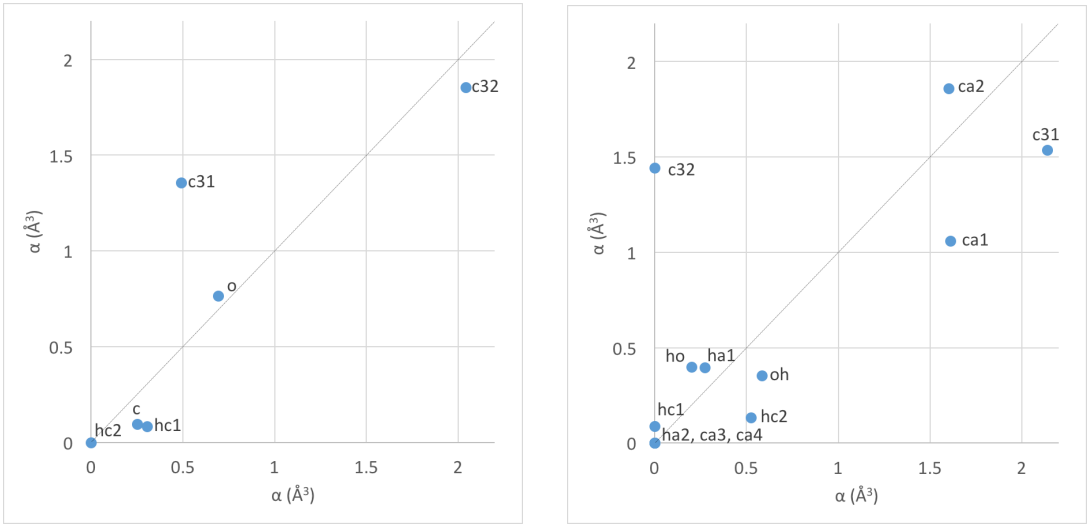
probably traces to the tendency of the present quantum calculations to underestimate the experimental molecular polarizabilities used (see above).

Table 3.3: Means and standard deviations of atomic polarizabilities by element, for Models 3-6 (Å^3).

	3. RESP, Dir Pol		4. Co-opt Chgs, Dir Pol		5. RESP, SC Pol		6. Co-opt Chgs, SC Pol	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
C	1.107	1.070	0.972	0.893	1.102	1.156	0.950	0.903
H	0.163	0.164	0.198	0.164	0.184	0.188	0.232	0.171
O	0.413	0.276	0.556	0.203	0.459	0.320	0.569	0.266
N	0.215	0.372	0.458	0.431	0.216	0.375	0.147	0.152

It is also instructive to examine how the optimized polarizabilities can change with molecular conformation. For the simple valine analog, all three conformations are quite similar to each other, and the three sets of polarizabilities agree to within 0.015Å^3. However, the picture is more complicated for the two conformations apiece of the aspartate and tyrosine analogs (Figure 3.6). The molecular conformations tested for the aspartate analog differ in the torsion of the side chain relative to the ethyl terminus. Thus, in one conformation the shortest oxygen to methyl carbon (*o-c31*) distance is 2.8Å while in the other conformation it is 3.2Å. The optimal polarizability of *c31* is larger for the structure with the shorter *o-c31* distance, but the polarizabilities of the other atoms are quite similar. The two conformations tested for the tyrosine analog are the same to within 0.078Å RMSD. Nonetheless, the optimized polarizabilities for certain atoms differ substantially between the two conformations; notably, the *c32* atom type has polarizability 0.000Å^3 in one structure and 1.442Å^3 in the other. However, this change is partly compensated by opposite shifts in the polarizabilities of atom types *ca1* and *c31*. Interestingly, when the optimized parameters

are swapped between conformers, the ESP errors, R, change by <0.05 kcal/mol-e; thus, the optimal solutions found here are degenerate.



(a) Aspartate Polarizabilities

(b) Tyrosine Polarizabilities

Figure 3.6: Comparison of Model 6 atomic polarizabilities from different molecular conformations of aspartate and tyrosine analogs

As detailed in the prior section, all four inducible dipole models (Models 3-6) yield rather similar levels of accuracy (Figure 3.2). Comparisons of the optimized parameters (Table 3.2) indicate that they, also, tend to be quite similar, for matching molecules and atom types. For example, polarizabilities optimized for the direct and self-consistent models with RESP charges (Models 3,5) agree well with each other (Figure 3.7). This agreement suggests that the optimized polarizabilities might be swapped between the two molecules with little loss of accuracy. In fact, using polarizabilities optimized for Model 5 in Model 3 increases the mean RMSE, R, by only 0.05 kcal/mol-e, while the reverse swap decreased the mean RMSE by 0.03 kcal/mol-e. Thus, the optimized values are effectively interchangeable. Similarly, the co-optimized charges of Models 4 and 6 are quite similar to corresponding baseline RESP charges (Figure 3.8). Overall, then, the four inducible dipole models end up

with similar parameters and yield similar levels of accuracy.

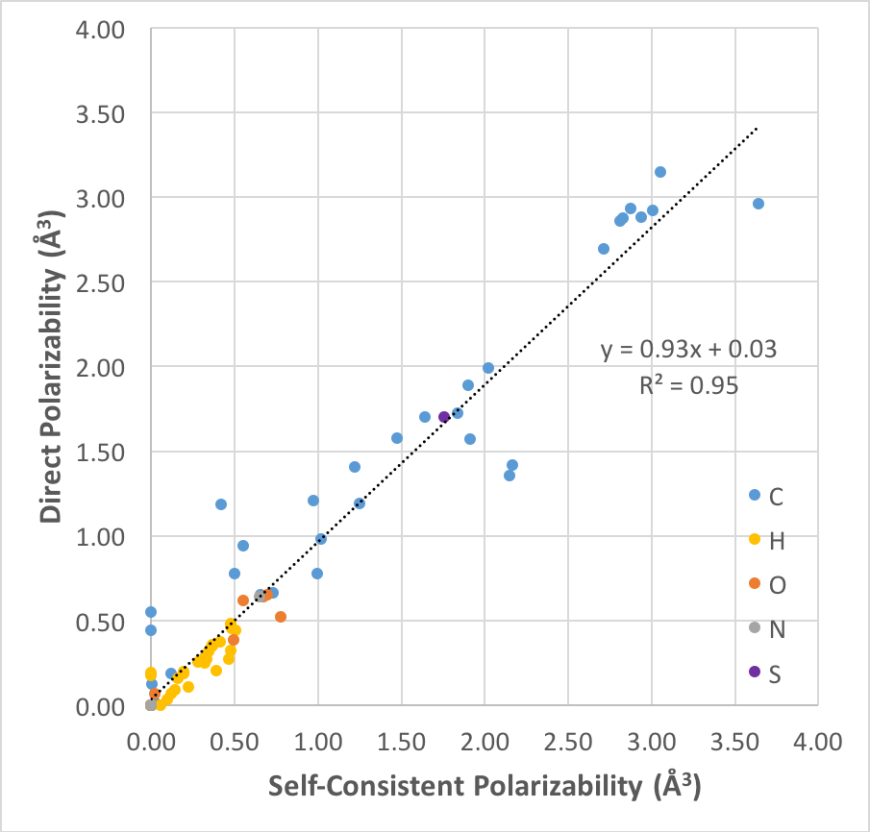


Figure 3.7: Comparison of self-consistent and direct polarizabilities for Models 3 and 5 across all molecules.

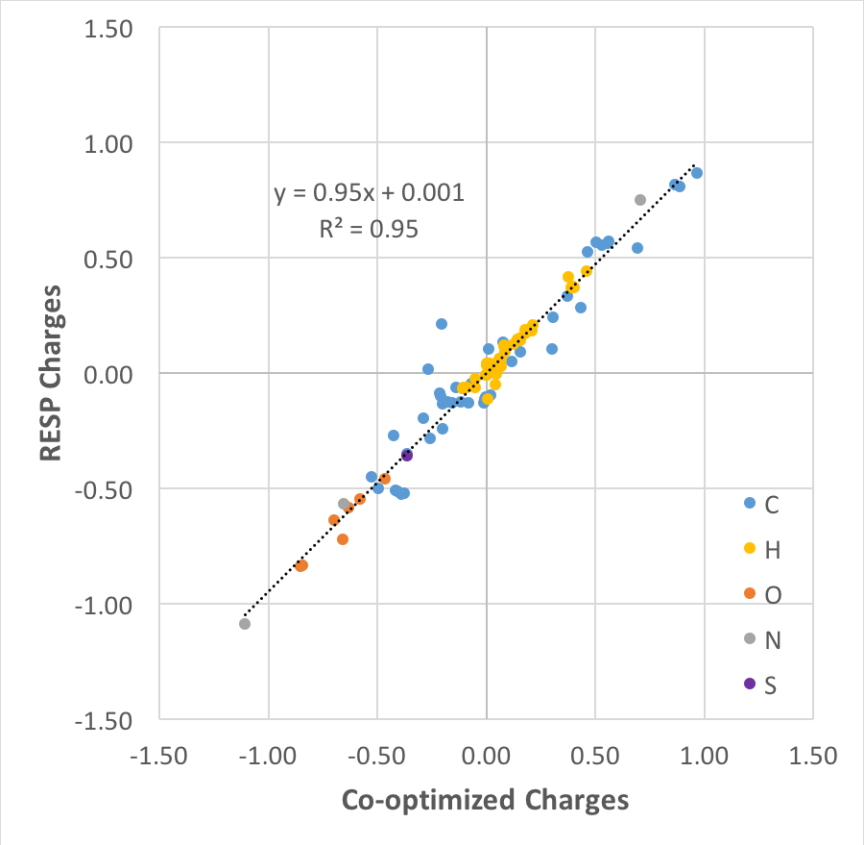


Figure 3.8: Comparison of co-optimized and RESP charges for Models 5 and 6 across all molecules.

3.4 Conclusions

Incorporating conformation-dependent electronic polarization into force fields promises to advance the predictive power of molecular simulations. Although various polarization models exist, the self-consistent atomic polarizability model has emerged as a well-regarded standard. This approach has been implemented in its full form, and approximated via the Drude oscillator method, as well as the first-order, or direct, approximation. Conformation-dependent polarization has also been modeled via a redistribution of charge among atomic point charges[149, 48], instead of using added point dipoles. (A combined approach has

also been reported[163].) Charge redistribution approaches cannot capture polarization that is not directed along bonds, but they are simpler, because they do not add point dipoles to the representation of the molecule.

In considering the relative merits of various polarization models, it is useful to distinguish between how a model represents polarization, and how it computes the polarization within this representation. The present study has considered two polarization representations, atom-centered point dipoles, and redistribution of atom-centered charges; and it has examined two response models for the point-dipole representation, namely direct and self-consistent inducible dipoles. (Although not studied here, there are also useful response models for the charge-based representation[149, 188].) It should be emphasized that, although the response models used in polarizable force fields are physically motivated, they are simplified representations of the mechanisms controlling how electrons in molecules shift in response to inducing fields. This holds not only for the charge-redistribution models, but also for the physically appealing picture of self-consistent, atom-centered inducible point dipoles. Thus, given that the parameters have been properly optimized, the accuracy of a polarization model can be limited by its representation of polarization, by its response model, or both.

A central finding of this study is that it is the response model, rather than the polarization representation, that limits the accuracy of the self-consistent atomic polarization model: point dipoles optimized independent of any response model yield much more accurate ESPs than the full polarization model. Indeed, the inducible dipole response model is so problematic that it yields ESPs less accurate than those achievable with a point-charge representation of polarization. Thus, although a key advantage of the dipole representation should be its ability to capture out-of-plane polarization, the induced dipole implementations

studied here are no more accurate than point charges at capturing out-of-plane induction of the planar molecule nitrobenzene.

This means that a polarization model using a point-charge representation could outperform the standard inducible dipole model, if it were outfitted with a good response model. Further accuracy might be available at modest computational cost by adding out-of-plane charge centers above and below aromatic rings; these could improve the representation both of the baseline potential and of polarization normal to the ring. On the other hand, because the point-dipole representation can yield greater accuracy than the point-charge representation, it would also be of great interest to seek improved response models for the point-dipole representation. Although any specific directions for improvements are currently speculative, several possible approaches may be considered. One is to allow for anisotropic polarizabilities. Another would be to modify the current electrostatic treatment of short-ranged induced dipole-induced dipole interactions, which clearly does not capture the complex details of quantum mechanical electronic reorganization. Finally, the success of Models 1 and 2 suggests that one might develop empirical response models for chemical fragments.

We also observed only a slight improvement in accuracy on going from the first-order, or direct, approximation of the induced dipole model to the fully self-consistent model, even though the self-consistent model is more physically complete. Perhaps greater improvement would be observed if a similar study could be done for molecules in the condensed phase, where there would be many more dipole-dipole interactions. Nonetheless, our results support prior suggestions [167, 151, 89, 183] that the direct approximation offers an advantageous combination of accuracy and computational efficiency. It is also interesting that polarizabilities optimized for the self-consistent model were essentially interchangeable

with those optimized for the direct model.

Somewhat greater improvement was observed upon replacing regular RESP charges with partial charges co-optimized with the atomic polarizabilities, particularly for the ionized compounds. Thus, although good results can be obtained with RESP charges combined with the inducible polarization models, it makes sense to re-optimize charges in the context of the inducible dipoles for charged compounds. Indeed, this affords to improvement in both the polarized and unpolarized baseline ESPs.

In the present study, the atomic polarizabilities were optimized to replicate the QM ESPs generated by molecules in the fields of external point charges. The agreement of molecular polarizabilities computed from these fitted atomic polarizabilities with molecular polarizabilities computed direction from the QM calculations supports the physical plausibility of the values assigned. On the other hand, some of the optimized polarizabilities differ significantly from previously published values. This often appears to result from compensating deviations of neighboring atoms. In addition, there is evidence that the solutions to the optimization problem can be degenerate, in the sense that equally good (or nearly so) fits can be obtained with different polarizabilities, much as observed in the optimization of point charges to match QM ESPs[14]. Procedures analogous to those used in RESP, such as the addition of weak restraints and/or the use of multiple conformations in fitting, could be used to generate more uniform polarizability assignments across chemically similar atoms.

When assessing the reliability of the present conclusions, it is reasonable to consider the degree to which the ability of a polarization model to fit the QM ESPs of polarized molecules is a useful metric of the model's quality. The central argument in support of this view is that this approach directly probes the relevant physics, and indeed other groups have used a similar approach [163, 9, 64]. In addition, RESP, one of the most successful

approaches to assigning partial atomic charges, works by fitting charges to QM ESPs, so a similar approach should also be suitable for adjusting and evaluating polarization models. More particuarly, using QM calculations at the HF/6-31G* level to calculate the reference QM ESPs is consistent with the standard RESP protocol. On the other hand, since this QM leads to molecular polarizabilities that underestimates experiment by roughly 30%, the reliability of the parameters might benefit from use of a larger basis set, such as aug-cc-pVTZ. It is also worth noting that partial charges fitted to the HF/6-31G* results in vacuum yield dipole moments that somewhat overestimate gas phase experimental results. These partial charges are regarded as suitable for simulations in the condensed phase, where some self-polarization occurs. However, an explicit treatment of polarizability should allow molecular dipole moments to adjust automatically to the environment, so it would be appropriate, when accounting explicitly for polarization, to set baseline gas-phase partial charges with a QM method that yields molecular dipole moments appropriate to the gas phase.

In summary, the accuracy of a polarization model is determined not only by how it represents polarization, but also by the response model it uses to compute polarization. Although atom-centered point dipoles can do an excellent job of representing molecular polarization, the inducible dipole response models typically used with this representation fall well short of the theoretical maximum accuracy it could attain. It should therefore be possible to develop more accurate polarization models not through more detailed representations of polarization, but instead through improved response models.

3.5 Acknowledgements

We thank Prof. Jan Jensen and an anonymous reviewer for valuable questions and comments that led to improvements in the present analysis. We also thank the National Institutes of Health (NIH) for Grant GM061300. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. MKG has an equity interest in, and is a cofounder and scientific advisor of, VeraChem LLC.

Chapter 3, in full, has been submitted for publication of the material as it may appear in The Journal of Physical Chemistry B 2016, Li, Amanda; Voronin, Alexey; Fenley, Andrew; Gilson, Michael K. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Attractive Interactions between Heteroallenes and the Cucurbituril Portal

4.1 Introduction

Molecular recognition and noncovalent interactions govern a wide range of chemical events[101, 136, 150, 120, 162, 113], including crystal growth[80], supramolecular chemistry[55, 178, 138, 40], self-assembly[186], catalysis[91] and almost every biochemical process[8], including protein-ligand binding, protein-protein binding, and DNA base-pairing. Noncovalent interactions encompass multiple binding mechanisms[109], such as hydrophobic interactions[169, 170, 19]; charge-charge, charge-dipole and dipole-dipole interactions[72, 42]; hydrogen-bonding[78]; and delocalization of electrons into antibonding orbitals[16, 131]. For the cucurbiturils, a class of host molecules with an already rich supramolecular chemistry, binding of guest molecules is thought to be dominated by three

fundamental mechanisms[97, 127, 77, 11]: (a) charge-dipole interactions between a strong dipole of the host's carbonyl-fringed portal and the positive charge of a guest; (b) hydrogen bonding between the portal carbonyls and a guest's donor moieties; and (c) hydrophobic interactions within the cucurbituril cavity, which is formed by the concave faces of the glycoluril subunits and their methylene bridges. Identification of new binding mechanisms accessible to the cucurbiturils would further enrich the uses of this important family of hosts, and would be of interest as another available fastener for use in the design of targeted molecules for many applications.

While investigating the structural and dynamic properties of bistable rotaxanes made of 1,4-bis(alkylaminomethyl)benzene and cucurbit[6]uril, **1**, we noticed a remarkable crystallographic feature in one of the complexes[160]. The azide moiety of the guest *N,N'*-bis-(azidoethyl)-p-xylylene diammonium chloride forms close contacts with the oxygen atoms of the host, and the azidoethyl group adopts a conformationally unfavorable gauche state (Figure 4.1). These observations suggested an interesting stabilizing attraction between the azide group and the carbonyls. We realized that this rare attractive interaction could provide a yet unexploited tool in supramolecular chemistry and therefore should be further explored.

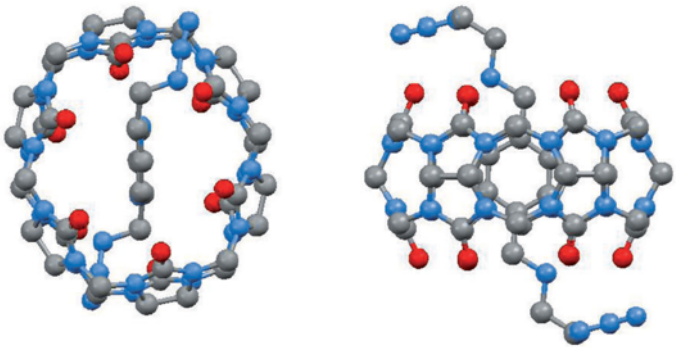


Figure 4.1: X-ray crystal structure of complex **9b** in ref. 13 (top and side views). Color code: C, gray; N, blue; O, red.

Here, we report attractive interactions between organic azides in a guest molecule and the portal carbonyls of cucurbit[6]uril, and characterize this interaction by crystallography, NMR and IR spectroscopy, and quantum chemical calculations. The results provide evidence that favorable interactions of heteroallenes with carbonyls are a general phenomenon that can be exploited for supramolecular applications.

4.2 Results and Discussion

4.2.1 Synthesis

In order to characterize the attractive interaction between organic azides and the CB[6] portals, we synthesized a series of guest molecules with two azidoethylamine end groups, the *N,N*-(2-azidoethyl)-α,ω-alkanediamines **2**-**6**, using a general four-step procedure (Figure 4.2). Reaction of sodium azide with bromoethylamine hydrobromide, **7**, in water produced 2-azidoethylamine **8**. Protection of the latter with Boc anhydride afforded **9**, which underwent *N*-alkylation with the appropriate α,ω-dibromoalkane to produce compounds **10**a-e. Finally, removal of the Boc protecting groups with ethanolic HCl afforded the guest molecules, **2**-**6**, in the form of their dihydrochloric salt in overall yields of 20-30%.

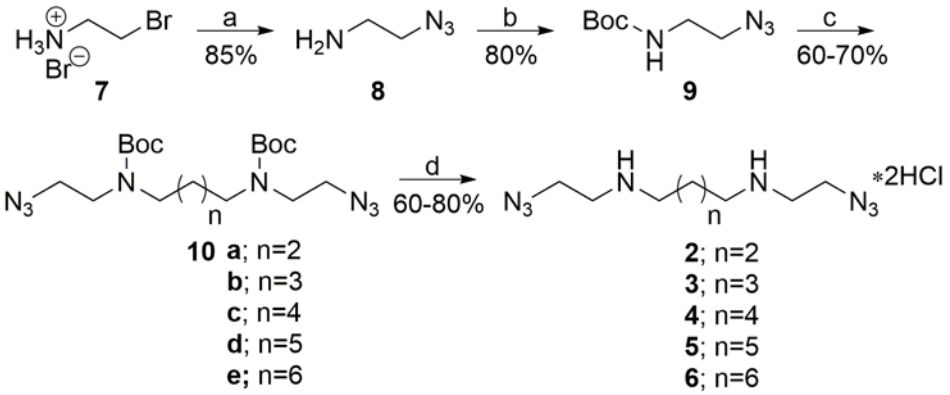


Figure 4.2: Synthesis of guest molecules **2**-**6**. Reagents and conditions: a) NaN_3, H_2O, 80°C, 24 h; b) $(Boc)_2O$, Et_3N, CH_2Cl_2, RT, 16 h; c) α,ω-dibromoalkane, NaH, DMF, RT, 48 h; d) HCl (4N), EtOH, RT, 16 h.

4.2.2 NMR studies

The stoichiometry of the inclusion complexes was determined by ^{1}H NMR. Each of the protonated guest molecules, **2**-**6**, was dissolved in D_2O-DCl at pH 5, then mixed with solid **1** (1 equiv) and the mixture was kept at room temperature for 16h. Formation of 1:1 inclusion complexes was evident by their ^{1}H NMR spectra, which exhibited significant changes in the chemical shifts of the guest molecules in comparison with their spectra in the absence of **1** (Figure 4.3). Consistent with previous observations[160, 114], all protons of the guest molecule residing in the host interior exhibit significant upfield shifts $\Delta\delta$, which increase with the depth of burial in the binding cavity. For example, a comparison between free **5** and its complex **5@1**, reveals that the upfield shifts of the hydrogen atoms along the oligomethylene chain are 0.02, 0.52, 0.88 and 0.98 ppm, beginning with the α-methylene attached to the ammonium groups and moving inward. This shielding presumably reflects the cumulative influence of the urea units, which form a hydrophobic wall of filled π orbitals, and make the host's cavity remarkably different from the acidic aqueous environment of the bulk solvent.

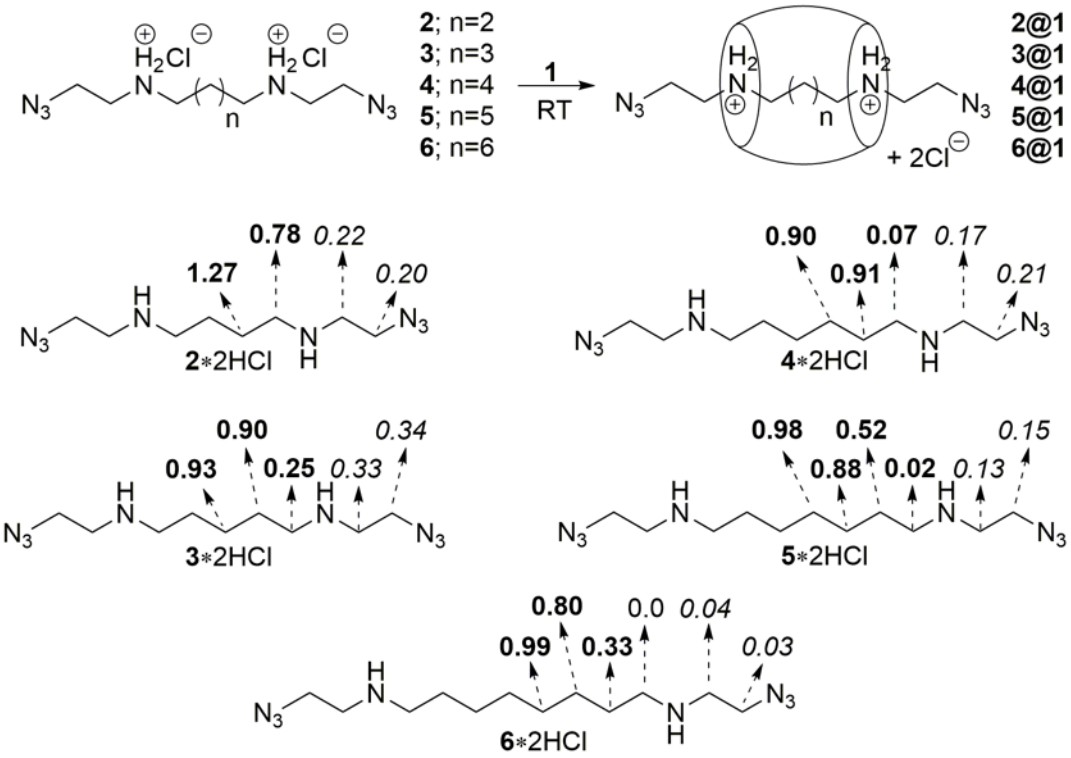


Figure 4.3: ^{1}H NMR induced chemical shift differences ($\Delta\delta$, ppm) upon formation of 1:1 complexes between guests **2**-**6** and **1** at room temperature in D_2O-DCl containing traces of DMSO (δ=2.71 ppm) as an internal standard. The shielding effect (ppm, upfield shift) is shown in bold, whereas deshielding (ppm, downfield shift) is shown in italics.

Interestingly, the upfield shift of the α-methylene protons, which reside at the portals, decreases with increased chain length: $-\Delta\delta$ = 0.78, 0.25, 0.07, 0.02 and 0.00 ppm for C_4, C_5, C_6, C_7 and C_8, respectively. This trend indicates that with chains of increasing length, the α-methylene group is pushed further out of the cavity. We have previously reported that all guest protons that reside outside the cavity in the vicinity of the portal undergo deshielding, probably due to the strong anisotropic effect exerted by the combined dipole of the portal of the carbonyl groups[160]. Thus, in case of the octa-methylene chain, **6**, the lack of any shift exhibited by α-methylene protons indicates that the shielding and deshielding effects completely offset one another.

The observation that the α-methylene group resides at the portal, regardless of the length of the guest's oligomethylene chain, may be understood in terms of induced fit[88], as further supported by our crystallographic data (*vide infra*). For example, whereas the penta-methylene chain exhibits an all-anti conformation, the hexa-methylene chain adopts a slightly folded conformation that allows it to retain favorable interactions with the host[96]. This phenomenon has been reported for alkyltrimethylammonium salts hosted by cucurbiturils[92], as well as for other host-guest complexes[193]. The ability of a molecule to shorten its length by adopting multiple conformations, which are achieved by multiple *gauche* interactions, also provides an entropic advantage to the complexation event[194]. The chemical shifts of the oligomethylene chain represent a valuable probe of the above-mentioned host-guest interactions, their conformation and relative orientation. In addition, the NMR data of the azidoethyl groups, which always reside outside the cavity, represent another helpful probe of these properties: $-\Delta\delta$ = 0.22-0.20, 0.33-0.34, 0.17-0.21, 0.13-0.15, and 0.04-0.03 ppm for C_4, C_5, C_6, C_7 and C_8, respectively.

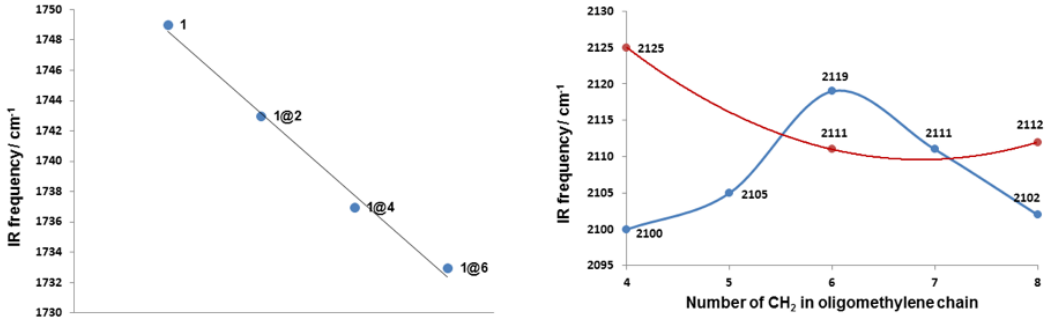
4.2.3 IR study

To further probe the host-guest interactions, we compared the solid state IR bands of the free guests and their host-guest complexes, and particularly those of the urea and azide groups at 2050-2150 and 1700-1750 cm^{-1}, respectively. Both bands can report on the local electrostatic environment. For example, the azide stretching vibration band of β-azidoalanine at 2000-2200 cm^{-1} is strongly red-shifted (14 cm^{-1}) in the hydrophobic environment of DMSO relative to water[123].

As evident from Figure 4.4a, the ureidocarbonyl vibration frequency becomes increasingly red-shifted (6-16 cm^{-1}) on going from free CB[6] to complexes with guests of

increasing size. This trend indicates that the exposure of the host carbonyl groups to the surrounding aqueous environment of water molecules is progressively attenuated by the hydrophobic parts of the guest molecules, which replace more water molecules.

The trends found for the azido stretching frequencies are more complex. In general, increased intermolecular interactions upon binding would weaken the internal N-N bonds, manifested by red shifts. The mixed effects shown in Figure 4.4b suggest that the azide groups in the free guest molecules are involved in inter- or intramolecular attractive interactions, which could be weaker or stronger than the azide-carbonyl interactions. Indeed, gas-phase MM2 dynamics calculations (Figure A.2) indicate that the free guest molecules feature intramolecular ion-dipole interactions between an azide group and a distant ammonium group, which is augmented by a dipole-dipole attractive interaction of two azides in an antiparallel orientation. Although gas-phase calculations may not fully represent the situation in the solid, the strong tendency of the free guest molecules to participate in inter- and intramolecular attractive interactions is self-evident. The loss of these interactions upon binding to **1** may not be fully compensated by the attractive host-guest interactions at the level of a single azide group.

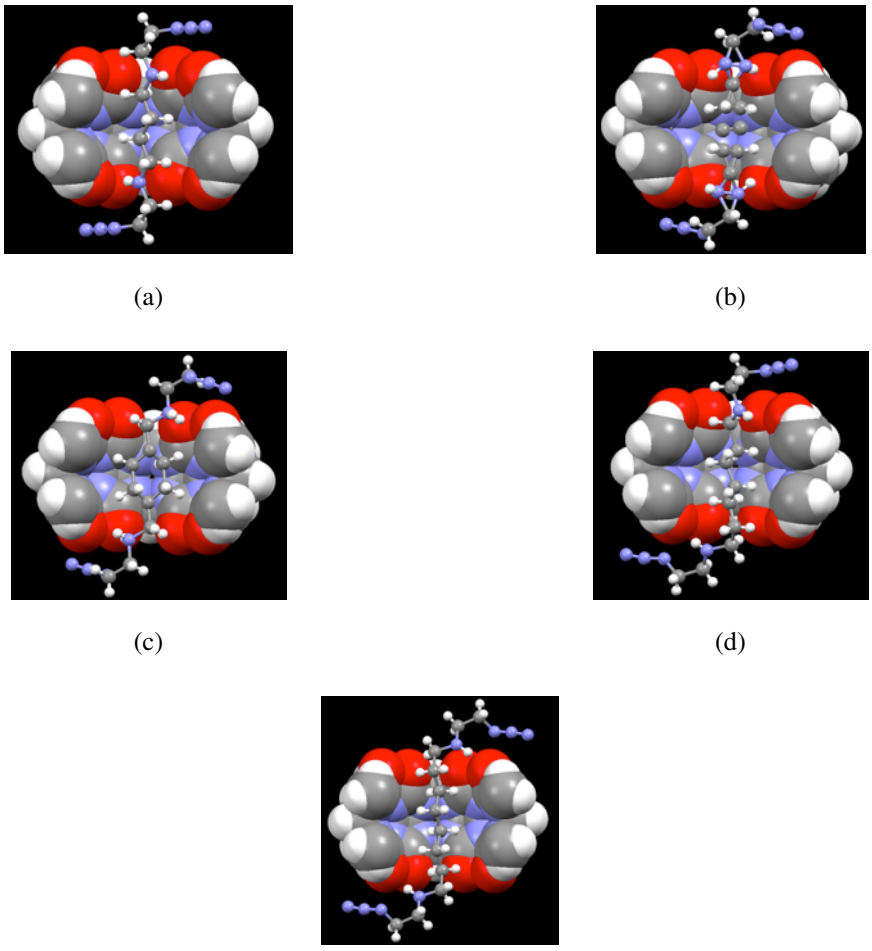


(a) (b)

Figure 4.4: IR vibrational frequencies (KBr pellet) of ureido and azide groups. (a) Ureido carbonyl stretching band for free host **1** and its complex with guests **2**, **4** and **6**. (b) Azido stretching band as a function of the number of methylene groups for guests **2**-**6** in the absence of the host (blue), and for guests **2**, **4** and **6** in complex with the host (red).

4.2.4 X-ray crystal structures

The crystallographic studies provide valuable structural information concerning specific interactions within the host-guest complexes was gained from X-ray crystallography. Single crystals of complexes **2@1**, **3@1**, **4@1**, **5@1** and **6@1** suitable for X-ray analysis were obtained from acidic (pH=6) aqueous solution by vapor diffusion. Crystallographic and refinement data of all structures are provided in Appendix A (Table A.5). Our structures (Figure 4.5) may be compared with the reported complexes of **1** hosting α,ω-alkanediammonium guests[88].



(a) (b)

(c) (d)

(e)

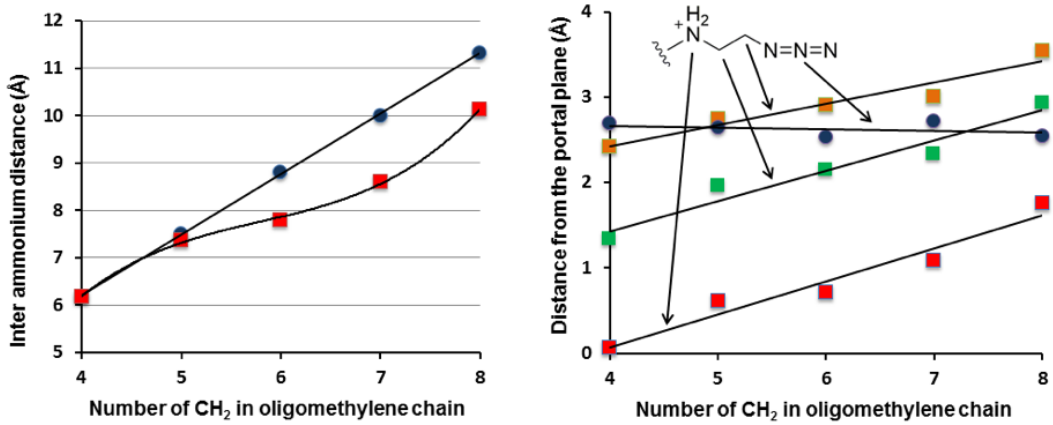
Figure 4.5: X-ray crystal structures of (a) **2@1**, (b) **3@1**, (c) **4@1**, (d) **5@1** and (e) **6@1**. The host, **1**, is presented in a cross-sectional, space-filling format. Atom doubling and missing bonds indicate disordered structures.

In both families of complexes the oligomethylene chain connecting the two ammonium groups adopts the same conformation within the CB[6] interior. Depending on the chain length, the guests assume an extended or partially bent conformation inside the cavity. The distance between the ammonium groups varies from 6.18 Å in **2@1** to 10.14 Å in **6@1**. Since the distance between the two portal planes, which accommodate the carbonyl oxygen atoms, is 6.1±0.1 Å, the oligomethylene chain in **2@1** adopts a fully extended

conformation, whereas the longer chains exhibit partially folded conformations. Thus, while the intramolecular distances between two ammonium nitrogen atoms in the fully extended conformation of free **3**, **4**, **5** and **6** are 7.51, 8.81, 10.0 and 11.33 Å, respectively,[1] these distances shrink to 7.36 and 7.80 Å in their corresponding complexes (Figure 4.6a). These folded conformations award the guest molecules with maximal charge-dipole interactions between the ammonium groups and the portals, along with favorable hydrophobic interactions between the oligomethylene chain and the interior of **1**.

Since this study aims at understanding the nature of the specific interactions between the host portals and the azide groups of the guest, their relative orientation is of particular interest. All structures reveal two consistent structural features. First, the azide group itself preserves a nearly linear geometry, as reflected by the consistent bond angles, CNN ($116\pm1^\circ$) and NNN ($172\pm1^\circ$). Second, all azide groups maintain short contacts with two carbonyl oxygen atoms through their central β- and terminal γ-nitrogen atoms (Figures 4.6b, 4.7, and 4.8). Remarkably, while the distance between the portal plane and most atoms at the guest end groups increase progressively with the molecular size, the β-nitrogen atoms maintain a constant distance from the portal plane in all homologs (Figure 4.6b), pointing at a strong attractive interaction between the azide group and the portal.

[1] The distance between the two ammonium groups of 1,5-pentanediammonium and 1,6-hexanediammonium is estimated according to molecular modeling study (MM2 force field).



(a) (b)

Figure 4.6: Crystallographic distances. (a) The intramolecular distances between the ammonium nitrogen atoms in the fully extended conformation of free guest molecules (blue circles) and in the corresponding complexes (red squares). (b) The intramolecular distances between the portal plane and selected atoms at the guest end group. For the non-symmetrical complex 5@1, the data points represent an average between the two sides of the complex.

The significantly short inter-atomic distances between the positively polarized nitrogen atoms of the azido groups and the negatively polarized carbonyl oxygen atoms approach the sum of the effective van-der-Waals radii of these atoms (∼3.07 Å, Figure 4.7)[196].[2] Such distances require a *gauche* conformation of the azidoethyl chain, which is reflected by the NCCN dihedral angle in all bound guest molecules, ranging between 64° and 71°.[3] This binding mode is modulated by the size of the guest. With the smaller guests, **2**, **3** and **4**, the host carbonyl groups interact mainly with the β- and γ- nitrogen atoms of the azide. In the non-symmetrical complex **5@1**, however, one of the two azides is pushed further away from the portal, so that its γ-nitrogen is further from the carbonyl oxygen, while hydrogen bonds

[2] Since different atomic radii are used in the van der Waals programs, we shall refer to Bondi radii of atoms

[3] An alternative driving force for this *gauche* interaction could be a potential intramolecular hydrogen bonding between the ammonium group and the α nitrogen atom of the azide group. Nevertheless, the contribution of this hydrogen bonding to the observed folded conformation seems negligible because the ammonium group is bound more strongly to the carbonyl groups.

are formed between the methylene group on the α-nitrogen and the carbonyls (Figure 4.5d). This trend is more pronounced with the symmetrical complex **6@1** where both azide groups are pushed away from the portal.

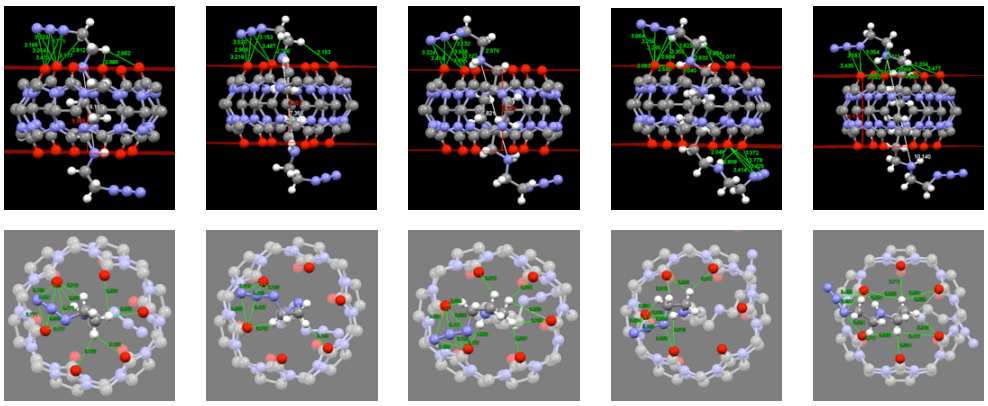
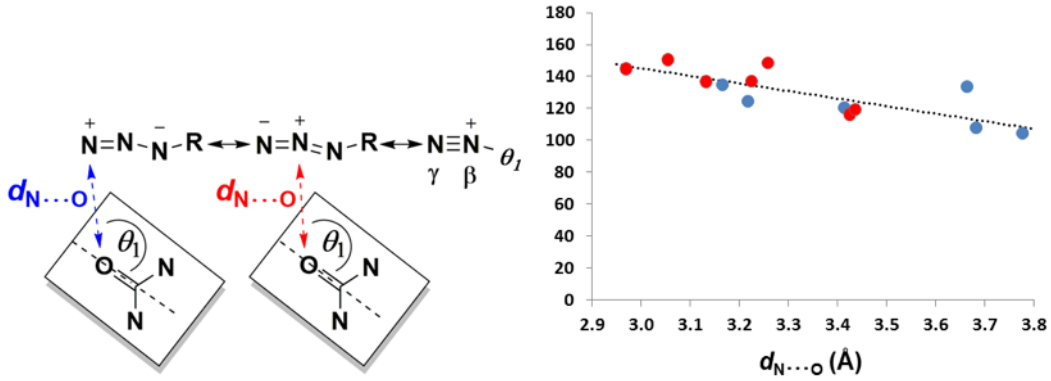


Figure 4.7: Representative short contacts between the guest azidoethyl groups and the carbonyl oxygen atoms of the host. From left: **2@1**, **3@1**, **4@1**, **5@1** and **6@1**.

Two significant structural parameters that characterize the attractive interaction between the carbonyl and azide are the distance ($d_{N\cdots O}$) between the two heteroatoms and the angle (θ_{NOC}) between the $d_{N\cdots O}$ vector and the carbonyl bond (Figure 4.8a). The distance $d_{N\cdots O}$ is of particular interest because it can shed light on the issue of binding mechanism, pointing at the relative importance of either $n\rightarrow\pi^*$ interaction[84] (*vide infra*) or the orthogonal dipolar description[129]. As can be concluded from the scatterplot correlation between $\theta_{N_\beta OC}$ and $d_{N_\beta\cdots O}$, the shortest interactions occur between the carbonyl and the β-nitrogen. Interestingly, the angle θ_{NOC} at short distances is narrowly distributed around 140° (Figure 4.8b), and the angle diminishes linearly with increased $d_{N\cdots O}$.

In order to set these results in context, we searched the Cambridge Structural Database (CSD)[34, 5] for short carbonyl-azide contacts (up to 3.6Å) and found 45 cases, which exhibited 84 interactions between a carbonyl oxygen and a β-nitrogen of an organic



(a) (b)

Figure 4.8: Geometrical parameters. (a) Definition of geometrical parameters $\theta_{NOC}(^\circ)$ and $d_{N_\beta\cdots O}$ (Å). (b) Scatterplot correlation between θ_{NOC} and $d_{N\cdots O}$, extracted from the X-ray structural data. The red circles refer to the interactions with the azide β-nitrogen and blue circles refer to the γ-nitrogen.

azide. The distribution of geometries for the intermolecular cases (Figure 4.9, blue circles) encompasses those seen in our host-guest complexes (Figure 4.8b, red circles), typically ranging within 140° ± 20° at a distance of 2.8-3.3Å. The distribution of angles for intramolecular interactions of this type in the CSD results is shifted and narrowed, relative to the intermolecular cases (Figure 4.9, red circles). This would be consistent with a view that tighter geometric constraints in the intramolecular setting prevent geometric optimization of an attractive carbonyl-azide interaction.

4.2.5 Computational analysis

The interactions between the host, **1**, and the azide moiety of guest **5** were further analyzed by quantum-mechanical (QM) electronic structure calculations. We examined the attractive forces between the azide group of the guest and the carbonyl group of the host, and compared these with the corresponding interactions of three geometrically similar groups, isocyanate, isothiocyanate, and propadiene. Like azide, isocyanate and isothio-

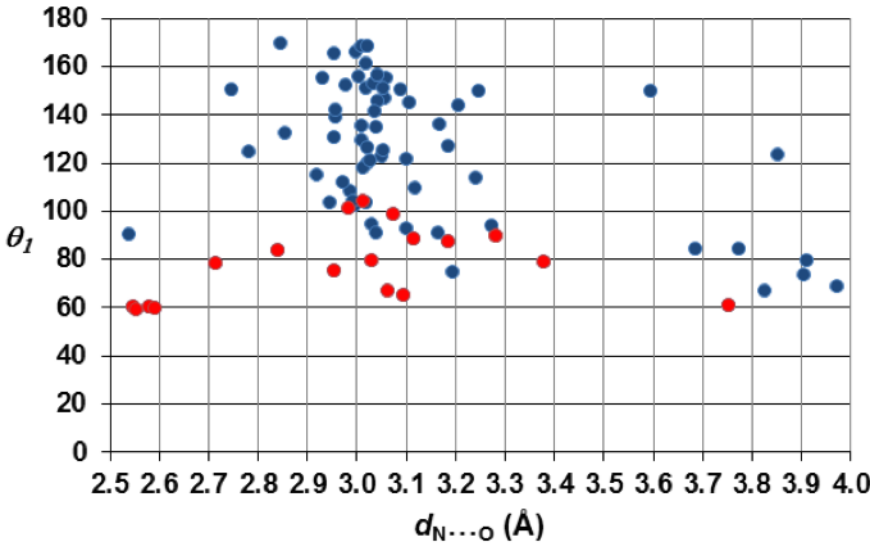


Figure 4.9: Scatterplot correlation between $\theta_{N_\beta OC}$ (°) and $d_{N_\beta \cdots O}$ (Å) extracted from the CSD database. The red circles refer to intramolecular interactions whereas the blue circles describe intermolecular interactions.

cyanate are heteroallenes, and thus might form similarly attractive interactions with the host. In contrast, propadiene is a nonpolar allene, and thus may not establish such favorable interactions. The character of these various interactions were further analyzed in terms of potential contributions from dispersion forces, electrostatic interactions, and $n \rightarrow \pi^*$ delocalization[84].

The crystal structure of the **5@1** system was modified, using the Maestro software [2] to generate models of complete host-guest complexes for the isocyanate, isothiocyanate and propadiene guests (**11**, **12**, **13**, respectively, in Figure 4.10, by direct substitution of the nitrogens in the azide moieties of **5** with the appropriate other elements, and addition of three hydrogens for propadiene. Each host-guest complex, **5@1**, **11@1**, **12@1** and **13@1**, was then geometrically optimized using the semi-empirical PM6-DH+[93] method with the COSMO implicit solvation model[90]. Higher-level quantum calculations, used to assess interaction energies, etc., were then carried out on fragments of these optimized

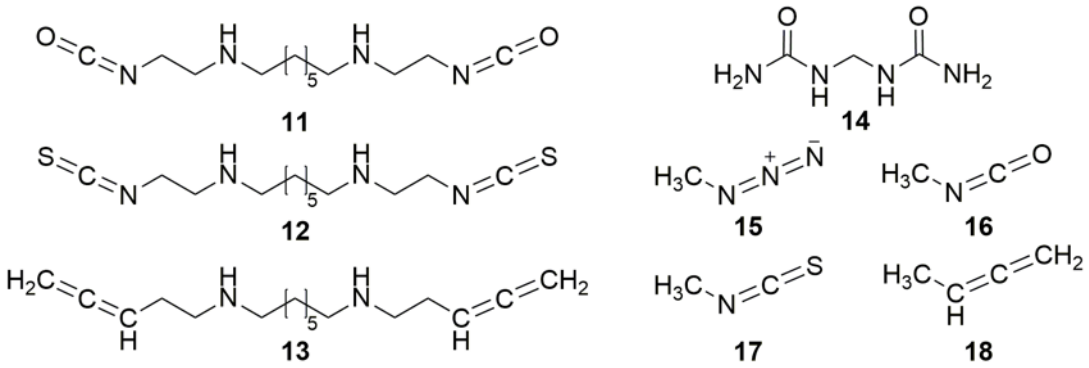


Figure 4.10: Structures used in computational studies.

systems where the host was represented by methylenediurea, **14**, and the guests by the small molecules **15-18** (Figure 4.10) without further changes in geometry. The azides at the two host portals adopt somewhat different geometries in the crystal structure of **5@1**, and the optimized host-guest structures retain these differences. We report computations for both geometries. Separate geometry optimizations, with PM6-DH+ and COSMO, were also carried out for each of the guest fragments, in order to look for possible geometric changes on binding.

Interaction energies were calculated for the representative dimers using both symmetry-adapted perturbation theory (SAPT)[79], implemented in the PSI4 program[174], and MP2[115], implemented in Gaussian D.01[52]. SAPT has been shown to accurately describe noncovalent interactions between molecules, including binding energies of large organic complexes[67]. We computed the total interaction energies as well as the decomposed energy terms resulting from electrostatic (elst), exchange (exch), induction (ind), and dispersion (disp) contributions using SAPT2+3/aug-cc-pVTZ[69]. It is worth noting that all orders of SAPT tend to overestimate attractive forces, and the performance of the SAPT approach depends strongly on the order of the SAPT expansion. The highest order, SAPT2+3, provides a full description of third-order interactions with accuracy that approaches the gold-

standard CCSD(T)/CBS level[103, 126]. Since current implementations of SAPT cannot include solvent effects, we also carried out similar calculations with MP2/aug-cc-pVTZ, both with and without the polarizable continuum method (PCM) implicit solvent model of water[172].

The nature of these host-guest interactions were further characterized by calculations by the NBO 3.0 program[54] as implemented in Gaussian 09 D.01. First, the electrostatic character of the allene moeities of the guests was evaluated by computing atom-centered natural charges, which are calculated using natural population analysis (NPA)[137], and assigning each atom a partial charge equal to its nuclear charge less the total population of its natural atomic orbitals. Second, the possibility that $n \rightarrow \pi^*$ interactions might play a role in the azide-carbonyl attraction was evaluated with natural bond orbital (NBO) analysis, which uses second-order perturbative analysis to estimate energies of donor-acceptor interactions[122].

The geometrically optimized **5@1** structure has $d_{\beta \cdots O}$ distances of 3.2 and 3.5Åfor the azide moieties at the two host portals, as measured between the β position of each azide group and the closest host carbonyl oxygen atom. These distances agree well with those observed in the crystal structure (Figure 3B) and are comparable with distances measured between groups engaged in orthogonal dipole interactions[129]. For the end groups of the **5** analogs, isocyanate, **11**, isothiocyanate, **12**, and propadiene, **13**, the corresponding distances are slightly increased, to 3.3 and 3.6Å for **11**, 3.4 and 3.6Å for **12**, and 3.5 and 3.6Å for **13**.

Interaction energies computed by various methods (Table 4.1) evidence significant attractive forces between the polar, heteroallene guest-representative molecules (methyl azide, methyl isocyanate, methyl isothiocyanate) and the host-representative molecule (methylenediurea), and weaker attractive forces for the nonpolar propadiene-containing

complexes. The comparison of SAPT energy decompositions reveals that the favorability of the azide-containing complexes is due to more than just dispersion. While MP2 tends to predict stronger binding with larger interaction energies than SAPT2+3, when the MP2 calculation is performed with the PCM solvent, the strength of the interaction is reduced. For the polar heteroallenes, this is consistent with the expectation that dipolar interactions will be weaker in a high dielectric solvent, like water.

Table 4.1: Interaction energies and energy decomposition of truncated host-guest complexes, all expressed in kcal/mol. Results are provided for the geometries of each end of each guest, as their geometries are somewhat different; the one with the shortest guest-host distance is reported first in each case.

	Total Interaction Energy			SAPT2+3 Decomposition			
Guest	MP2	MP2-PCM	SAPT2+3	elst	exch	ind	disp
15	-5.6	-2.9	-4.9	-2.9	3.6	-1.3	-4.3
	-5.3	-3.2	-4.3	-1.7	2.3	-0.8	-4.0
16	-5.1	-2.3	-5.0	-2.9	3.0	-1.2	-3.9
	-4.9	-2.7	-4.5	-2.1	2.3	-0.8	-3.8
17	-6.9	-3.1	-6.4	-3.6	2.9	-1.4	-4.3
	-7.3	-3.6	-6.7	-3.8	2.8	-1.1	-4.6
18	-2.6	-1.3	-1.8	-0.1	4.2	-1.4	-4.5
	-4.6	-3.0	-3.6	-1.9	4.7	-1.3	-5.1

The SAPT2+3 energy decompositions offer further insight regarding the attractive host-guest interactions. While the largest attractive component for all guests is dispersion, the electrostatic component is stronger in all heteroatom-containing functional groups than in the propadiene analog. This is congruent with the fact that the azide, isocyanate and isothiocyanate are polar, while the propadiene is nonpolar. The induction energy component is comparably small for all guests, indicating that mutual polarizing effects only have a minor influence on the overall stabilizing energies. The exchange term, which includes exchange-induction and exchange-dispersion effects, measures repulsion, and is stronger for the propadiene than for the polar functional groups. Thus, in the SAPT decomposition,

the weaker repulsion and stronger electrostatic attraction between those groups and the host account for their overall greater attraction relative to that of propadiene.

The role of electrostatics is further elucidated by the natural atomic charges computed for all guest fragments in complex with methylenediurea (Table 4.2). The structures correspond to those used in Table 1, and two sets of charges are listed, as the fragments adopt slightly different geometries at the two portals of the host. While the methyl propadiene has partial charges less than $0.11e$ at each position, the azide, isocyanate, and isothiocyanate analogs have partial charges at the α and β positions whose magnitudes are greater than $0.3e$. The substantial localization of positive charge at the β position in all three heteroallenes is consistent with a favorable electrostatic interaction with the negative charge on the nearby carbonyl oxygen of the host.

We also considered whether the attractive interactions between the polar guest groups and the host carbonyl might result in part from $n\rightarrow\pi^*$ interactions[84], which are characterized by the delocalization of a lone pair (n) of a donor group, typically a heteroatom nucleophile, into an antibonding orbital (π^*) of an acceptor group, typically a carbonyl group[29]. In the host-guest systems, we would expect delocalization of a lone pair of the ureidocarbonyl oxygen atom donor in the host to the antibonding orbital of an acceptor in the guest functional group. However, the systems studied here do not demonstrate the characteristic out-of-plane bending that results from attractive $n\rightarrow\pi^*$ interactions[28].

We used two criteria to check whether $n\rightarrow\pi^*$ interactions might play a role in these stabilizing interactions. First, recognizing that such interactions would make the α-β-γ angle, $\theta_{\alpha\beta\gamma}$, deviate from linearity, we carred out PM6-DH+ geometry optimizations in implicit solvent (COSMO model) for the various guests free in solution, and compared the resulting structures with the optimized host-guest structures. We observed no significant host-

Table 4.2: Natural atomic charges of guest functional groups, expressed in *e*.

Guest	Natural Atomic Charge					
	α		β		γ	
15	N	-0.4297 -0.4336	N	0.3224 0.3088	N	-0.1218 -0.1022
16	N	-0.6573 -0.6547	C	1.0407 1.0290	O	-0.6430 -0.6348
17	N	-0.5052 -0.5112	C	0.4292 0.4176	S	-0.2318 -0.2167
18	CH	-0.0219 -0.0286	N	0.0889 0.0744	CH_2	-0.1081 -0.0856

induced bending of either azide or the other analogs, as $\theta_{\alpha\beta\gamma}$ changed by at most 0.7° for the polar groups and 4.6° for the propadiene on going from solvent to the bound state (Table A.6). This result is consistent with the near linearity of the azide groups in the X-ray structures of **2**-**6** in complex with **1**. Second, we used NBO calculations to look fordono-acceptor interactions between carbonyl oxygen lone pairs of the host and antibonding π orbitals of the guest-representative fragments. Initial calculations on formamide-formaldehyde complexes previously studied in the Raines group[122] served to validate the present approach, as the NBO results confirmed the interaction of the oxygen lone pair of the formaldehyde donor with the antibonding orbital of the C=O acceptor in formamide (Table A.3). In contrast, NBO analysis of the solvent-optimized methylenediurea systems complexed with the truncated guest molecules indicates no significant $n\rightarrow\pi^*$ interaction, as detailed in Figure A.4, which compares the donor-acceptor interactions between carbonyl oxygen lone pairs of the host and antibonding π orbitals across the various guest-representative fragments. In particular, no $n\rightarrow\pi^*$ delocalizations above 0.07 kcal/mol were recorded. Thus, the present results indicate that $n\rightarrow\pi^*$ interactions do not contribute significantly to the attractive interactions studied here.

Altogether, these computational results point to a substantial azide-carbonyl attrac-

tion, which is attributable largely to dispersion and electrostatics interactions, and which goes beyond the weaker, primarily dispersive, attraction of a simple propadiene group for the host. The electrostatic component of the interaction traces largely to localization of positive charge on the β-nitrogen of the azide, and has the character of an orthogonal dipole interaction[129]. The calculations do not support a significant role for $n\rightarrow \pi^*$ interactions. Analogous calculations for two other heteroallenes, isocyanate and isothiocyanate, suggest that these groups can interact with the host in much the same way as azide.

4.3 Conclusions

This study reports the discovery of a remarkable attractive interaction between organic azides and the portal carbonyls of cucurbiturils. Since this yet unexploited interaction could be more broadly useful as a driver of supramolecular assembly, we investigated it using a set of homologous bis-α,ω-azidoethylammonium alkanes. The interactions between these molecules and cucurbit[6]uril were studied by NMR, IR, ITC, X-ray crystallography and by computational methods. The results indicate that the attractive azide-carbonyl interaction is a general phenomenon that can be exploited for supramolecular applications in the cucurbituril family and other systems. In addition, computational studies indicate that the interaction is not limited to azides, but generalizes to other isoelectronic heteroallenes, such as isocyanate and isothiocyanate. Further computational analysis points to electrostatics as the main driver for this interaction; in particular $n\rightarrow \pi^*$) delocalization does not play a significant role. Further studies with other functional groups are currently under way in our laboratories.

4.4 Acknowledgements

Chapter 4, in full, is currently being prepared for submission for publication of the material, Reany, Ofer; Li, Amanda; Yefet, Maayan; Gilson, Michael K.; Keinan, Ehud. The dissertation author was the secondary investigator and author of this paper.

Chapter 5

Calculation of Relative Binding Enthalpies for Constrained and Flexible Ligands of the Grb2 SH2 Domain

5.1 Introduction

The growth factor receptor protein 2 (Grb2) is an adapter protien that promotes cellular signal transduction[112]. The *Src* homology 2 (SH2) domain of Grb2 binds phosphotyrosine-containing proteins with a pYXNX recognition sequence. Thus, the Grb2 SH2 domain also binds synthetic phosphotyrosine-containing peptides of the same sequence. Matched pairs of flexible and constrained pYXN ligands for the Grb2 SH2 domain have been previously studied experimentally by Stephen Martin's group, with the goal of elucidating the thermodynamic consequences of ligand preorganization [39]. In general, preorganization of a ligand into its bound conformation may be expected to make the binding free energy more favorable, by reducing the entropic penalty on binding. Unexpectedly, although Martin

and coworkers did find that the constrained ligands bound with higher affinity, the binding entropies were less favorable and the enthalpies were more favorable.

To further investigate the thermodynamics of this system, we used molecular dynamics (MD) simulations of a series of protein-ligand complexes and the corresponding unbound ligands, under conditions mimicking those at which the experiments were done, to estimate the relative binding enthalpies of a series of constrained and unconstrained peptides for this protein. While related simulation studies using the AMOEBA polarizable force field have previously been published[157], the present study is distinct, as we use much longer simulations, 20-120 μs, with a more traditional, fixed-charge potential function, and we estimate binding enthalpies by a direct method [66], rather than via binding free energy simulations at multiple temperatures. With the multiple-GPU version of PMEMD, and long (4 fs) time steps made possible by hydrogen mass repartitioning [73], we were able to achieve up to 450 ns of simulation per day. This study thus presses the state of the art in protein-ligand simulations accessible with GPU computing, and the results are informative regarding the convergence of enthalpies by the direct method.

5.2 Methods

5.2.1 Calculation of Relative Binding Enthalpies

Relative binding enthalpies ($\Delta\Delta H$) were estimated by the direct method which only requires the mean potential energy of converged simulations. For the systems studied, $\Delta\Delta H$ are calculated to describe the energetic consequences of mutating of the central residue from valine to glutamine or incorporating conformational constraints to the phosphotyrosine. The relative binding enthalpies considered in this study are computed as in Equations 5.1-5.4,

where $\langle U_{PL} \rangle$ is the mean potential energy for the protein-ligand complex, and $\langle U_L \rangle$ refers to the same for the unbound ligand. A shorthand is used for the ligands so that fV refers to fpYVN, cV refers to cpYVN, and so on.

$$\Delta\Delta H_{cV-fV} = (\langle U_{PL,cV} \rangle - \langle U_{L,cV} \rangle) - (\langle U_{PL,fV} \rangle - \langle U_{L,fV} \rangle) \tag{5.1}$$

$$\Delta\Delta H_{cQ-fQ} = (\langle U_{PL,cQ} \rangle - \langle U_{L,cQ} \rangle) - (\langle U_{PL,fQ} \rangle - \langle U_{L,fQ} \rangle) \tag{5.2}$$

$$\Delta\Delta H_{fQ-fV} = (\langle U_{PL,fQ} \rangle - \langle U_{L,fQ} \rangle) - (\langle U_{PL,fV} \rangle - \langle U_{L,fV} \rangle) \tag{5.3}$$

$$\Delta\Delta H_{cQ-cV} = (\langle U_{PL,cQ} \rangle - \langle U_{L,cQ} \rangle) - (\langle U_{PL,cV} \rangle - \langle U_{L,cV} \rangle) \tag{5.4}$$

Since the force field energy terms are additive, one can also decompose relative binding enthalpies by structural components. For examples, the ligand structure alone can be isolated from both the complex and unbound ligand simulations, without any solvent molecules, to determine the energy differences that are specific to the ligand.

5.2.2 Molecular Dynamics Simulations

Relative binding enthalpies for the Grb2 SH2 domain and pYXN ligands were estimated with molecular dynamics (MD) simulations designed to mimick the experimental ITC conditions. All pYXN ligands are pseudopeptides which contain a varied central amino acid (X) flanked by a modified phosphotyrosine (pY) and an asparagine (N) (see Figure 5.1). For each flexible ligand structure (fpYXN, Fig. 5.1a), a constrained analog is also compared

(cpYXN, Fig. 5.1b). The cpYXN ligands are conformationally constrained by cyclization of the phosphotyrosine to form a cyclopropane ring. Here, we examine two pairs of flexible and constrained ligands, where each pair contains the same central residue, either valine (V) or glutamine (Q). The structures of these ligands in complex with the Grb2 SH2 domain were previously solved citeDeLorbe2009 and are available in the RSCB PDB[18]: fpYNVN (PDB ID: 3C7I) and cpYVN (2HUW); fpYQN (3IMD) and cpYQN (3IN7).

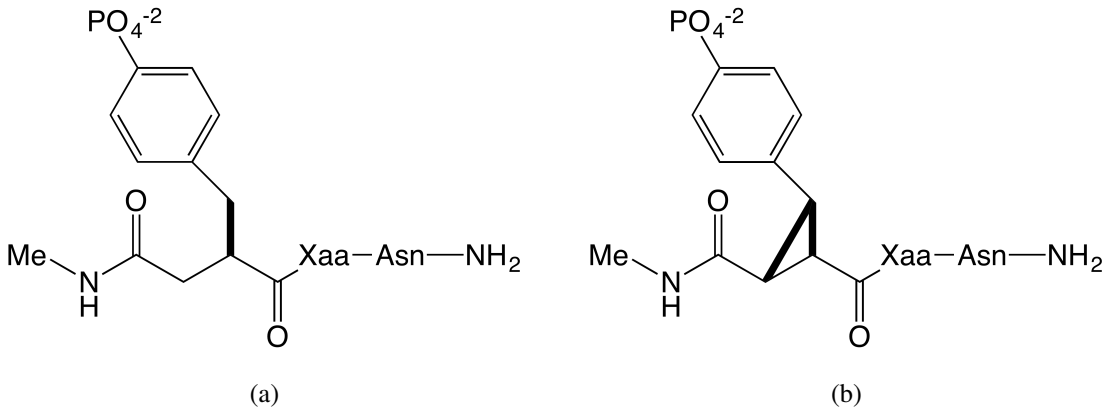


Figure 5.1: Flexible (a) and constrained (b) ligand structures. (a) fpYXN (b) cpYXN, where Xaa (X) is either valine (V) or glutamine (Q).

Each of the 4 ligands was simulated in complex with the Grb2 SH2 domain as well as free in solution (without the protein). The starting coordinates were prepared from the crystal structures of the complexes using Maestro[2] to remove waters beyond 5Å and add missing hydrogens. For the free ligand simulations, the coordinates of the ligand alone were extracted from the crystal structure. The complex (with the retained crystal waters) or the free ligand was then solvated with TIP3P water, buffer and ions to approximate the experimental conditions[39]. To facilitate the calculation of relative binding enthalpies, the contents of each simulation box are kept identical with the exception of the solute (either the complex or free ligand). Thus, each truncated octahedral simulation box (measuring

12Å from the solute to the box edge for the complex, and 23Å for the free ligand), was populated with 1 solute, 6154 waters, 6 HEPES molecules and 17 NaCl. Additional Na^+ and Cl^- ions were added for neutralization, dependent on the charge of the solute. The charges on the residues of Grb2 SH2 and the protonations of its histidines at pH 7.45 were determined using the H++ 3.0 server (http://biophysics.cs.vt.edu)[7, 121, 56]. Based on the *pKa* values predicted by MarvinSketch 16.3.7.0, 2016, ChemAxon (www.chemaxon.com), three different ionization states of HEPES were included (see Appendix A).

The RESP charges for both the HEPES molecules and the modified phosphotyrosine residues were determined using the R.E.D. Server[177, 45] and Gaussian09 C.01[51]. The forcefield parameters for the modified phosphotyrosine (PTY) were taken from the set determined by Steinbrecher et al[161, 71], available in the AMBER frcmod.phosaa10 file. The cyclized phosphotyrosine (CPY) parameters were the same as PTY with the exception of the cyclopropyl moeity, which used the GAFF parameters for sp^3 carbons in triangle systems (cx). The complete parameter sets are made available in the supplementary material. The systems were prepared by the LEaP program using the ff12SB force field.

The MD simulations were performed by using the multiple GPU version of PMEMD (pmemd.cuda.MPI) available in the AMBER simulation package[25]. The systems were first NVT heated to 300K and then NPT equilibrated for 5ns. The equilibrated coordinates were then used as the initial coordinates for the production simulations. The production simulations were performed using periodic boundary conditions with a nonbonded cutoff of 9Å. The SHAKE algorithm was used to constrain the bond lengths containing hydrogen atoms. Constant pressure and temperature were regulated by a Monte Carlo barostat and a Langevin thermostat. Hydrogen mass repartitioning was enabled for longer timesteps (4 fs). A prior study showed that this approach does not lead to a statistically significant change in

computed binding enthalpies [66]. The simulations were performed in 200 ns blocks, with each block seeded by a new random number. Coordinates and energies were recorded every 500 steps (2 ps).

For each system, two replicate simulations were initiated using the same equilibrated starting coordinates, but the trajectories are non-identical due to different random seeds. Each replicate was simulated for over 20 μs for the free ligands and over 120 μs for the complexes, so that the total simulation time was over 40 μs for each free ligand and over 240 μs for each complex.

5.2.3 Evaluation of Uncertainty

There are many approaches to estimating the uncertainties associated with calculating quantities that rely on means of the potential energies, such the binding enthalpy. Here we apply two of the methods previously detailed in [66]. Using the approach described in [158], the statistical inefficiency is determined from the autocorrelation function to create subsampled data series that is uncorrelated, at least in principle. The standard error of the mean (SEM,σ) is then computed for the resulting uncorrelated series. Blocking analysis [50] is another approach to estimating the uncertainty of the time series of potential energies, where block-wise SEMs are computed for successively longer blocks of energies. On a plot of block size versus SEM, a plateau is generally seen for simulations that are considered converged, and the SEM value corresponding to the plateau is taken as the error of the estimation. For a more conservative (i.e., larger) error estimate, one may also read off the largest SEM reached for any of the block sizes tested [66].

5.2.4 Structural Decomposition of Trajectories

Since the total energies of the simulations include the effects of water, buffer and ions, it is of interest to examine the energies specific to relevant substructures, such as the ligand or the binding site. These can be calculated by isolating, or decomposing, the component structures from the trajectories. We anticipated that the fluctuations of the solvent molecules would be larger than those of the components, and that the energies of the components would converge more rapidly than those of the total system.

Both complex and free ligand trajectories were processed using cpptraj[25] to generate decomposed trajectories. To generate trajectories containing only the ligand, the protein and all solvent molecules, including water, HEPES buffer and ions, were stripped. Similarly, to generate trajectories containing only 13 binding site residues (Arg13, Arg32, Ser34, Glu35, Ser36, Ser42, Val51, Gln52, His53, Phe54, Lys55, Leu66, Trp67) with and without the ligand, the rest of the protein and all solvent molecules were stripped. The potential energies of the decomposed systems were then evaluated by specifying *imin*=5 and *maxcyc*=1 in sander[25] to read in the trajectories and calculate a single-point energy at each frame. PME was disabled by using *ntb*=0 so that no periodicity is applied and long-ranged interactions were accounted for instead by increasing the nonbonded cutoff to 100.0Å.

5.2.5 Principal Component Analysis

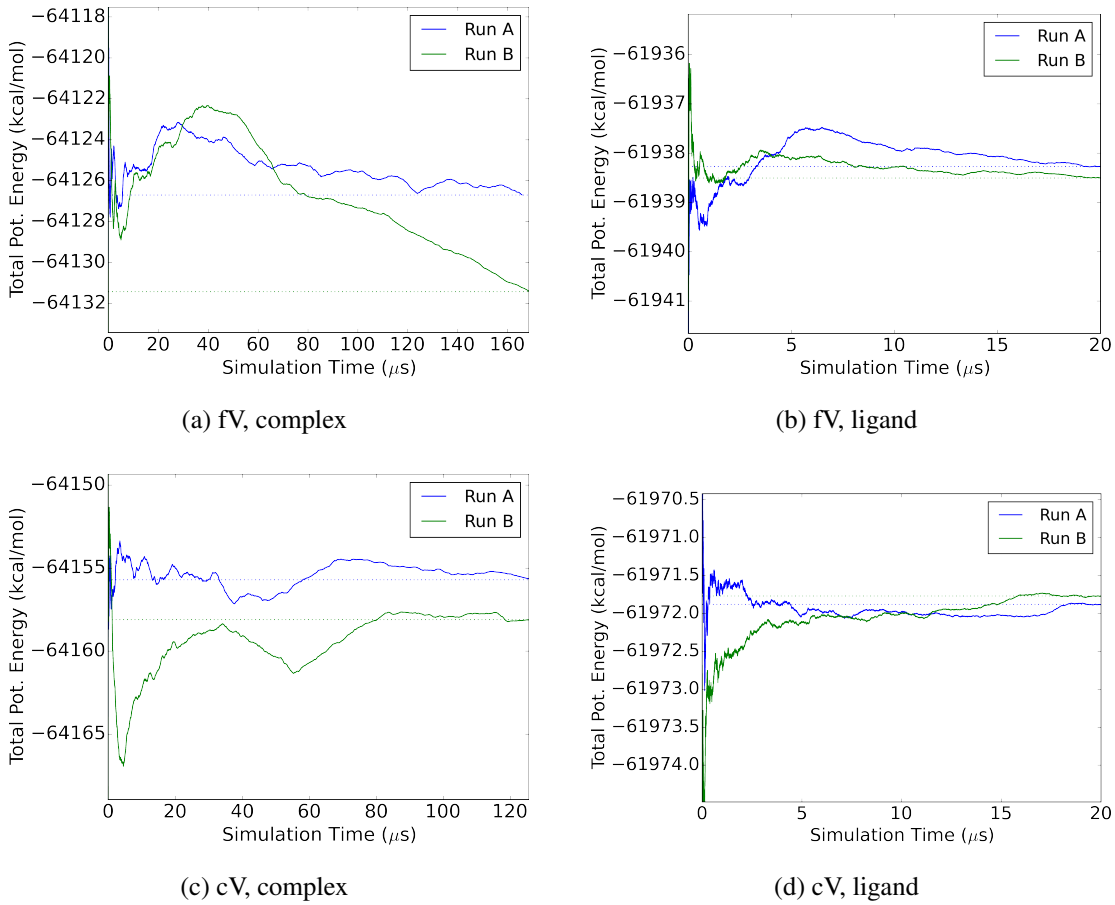
In order to look for large-scale, slow protein motions that might account for convergence issues, we applied principal component analysis (PCA) to the simulation trajectories. PCA is commonly used to determine the essential dynamics of a simulation [6, 65] by reducing the dimensionality of the trajectory motions. Principle components, or PCs, are the eigenvectors obtained from diagonalizing the covariance matrix of a trajectory. We used

the cpptraj program to obtain the first three PCs for combined trajectories of the simulation replicates (Run A and Run B) and then to project the individual trajectories onto each PC. Only the Cα atoms of the protein and all atoms of the ligand were considered.

5.3 Results

5.3.1 Mean Potential Energies

The cumulative and running means of the total potential energies of the systems are plotted in Figures 5.2-5.4). For the free ligand simulations, the plots of the cumulative means of both runs converge to values within 0.5 kcal/mol of each other by roughly 4,000,000 frames (8μs). For the complex simulations, there is a difference of at least 2.0 kcal/mol between the final means of the runs.



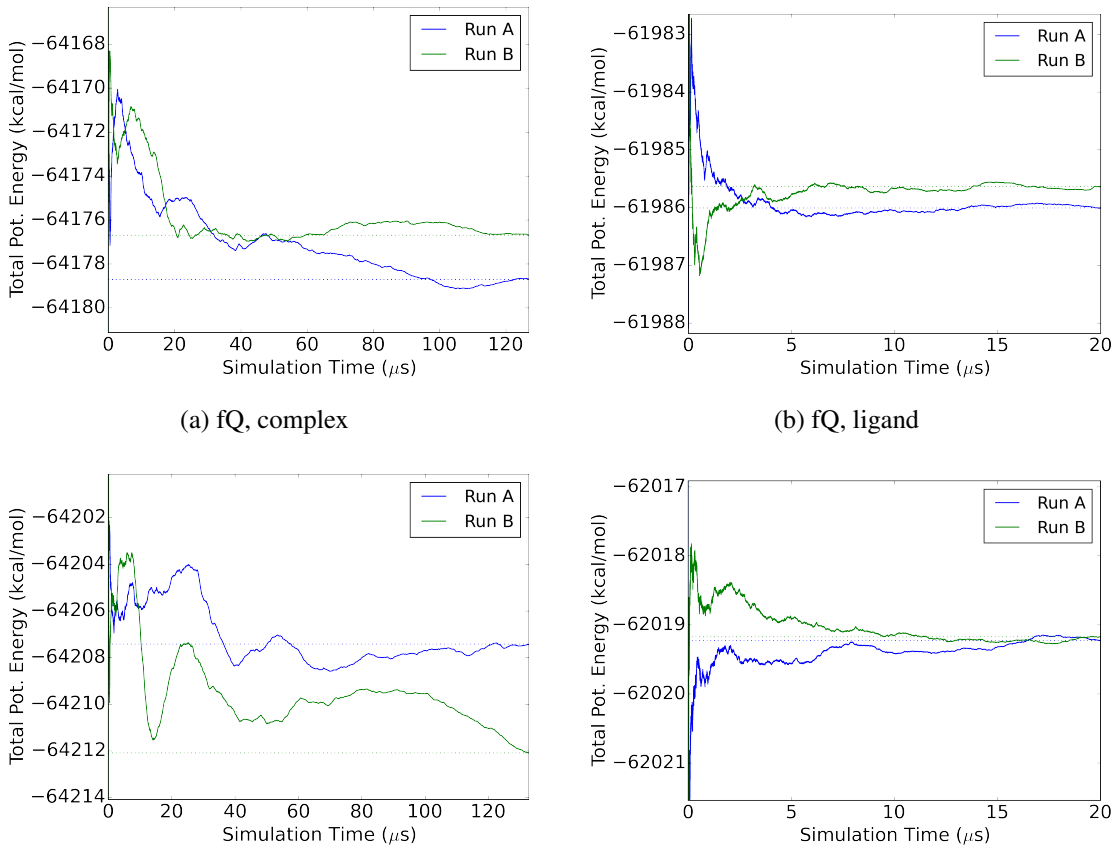
(a) fV, complex

(b) fV, ligand

(c) cV, complex

(d) cV, ligand

Figure 5.2: Cumulative means of total potential energy for fV and cV. The dotted line indicates the mean, calculated across all frames.



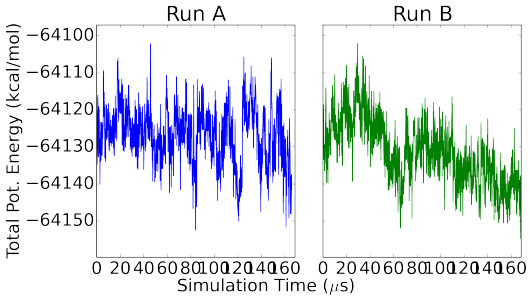
(a) fQ, complex

(b) fQ, ligand

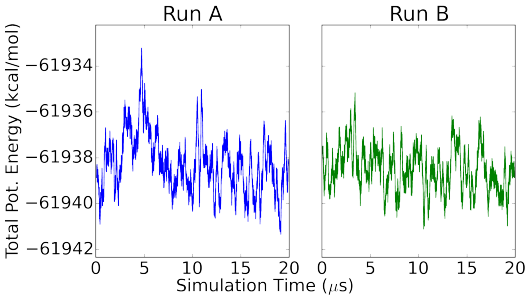
(c) cQ, complex

(d) cQ, ligand

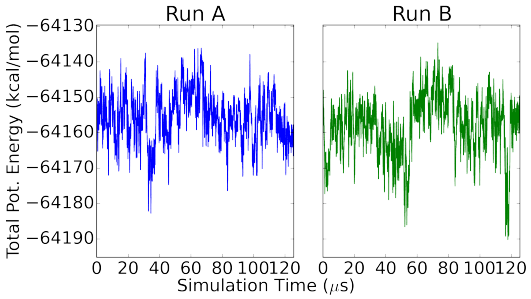
Figure 5.3: Cumulative means of total potential energy for fQ and cQ. The dotted line indicates the mean, calculated across all frames.



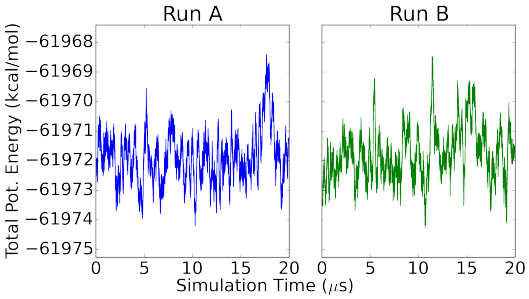
(a) fV, complex



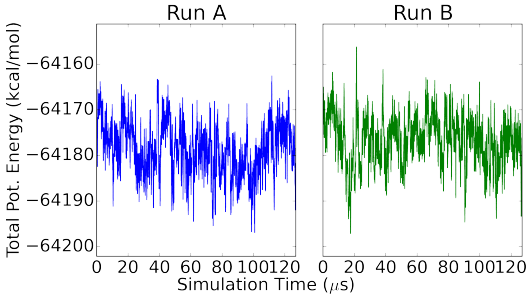
(b) fV, ligand



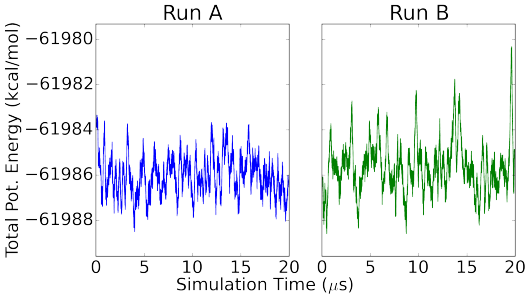
(c) cV, complex



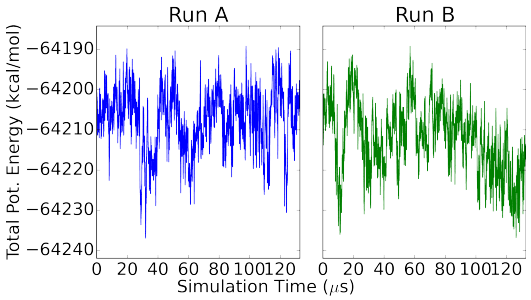
(d) cV, ligand



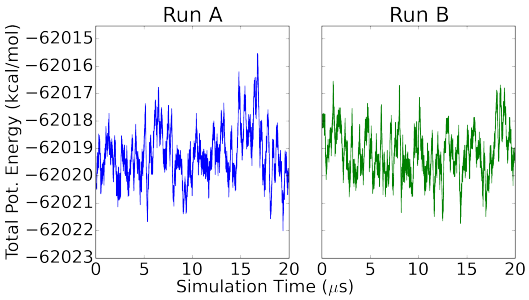
(e) fQ, complex



(f) fQ, ligand



(g) cQ, complex



(h) cQ, ligand

Figure 5.4: Running means of total potential energy, calculated using a window size of 10,000 frames.

5.3.2 Relative Binding Enthalpies

The relative binding enthalpies $\Delta\Delta H$ are calculated from the mean potential energies of the simulations (Equations 5.1-5.4). The relative enthalpy differences are evaluated for pairs of flexible and constrained ligands and for pairs of ligands with different central residues, and compared to experimental ITC results in Table 5.1.

Table 5.1: Relative Enthalpies ($\Delta\Delta H$, kcal/mol). The average of the four possible differences between parallel runs listed for each of the differences. For example, the mean $\Delta\Delta H_{cV-fV}$ is the average of $\Delta\Delta H_{cV,A-fV,A}$, $\Delta\Delta H_{cV,A-fV,B}\Delta\Delta H_{cV,B-fV,A}$, and $\Delta\Delta H_{cV,B-fV,B}$. The SEMs are propagated from the SEMs estimated using the statistical inefficiency for each run; the experimental errors are taken from [39], which are propagated from the errors in ligand concentration

	Calculated		Experimental [39]	
	Mean $\Delta\Delta H$	SEM σ	ITC	Error
cV - fV	5.59	2.63	-2.5	0.32
cQ - fQ	1.34	1.47	-1.1	0.30
fQ - fV	-1.42	2.25	-3.3	0.27
cQ - cV	-5.67	2.00	-1.9	0.35

The computed relative binding enthalpies do not agree well with experiment: the root mean square error across all four values is 4.72 kcal/mol, and the Pearson correlation coefficient is -0.03. Also, the large uncertainties in the computed results, relative to the differences of interest, reduce the significance of these comparisons. Nonetheless, some interesting patterns emerge from the calculations. First, although experiments saw that constrained ligands had more favorable binding enthalpies than their flexible counterparts, the simulated results indicate the opposite, even accounting for the deviations between the matched A and B runs, and for the numerically estimated errors (see next section). The enthalpic changes associated with mutating the central residue also differ between the

computational and experimental estimates. While both experiment and computation indicate that ligands with glutamine have more favorable binding enthalpies than those with valine, the experimental measurements indicate that glutamine leads to more favorable binding for the flexible ligands, while the computational results indicate the opposite.

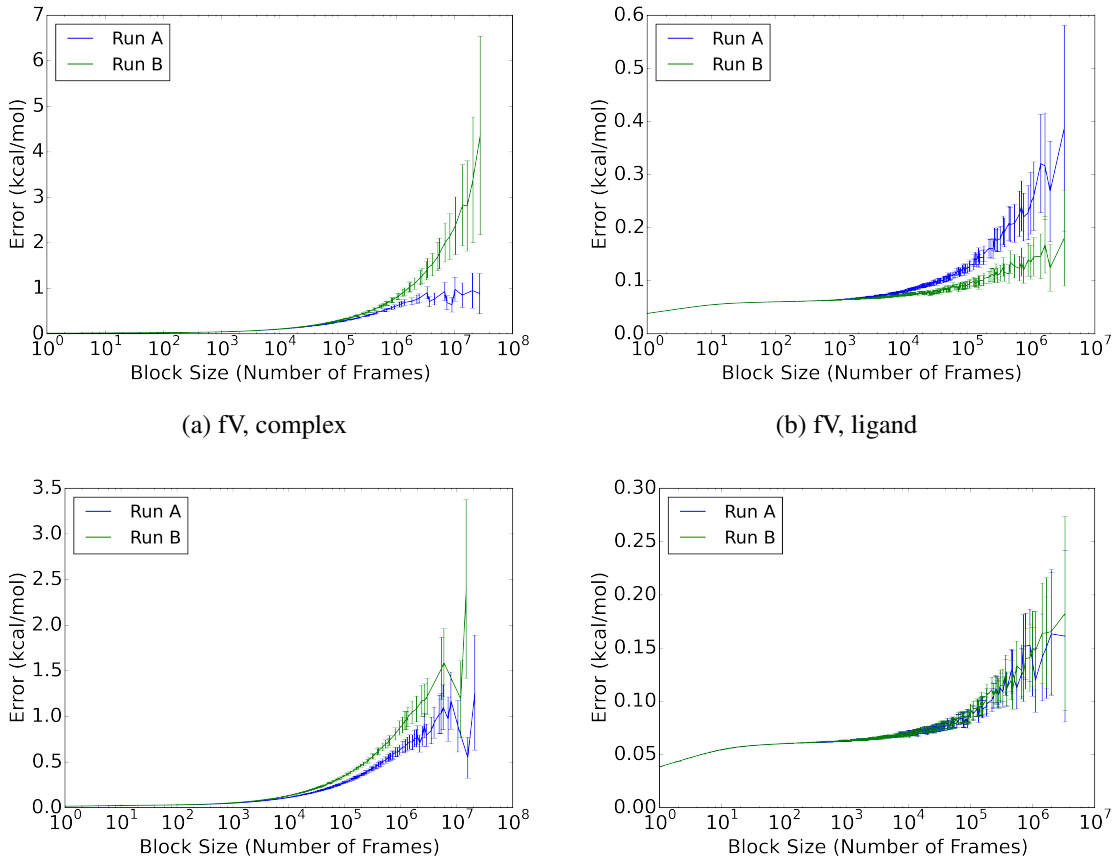
5.3.3 Evaluation of Uncertainty

When considering the estimated binding enthalpies, it is important to also evaluate the uncertainty of the estimate. Since the binding enthalpies are estimated using the means of the total potential energies from a simulation, it is useful to examine the standard error of the mean (SEM, σ) to quantify the uncertainty. The estimates of the SEM based on the statistical inefficiency[158] are listed in Table 5.2. The SEMs of the complex simulations are much greater than those of the free ligand simulations, and whereas the replicate runs of the complex simulations have different SEMs, those of the free ligand simulations have nearly identical SEM values. Altogether, this provides evidence that the simulations of the free ligands are converged to within approximately 0.06 kcal/mol error, while the simulations of the complexes are not well enough converged to support a detailed comparison with experiment, with errors up to 3.0 kcal/mol.

Table 5.2: SEMs based on statistical inefficiency (kcal/mol).

		SEM σ	
Ligand	Run	Complex	Free Ligand
fV	A	0.8571	0.0576
	B	2.9865	0.0577
cV	A	0.7833	0.0603
	B	1.8932	0.0615
fQ	A	0.5417	0.0591
	B	0.4248	0.0596
cQ	A	1.0936	0.0608
	B	1.6178	0.0607

The statistical uncertainties were also evaluated by blocking analysis (Figures 5.5 and 5.6). Plateaus are observed in the curves for all the simulations of the ligand alone (right column), but not in the simulations of the full complex (left column). Congruent with the statistical inefficiency results, the errors for the ligand simulations are determined to be no more than 0.07 kcal/mol, and there is notably good agreement between the parallel runs. In the absence of a plateau on the blocking curves, it is difficult to quantify the errors for the complex simulations.



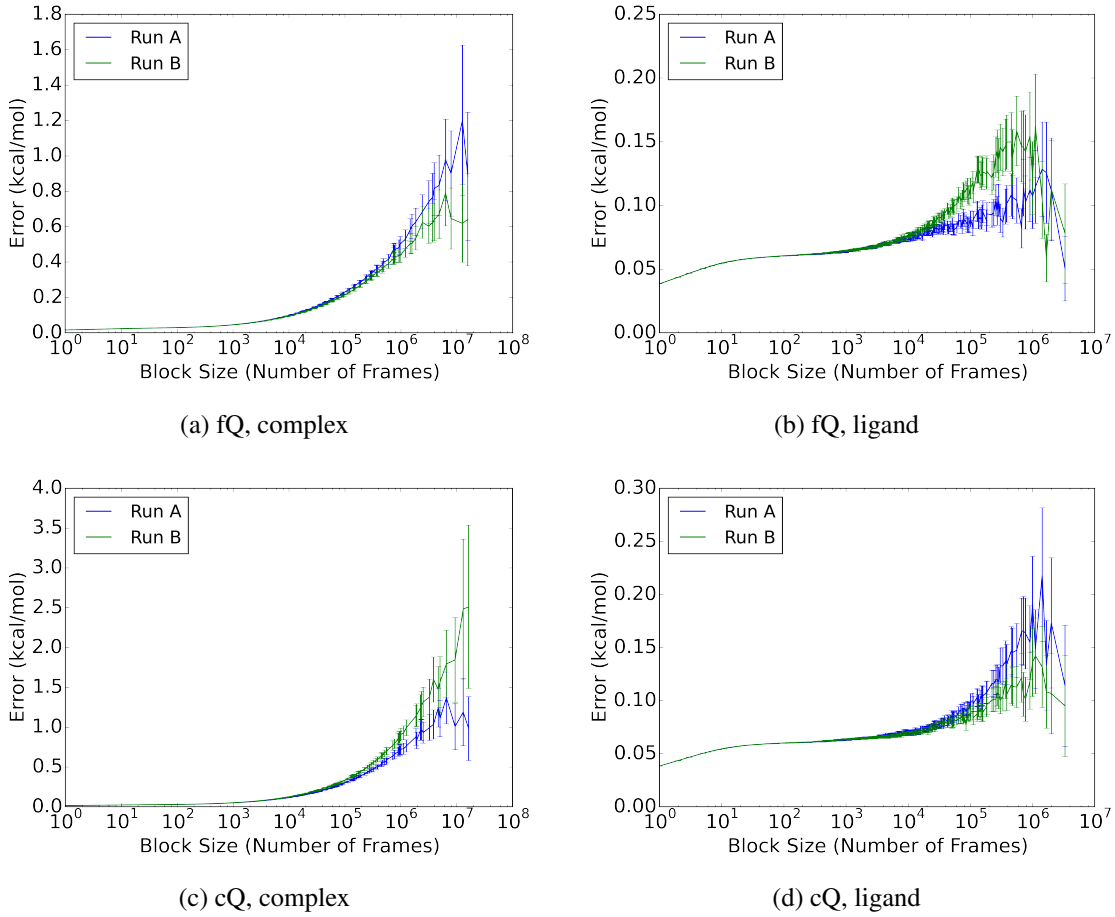
(a) fV, complex

(b) fV, ligand

(c) cV, complex

(d) cV, ligand

Figure 5.5: SEM blocking curves for fV and cV total potential energies.



(a) fQ, complex

(b) fQ, ligand

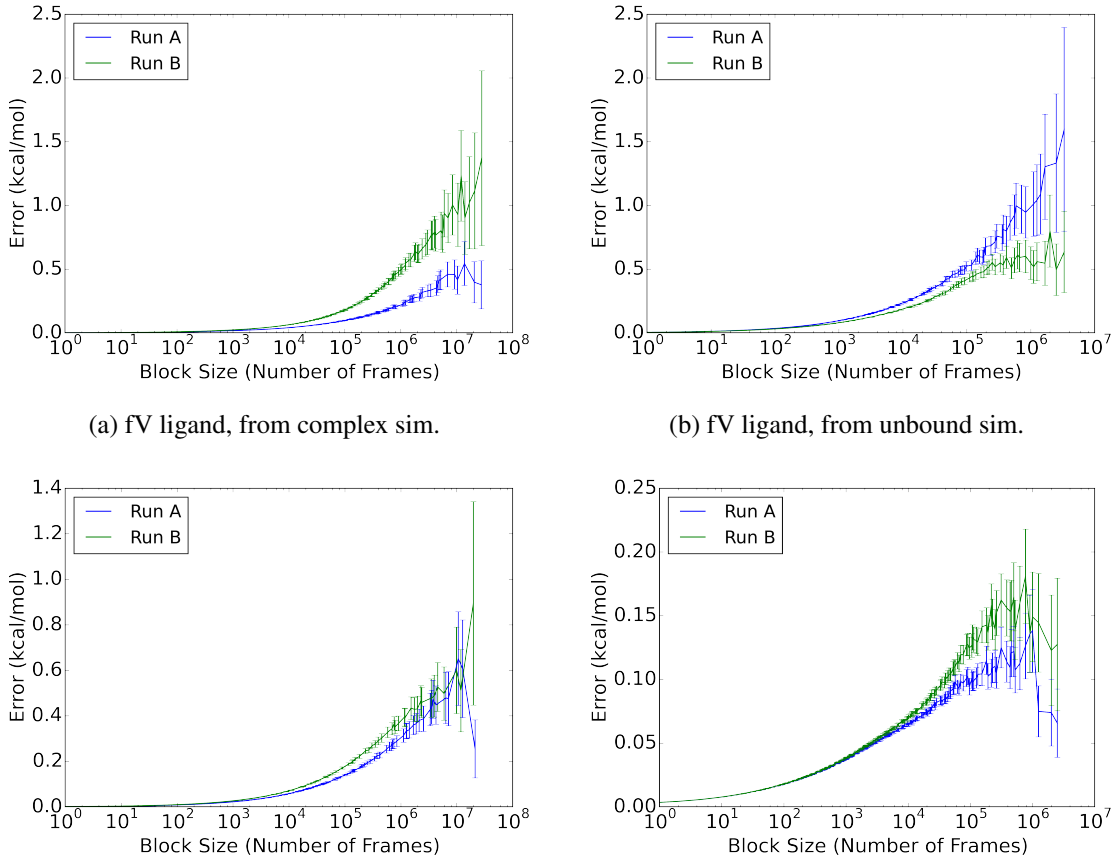
(c) cQ, complex

(d) cQ, ligand

Figure 5.6: SEM blocking curves for fQ and cQ total potential energies.

We conjectured that the mean potential energies of parts of these large systems would converge more rapidly than the total energies, and that these component energies would be informative regarding the mechanistic determinants of the computed relative binding enthalpies. Thus, blocking analysis was also applied to the energies of decomposed systems of the ligand alone (Figures 5.7 and 5.8) and binding residues with and without the ligand (Figure 5.9 and 5.10). For the energy components, plateaus are only seen for the free cQ ligand trajectories (Figures 5.5 and 5.6); in all other cases, the errors for the component terms appear to be larger than those for the total potential energies of the same systems. The blocking analysis of the decomposed trajectories of 13 binding site residues

with and without the ligand also demonstrate larger errors for the component energies than for the total energies of the solvated systems. Altogether the larger errors of the component energies indicate that these actually experience greater fluctuations (see Table 5.3) and are less converged than the energies of the full systems that include water, buffer and ions. This result suggests that there is strong anticorrelation of energies among the various components of each system. For example, large fluctuations of the internal energy of the free ligands are effectively opposed by anticorrelation fluctuations of the ligand-solvent and solvent-solvent energies.



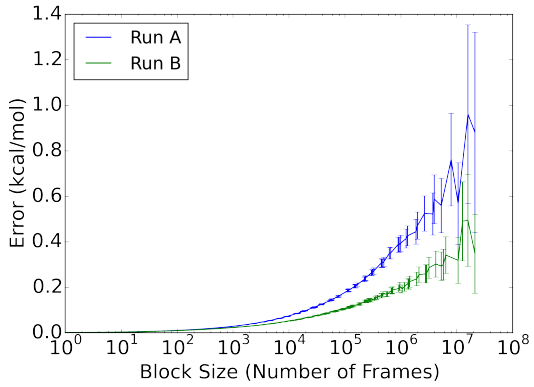
(a) fV ligand, from complex sim.

(b) fV ligand, from unbound sim.

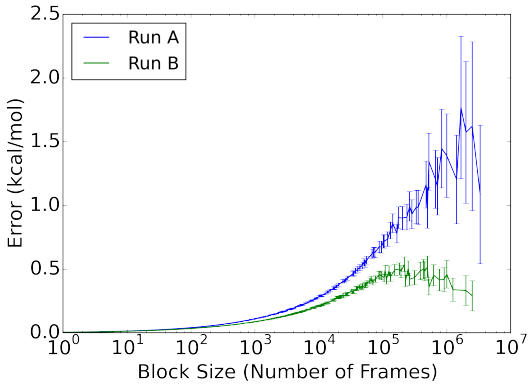
(c) cV ligand, from complex sim.

(d) cV ligand, from unbound sim.

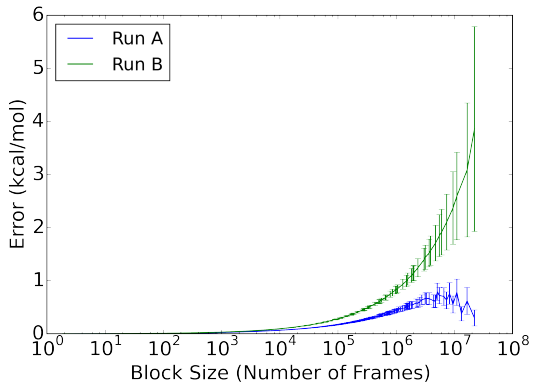
Figure 5.7: SEM blocking curves of the potential energies of the fV and cV ligands, decomposed from simulations of the complex and of the unbound ligand.



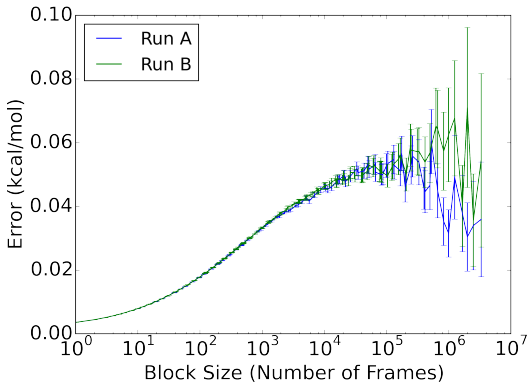
(a) fQ ligand, from complex sim.



(b) fQ ligand, from unbound sim.

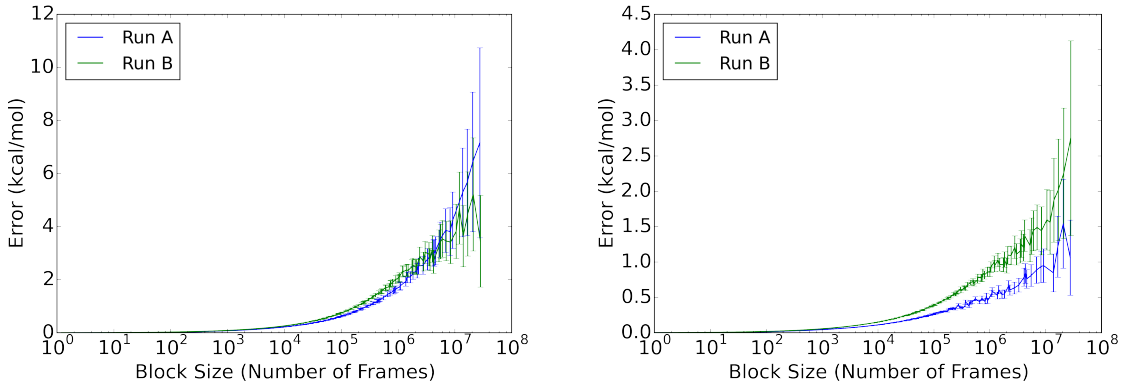


(c) cQ ligand, from complex sim.



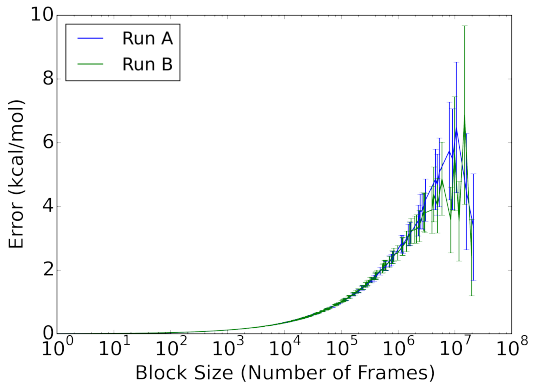
(d) cQ ligand, from unbound sim.

Figure 5.8: SEM blocking curves of the potential energies of the fQ and cQ ligands, decomposed from simulations of the complex and of the unbound ligand.

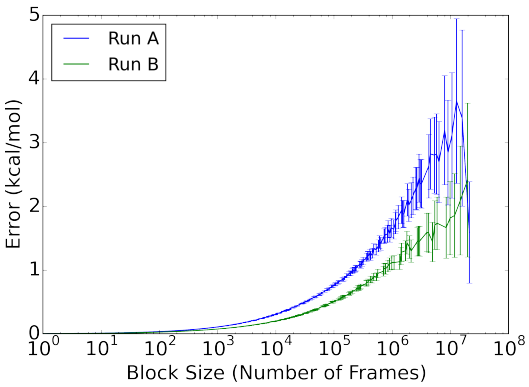


(a) fV binding site complex

(b) fV binding site alone

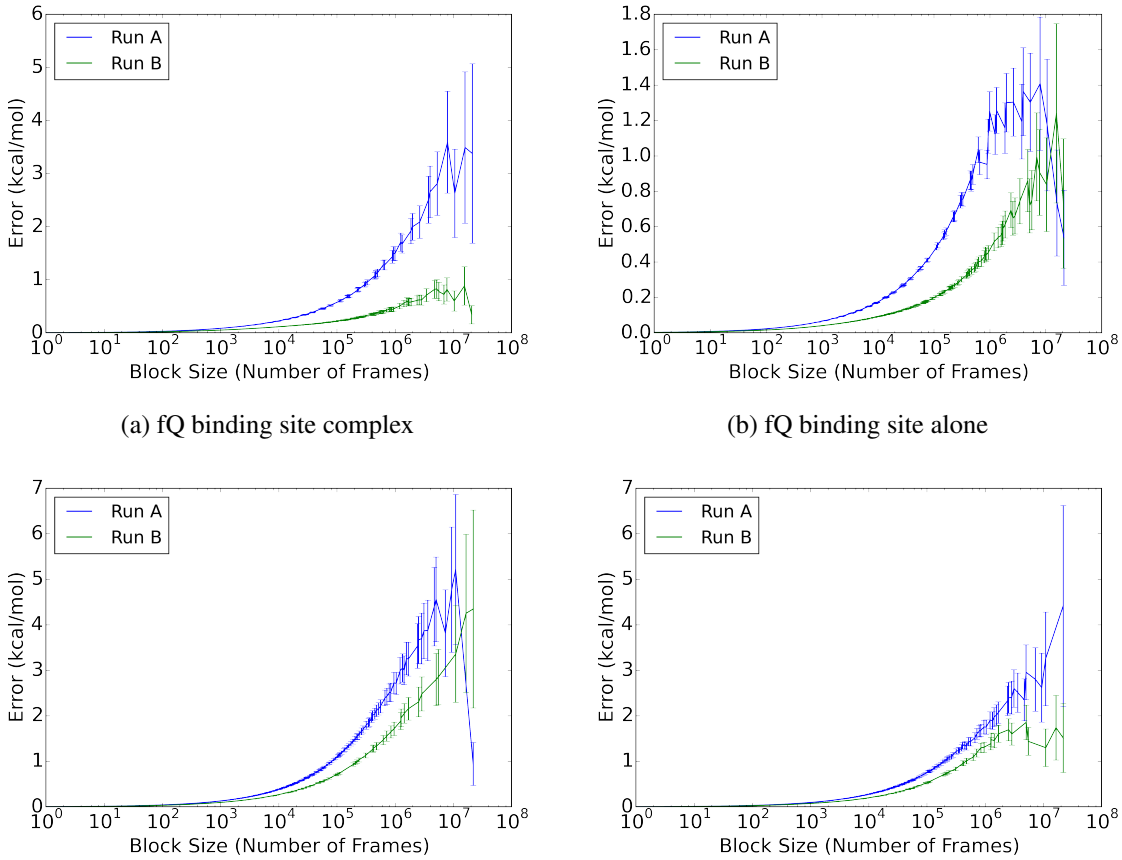


(c) cV binding site complex



(d) cV binding site alone

Figure 5.9: SEM blocking curves for the potential energies of the binding site with and without ligand, for fV and cV, decomposed from simulations of the complex



(a) fQ binding site complex

(b) fQ binding site alone

(c) cQ binding site complex

(d) cQ binding site alone

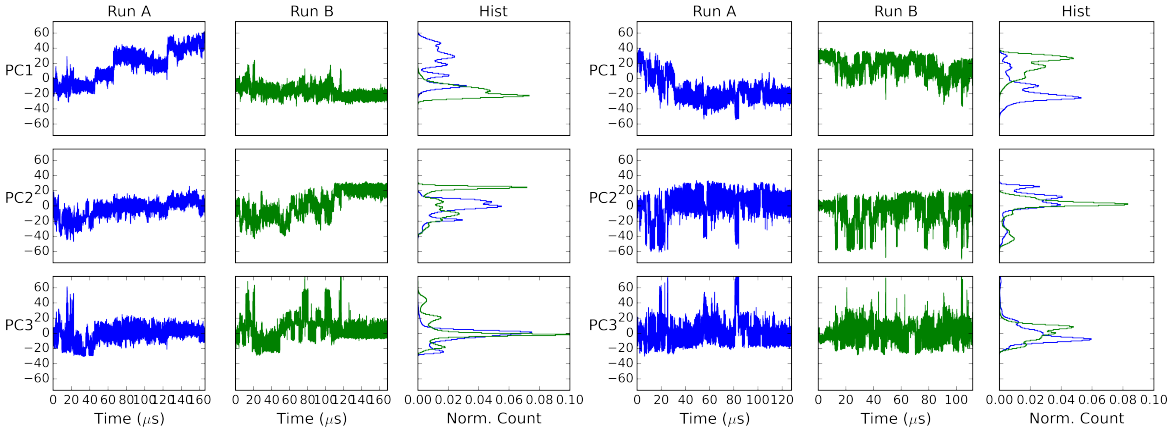
Figure 5.10: SEM blocking curves for the potential energies of the binding site with and without ligand, for fQ and cQ, decomposed from simulations of the complex

Table 5.3: SEMs based on statistical inefficiency for component energies (kcal/mol).

Ligand	Run	SEM σ from Complex	SEM σ from Free Ligand
fV	A	0.4465	1.0908
	B	1.0566	0.5949
cV	A	0.9606	0.1059
	B	0.5161	0.1558
fQ	A	0.6525	1.3140
	B	0.2831	0.5319
cQ	A	0.7362	0.0490
	B	2.9349	0.0496

5.3.4 Principal Component Analysis

In light of the significantly different mean potential energies for parallel runs of the same complex simulation, and the large errors and lack of plateaus in the blocking curves, we used principal component analysis (PCA) to look for slow motions which might contribute to the slow convergence of the simulations. The first three principal components (PCs) were determined for the combined parallel simulations (Run A and Run B). The projection of each individual run on each PC for the complex and ligand simulations as well as their normalized histograms are shown in Figures 5.11 and 5.12



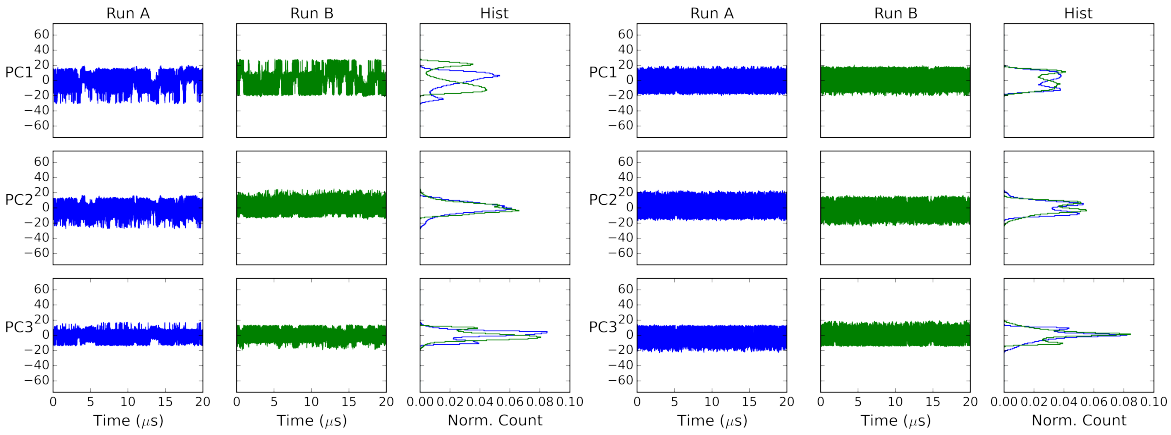
(a) fV, complex

(b) cV, complex

(c) fQ, complex

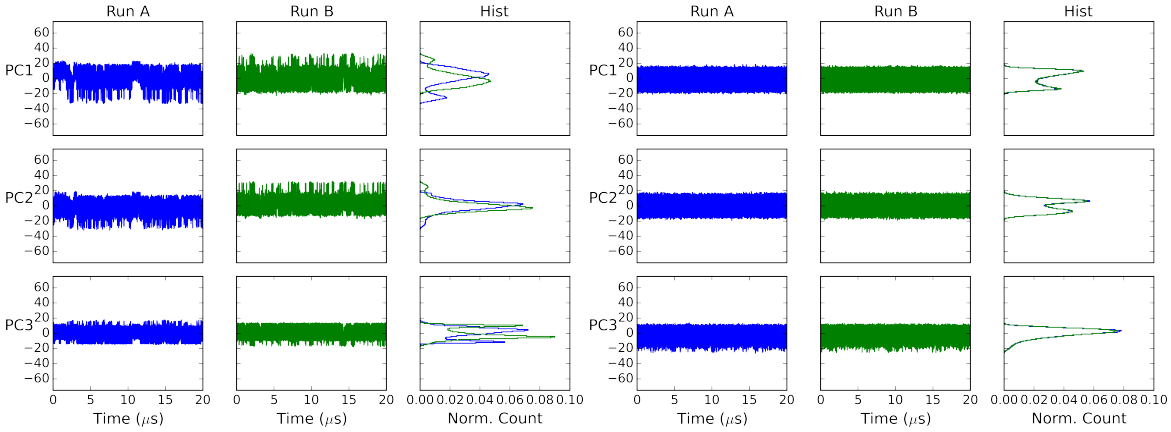
(d) cQ, complex

Figure 5.11: PC projections for complex simulations



(a) fV, ligand

(b) cV, ligand



(c) fQ, ligand

(d) cQ, ligand

Figure 5.12: PC projections for free ligand simulations

The histograms of the projections for the complex simulations (right column of Fig. 5.11) illustrate that the two matched runs explore largely non-overlapping regions of the leading PCs. In contrast, for the ligand simulations, the histograms of the projections for the parallel runs (right column of Fig. 5.12) have more overlap, indicating that the motions along the PC are represented equally in each runs. The lack of overlap in the PCs observed for the complex trajectories is consistent with previous observations of poorly sampled sub-nanosecond MD simulation of proteins[13, 33] and points to relatively large-scale, slow,

motions as a limiting factor in the slow convergence of the mean potential energies of the protein-ligand complexes.

5.4 Conclusions

Although experimental results[39] indicate that the constrained Grb2 SH2 ligands studied have more favorable binding enthalpies than their flexible analogs, our simulations point in the opposite direction. However, since the mean energies of the complexes are not well converged, these results could change with longer simulation time. Although the simulated complex and unbound ligand systems are of similar size ($\sim$20,000 atoms for the complexes and $\sim$18,000 atoms for the unbound ligands), the unbound ligand simulations yield mean energies that converge to fractions of a kcal/mol within $\sim$20 μs while the complex simulations are converged to within 0.4-3.0 kcal/mol after over $\sim$120 μs. This observation suggests that the slow convergence of the complexes does not result solely from the large number of degrees of freedom, but also from the fact that the protein undergoes slow motions that the free ligands and water do not. This idea is supported by the lack of convergence of the leading principal components of motion of the complexes. Future directions of this study will include further investigation to better characterize the dynamics which may be responsible for the slow convergence of the protein simulations.

It is also interesting to note that the estimated errors for the decomposed trajectories of just the ligand or just the selected binding site residues are greater than those of the whole system. The fact that the fluctuations of the component energies are greater than those of the total energies implies that there is strong anticorrelation among component terms. Thus, it is more difficult than anticipated to informatively dissect the differences in overall binding enthalpies into components.

While the current simulations of the complexes are not sufficiently converged to allow clear comparisons with experimental results, the estimated errors are within 3 kcal/mol, which is only 0.004% of the total energy. This is, arguably, an impressive and ultimately encouraging result, and we would anticipate future attempts with faster MD implementations to be more successful. It is also of interest to ask whether better convergence might be reached by alternative methods. In particular, the van't Hoff relation, and related expressions, allow the binding enthalpy to be obtained by analysis of the variation of binding free energy as a function of temperature. Such approaches have the advantage of not requiring tight convergence of the full system potential energy. On the other hand, computing the numerical derivative of the binding free energy with respect to temperature requires very tight convergence of the free energy, and this poses its own challenges.

Chapter 6

Conclusions and Future Directions

The research described in the prior chapters addresses a range of issues relating to noncovalent interactions between molecules. One thing that comes across in all chapters is that there is no definitive way to model any of the biochemical phenomena examined. There is a "zoo" of different QM methods for computing interaction energies, a variety of ways to model polarization, and many different approaches to characterizing binding. Common to all the studies presented is the use of approximations in the interest of allowing calculations to be more tractable within computer resource and time constraints. Thus, measuring the error of these approximations is important, especially for understanding the tradeoffs between accuracy and computation speed.

For a given system, there should only be one correct value for a physical quantity of interest, such as an interaction energy or electrostatic potential, and an accurate calculation should be able to reproduce it. However, computational chemistry seems to be saturated with approximate methods of comparable accuracy. In Chapter 2, we see that the DFT methods with dispersion corrections and the linearly-scaled SAPT0 methods all have similar accuracies. In Chapter 3, we show that using different response models for inducible

dipoles produces only slight differences in error, and note that the use of different screening functions did not significantly change our results. Thus, the choice between methods or models, especially amongst those with similar accuracies, is usually discerned by their computational efficiency. For example, in Chapter 2, the linearly-scaled SAPT0 is twice as fast as the DFT methods, and in Chapter 3, direct polarization requires less calculation time than self-consistent polarization.

Overall, the evaluations of QM methods and polarization models in Chapters 2 and 3 yield informative considerations for modeling noncovalent interactions. A fast and accurate QM method may be used for parameterizing force fields. One promising application of the linearly-scaled SAPT0 method would be for fitting parameters of the physically-motivated force field introduced by McDaniel and Schmidt[111]. Chapter 3 reveals that the inducible dipole models found in popular polarizable force fields improve upon traditional force fields which model polarization implicitly, but fall short of the theoretical maximum accuracy achievable for both point charge and point dipole representation. These results imply that polarizable force fields would benefit from improved response models for both these representations. The future direction of this work would endeavor to develop such response models that could be readily implemented into a force field for simulation. This could be done by adding a term to the force field potential energy equation that would generate optimal point charges or optimal point dipoles, likely dependent on changing electric field. Either a mathematical model or a lookup table would be needed to map from the electric field to the parameters.

The characterization of binding between two molecules also has diverse set of approaches. Binding is a complex process, resulting from many noncovalent interactions and molecular motions acting together. In Chapters 4 and 5 a few different approaches are applied

to specific systems, including equilibrium quantum mechanical calculations and dynamic molecular mechanical simulations. We note that there are many other ways to assess binding, but the comparison of these is beyond the scope of this dissertation. The QM calculations used in Chapter 4 support that there is an attractive interaction between heteroallene moieties in guests and carbonyl groups in cucurbituril portals. The SAPT2+3 decompositions and NBO analyses provide further detail regarding the dominance of electrostatic and dispersion components in the favorable interaction. On the other hand, since the systems examined were optimized equilibrium geometries, these calculations give no sense of how those interactions would behave over time.

In contrast, in Chapter 5, the time-varying interactions and motions of the Grb2 SH2 domain are simulated using molecular dynamics, whereas the finer details of the noncovalent interactions are too computationally costly and not explicitly calculated. As with all MD, these details are coarse-grained by the parameters for force field terms, but the ability to simulate molecular motions over time allows the calculation of thermodynamic quantities such as binding free energies, entropies or enthalpies. There are different ways to estimate relevant thermodynamic quantities from simulations, but Chapter 5 focuses on the direct approach for calculating relative binding enthalpies from mean potential energies.

Since the protein-ligand simulations of Chapter 5 were not well-converged, future work will involve additional analyses to understand and characterize the slow convergence. While the principal component analyses demonstrated that replicate runs of the complex simulations exhibit different essential dynamics, they do not specify which motions are responsible for the slow convergence. To answer this question, we plan to investigate representative structures along the principal components and perform more structural analyses on the simulations. Additionally, the structural decomposition of the entire Grb2 SH2 protein

should be interesting to look at, since previously we had only decomposed selected residues of the binding site. The component energies of the protein would then be calculated and compared to the principal component projections. We do anticipate, though, as computers become more powerful and simulations become more efficient, that the full convergence of the complex simulations will be possible in the next few years.

In this field of molecular modeling, we are always looking to make artful approximations that capture physics as well as possible within current computational constraints. Continual growth in computing power promises inevitable improvements for the whole field over time, but the choice of best approximation will always be necessary. In addition to technological progress, the development of next generation models for simulation, such as more accurate polarizable force fields, should also advance the field. Ultimately, advances in computational chemistry theory, software and hardware will eventually enable more adept manipulation of biochemical system, as in drug design and protein engineering.

Appendix A

A.1 Supplementary Information for Chapter 2

Table A.1: Definitions of SAPT truncations. The subscripts "elst," "exch", "ind", and "disp" refer to electrostatics, exchange, induction, and dispersion. The subscript "resp" denotes that the orbital response of the perturbed system is taken into account. The superscripts refer to the orders of the interaction and Møller-Plesset fluctuation potential operators. A complete explanation of the SAPT theory and terms is available from Szalewicz[79].

SAPT0	$E_{SAPT0} = E_{elst}^{(10)} + E_{exch}^{(10)} + E_{ind,resp}^{(20)} + E_{exch\ ind,resp}^{(20)} + E_{disp}^{(20)} + E_{exch\ disp}^{(20)}$
SAPT2	$E_{SAPT2} = E_{SAPT0} + E_{elst,resp}^{(12)} + E_{exch}^{(11)} + E_{exch}^{(12)} + {}^{t}E_{ind}^{(22)} + {}^{t}E_{exch\ ind}^{(22)}$
SAPT2+	$E_{SAPT2+} = E_{SAPT2} + E_{disp}^{(21)} + E_{disp}^{(22)}$
SAPT2+(3)	$E_{SAPT2+(3)} = E_{SAPT2+} + E_{elst,resp}^{(13)} + E_{disp}^{(30)}$
SAPT2+3	$E_{SAPT2+3} = E_{SAPT2+(3)} + E_{exch\ disp}^{(30)} + E_{ind\ disp}^{(30)} + E_{exch\ ind\ disp}^{(30)}$

Table A.2: Evaluation of PMx methods with and without halogen corrections across various dissociation separations. Errors are presented as RMSE values, in kcal/mol.

	X40x10			**X31x10**		
	PM6-DH2	*PM6-DH2X*	*PM7*	*PM6-DH2*	*PM6-DH2X*	*PM7*
0.80	9.57	22.61	6.46	7.81	6.61	4.77
0.85	6.10	5.87	5.12	5.51	3.91	3.88
0.90	4.01	3.02	4.19	4.10	3.13	3.29
0.95	2.97	2.51	3.52	3.23	2.71	2.83
1.00	2.46	2.30	3.02	2.65	2.39	2.44
1.05	2.18	2.14	2.66	2.23	2.12	2.12
1.10	1.99	2.00	2.39	1.91	1.88	1.84
1.25	1.61	1.62	1.81	1.29	1.29	1.24
1.50	1.09	1.09	1.14	0.69	0.69	0.65
2.00	0.45	0.45	0.51	0.22	0.22	0.26

Table A.3: Curve fitting results for scaling of calculation time with system size. Timings for the A24 dataset are fit to the equation $t = an^b$, where n is the number of atoms or electrons in each dimer.

	n = number of atoms			***n = number of electrons***		
	a	**b**	**R^2**	**a**	**b**	**R^2**
B3LYP	0.36	2.52	0.86	0.04	2.37	0.39
B97-D	0.42	2.44	0.86	0.05	2.26	0.39
M062X	0.40	2.54	0.88	0.07	2.23	0.35
ωB97X-D	3.13	1.65	0.67	0.82	1.51	0.29
SAPT0	1.46	1.49	0.81	0.22	1.57	0.47
SAPT2	1.02	1.96	0.83	0.09	2.06	0.47
SAPT2+	1.15	2.09	0.82	0.08	2.22	0.48
SAPT2+(3)	0.67	2.44	0.82	0.03	2.61	0.48

Table A.4: Comparison of SAPT0 scaled energy components with corresponding SAPT2+(3) energy components. Electrostatics: $E_{elst}^{(10)}$, $E_{elst,resp}^{(12)}$, $E_{elst,resp}^{(13)}$, Exchange: $E_{exch}^{(10)}$, $E_{exch}^{(11)}$, $E_{exch}^{(12)}$, Induction: $E_{ind,resp}^{(20)}$, $E_{exch-ind,resp}^{(20)}$, ${}^{t}E_{ind}^{(22)}$, ${}^{t}E_{exch-ind}^{(22)}$, Dispersion: $E_{disp}^{(20)}$, $E_{exch-disp}^{(20)}$, $E_{disp}^{(21)}$, $E_{disp}^{(12)}$, $E_{disp}^{(30)}$. Results are presented as a linear regression, $y = mx + b$.

	m	**b**	**R²**
fSAPT0(1)			
Electrostatics	1.04	-0.05	1.00
Exchange	0.95	-0.13	1.00
Induction	0.88	-0.05	0.99
Dispersion	0.82	0.05	0.97
fSAPT0(2)			
Dispersion	0.77	-0.02	0.95
fSAPT0(3)			
Dispersion	0.82	0.19	0.98

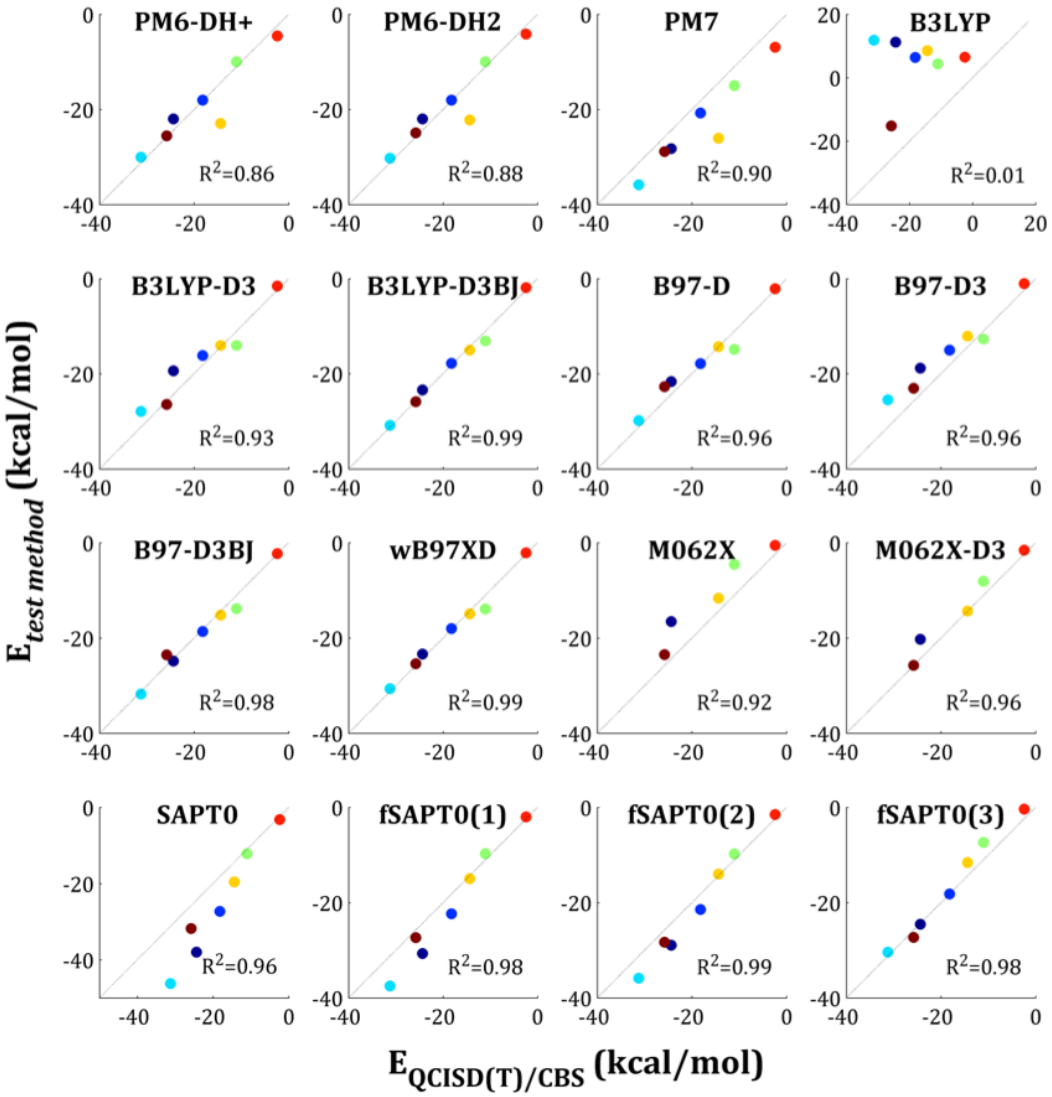


Figure A.1: Correlation of QM methods with reference results for complexes in L7 dataset. All reference energies were obtained using QCISD(T)/CBS, except for the guanine-cytosine···guanine-cytosine complex, for which the reference energy was obtained using CCSD(T)/CBS[152]. M062X and M062X-D3 calculations were not completed for the circumcoronene systems.

A.2 Supplementary Information for Chapter 4

A.2.1 Experimental details

General methods:

All reactions were carried out in anhydrous solvents under inert atmosphere. Starting materials including compound **7** and α,ω-alkanediamines (C_mDA, m=4-8) were purchased and used without further purification. CB[6] **1**, 2-azidoethylamine **8**, *N*-Boc-2-azidiamineethane **9**, *N,N'*-Di-Boc-bis-(2-azidoethyl)-1,6-diaminohexane, **10c**, *N,N'*-Bis-(2-azidoethyl)-hexane-1,6-diammonium chloride salt, **4** and **4@1** were prepared as described before.[1] Flash chromatography was performed on Merck silica gel 60 (230-400 mesh). ^{1}H and ^{13}C NMR spectra were recorded in the solvents indicated by using either AVIII400 Bruker spectrometer. Chemical shifts (δ) are given in ppm relative to TMS. The residual solvent signals were used as references and the chemical shifts were converted to the TMS scale: $CDCl_3$: δ_H = 7.26 ppm, δ_C = 77.0 ppm, D2O-DCl: (with trace DMSO) δ_H = 2.70 ppm, δ_C = 39.5 ppm. The following abbreviations or combinations thereof were used to explain the multiplicities: s = singlet, d = doublet, t = triplet, q = quartet, qu = quintet, m = multiplet, br = broad. Mass spectra were recorded by using either Waters MALDI microMX (TOF) or Waters LCT Premier microMax spectrometers (TOF-ESI, with MeCN/water, 1:1). Crystals of all five complexes, **2@1**, **3@1**, **4@1**, **5@1** and **6@1** were obtained by slow diffusion of iso-propanol into aqueous solutions of the complexes. The Single crystals were mounted on a Nonius KappaCCD diffractometer and data was collected using graphite monochromatized MoKα radiation (λ=0.71073) at 293 K (**2@1**, **5@1** and **6@1**), 150 K (**3@1**) or at 120 K (**4@1**). The following program was used for data collection and reduction: Nonius 1997 Collect,[2] HKL DENZO, and Scalepack.[3] The structures were solved by direct methods using the program package maXus[4] and refined in the usual way using SHELXL97.[5] Non-hydrogen atoms were refined anisotropically and hydrogen atoms isotropically.

General procedure for *N*-alkylation:

The synthesis was performed by addition of NaH (60% suspended in oil, 0.7g, 18 mmol) to solution of **9** (1.15g, 6.1 mmol) in DMF (25 mL) at 0 ºC for 1 h. Dibromide (1,4-butane, 1,5-pentane, 1,7-heptane and 1,8-octane) was added (2.8-3.0 mmol) and the mixture was stirred overnight at RT, then quenched with aq. NH_4Cl solution, extracted with Et_2O, washed with water and brine and dried over Na_2SO_4. Removal of the solvent followed by column chromatography afforded the desired products as colorless to pale yellow viscous oils.

***N,N'*- Di-Boc-bis-(2-azidoethyl)-butane-1,4-diammonium chloride (10a)**:

Product was isolated by column chromatography (hexane/EtOAc, 85:15) in 60% yield (1.05 g).

^{1}H NMR (400 MHz, $CDCl_3$): δ 3.42 (bs, 4H), 3.34 (bs, 4H), 3.25 (bs, 4H), 1.50 (bs, 4H), 1.46 (bs, 18H). **^{13}C NMR** (100 MHz, $CDCl_3$): δ 155.4, 80.0, 49.8, 48.3, 47.5, 46.8, 28.4, 25.5. **MS** (TOF-MS-ES$^+$): *m/z* calcd for $C_{18}H_{34}N_8O_4$ $[M]^+$: 426; found: 427 $[M+H]^+$.

***N,N'*- Di-Boc-bis-(2-azidoethyl)-pentane-1,5-diammonium chloride (10b)**:

Product was isolated by column chromatography (hexane/EtOAc, 85:15) in 60% yield (2 g).

^{1}H NMR (400 MHz, $CDCl_3$): δ 3.35 (bs, 8H), 3.24 (bs, 4H), 1.69 (bs, 4H), 1.46 (bs, 18H), 1.27 (bs, 2H).

***N,N′*- Di-Boc-bis-(2-azidoethyl)-heptane-1,7-diammonium chloride (10d)**:

Product was isolated by column chromatography (hexane/EtOAc, 9:1) in 80% yield (1.05 g).

^{1}H NMR (400 MHz, $CDCl_3$): δ 3.42 (bs, 4H), 3.34 (bs, 4H), 3.22 (bt, *J*=6.8 Hz, 4H), 1.50 (bs, 4H), 1.46 (bs, 18H), 1.25 (m, 6H). **^{13}C NMR** (100 MHz, $CDCl_3$): δ 155.4, 79.7, 49.6, 48.2, 46.6, 29.2, 28.5, 26.2. **HRMS** (TOF-MS-ES$^+$): *m/z* calcd for $C_{21}H_{41}N_8O_4$ $[M+H]^+$: 469.3251; found: 469.3257 $[M+H]^+$.

***N,N′*- Di-Boc-bis-(2-azidoethyl)-octane-1,8-diammonium chloride (10e)**:

Product was isolated by column chromatography (hexane/EtOAc, 95:5) in 80% yield (1.65 g).

^{1}H NMR (400 MHz, $CDCl_3$): δ 3.40 (bs, 4H), 3.34 (bs, 4H), 3.2 (bt, *J*=6.8 Hz, 4H), 1.60 (bs, 4H), 1.46 (bs, 26H), 1.29 (bs, 4H). **^{13}C NMR** (100 MHz, $CDCl_3$): δ 155.4, 79.7, 49.6, 48.2, 46.6, 29.2, 28.5, 26.6. **HRMS** (TOF-MS-AP$^+$): *m/z* calcd for $C_{22}H_{43}N_8O_4$ $[M+H]^+$: 483.3407; found: 483.3417 $[M+H]^+$.

General procedure for the removal of Boc protecting groups:

A mixture of each starting material, *i.e*, **10a-e** (1.2-2.0 mmol) in EtOH (35-50 mL) and 4N HCl (15-20 mL) was stirred overnight at RT. After removal of the solvent the residue was dissolved in hot MeOH (5-10 mL) and upon standing overnight the resultant precipitate was collected. If precipitation was not occur or resulted in a poor precipitate, Et_2O was added to the hot solution of methanol and resulted in precipitation upon cooling and afforded the desired products as off-white solids.

***N,N′*- Bis-(2-azidoethyl)-butane-1,4-diammonium chloride (2)**:

Product was isolated in 60% yield (0.25 g).

^{1}H NMR (400 MHz, D_2O): δ 3.74 (m, 4H), 3.23 (m, 4H), 3.11 (m, 4H), 1.77 (m, 4H). **^{13}C NMR** (100 MHz, D_2O): δ 47.6, 47.1, 23.4. **HRMS** (TOF-MS-ES$^+$): *m/z* calcd for $C_8H_{19}N_8$ $[M\text{-}HCL\text{-}Cl^-]^+$: 227.1733; found: 227.1721 $[M\text{-}HCl\text{-}Cl^-]^+$.

***N,N′*- Bis-(2-azidoethyl)-pentane-1,5-diammonium chloride (3)**:

Product was isolated in quantitative yield (0.5 g).

^{1}H NMR (400 MHz, D_2O) δ 3.74 (bt, *J*=5.6 Hz, 4H), 3.22 (bt, *J*=5.6 Hz, 4H), 3.07 (bt, *J*=8 Hz, 4H), 1.72 (bq, 4H), 1.44 (bq, 4H). **^{13}C NMR** (400 MHz, D_2O): δ 48.0, 47.6, 47.0, 25.7, 23.6. **HRMS** (TOF-MS-ES$^+$): *m/z* calcd for $C_9H_{21}N_8$ $[M\text{-}HCl\text{-}Cl^-]^+$: 241.1889; found: 241.1896.

***N,N′*- Bis-(2-azidoethyl)-heptane-1,7-diammonium chloride (5)**:

Product was isolated in 70% yield (0.5 g).

^{1}H NMR (400 MHz, D_2O): δ 3.75 (bt , *J*=5.6 Hz, 4H), 3.22 (bt , *J*=5.6 Hz , 4H), 3.06 (bt, *J*=8.0 Hz, 4H), 1.68 (bs, 4H), 1.37 (bs, 6H). **^{13}C NMR** (100 MHz, D_2O): δ 48.3, 47.6 46.9, 28.3, 26.1, 25.9. **HRMS** (TOF-MS-ES$^+$) m/z calcd for $C_{11}H_{25}N_8$ $[M\text{-}HCl\text{-}Cl^-]^+$: 269.2202; found: 269.2175.

***N,N′*- Bis-(2-azidoethyl)-octane-1,8-diammonium chloride (6)**:

Product was isolated in 70% yield (0.5 g).

^{1}H NMR (400 MHz, D_2O) δ 3.75 (bt, *J*=5.2 Hz, 4H), 3.21 (bt, *J*=5.2 Hz, 4H), 3.06 (bt, *J*=7.6 Hz, 4H), 1.66 (bs, 4H), 1.34 (bs, 8H). **^{13}C NMR** (100 MHz, D_2O): δ 48.4, 47.6, 46.9, 28.6, 26.2, 26.0. **HRMS** (TOF-MS-AP$^+$): m/z calcd for $C_{12}H_{27}N_8$ $[M\text{-}HCl\text{-}Cl^-]^+$: 283.2359; found: 283.2344.

General procedure for the complex formation of bis-(2-azodoethyl)-α,ω-alkanediammonium salts with 1:

To a stirred solution of selected guest (0.05 mmol) in water (5 mL) and conc. HCl (150 μL), **1** (0.1 mmol) was added and the solution was left stirring at RT overnight. The resulting mixture was filtered and few drops of co-solvent was added slowly to the filtrate to afford a single crystal of **2@1**, **3@1**, **5@1** and **6@1** within several days and then subjected to X-ray analysis.

2@1: Crystals were obtained by using *i*-PrOH as the co-solvent.

^{1}H NMR (400 MHz, D_2O-DCl): δ 5.73 (d, *J*=15.6, 12H), 5.66 (s, 12H), 4.39 (d, *J*=15.6, 12H), 3.94 (m, 4H), 3.44 (m, 4H) , 2.33 (m, 4H) , 0.57 (m, 4H). **^{13}C NMR** (100 MHz, D_2O- DCl): δ 157.4, 71.1, 52.4, 49.4, 48.2, 47.5, 24.5.

3@1: Crystals were obtained by using ethylenglycol as the co-solvent.

^{1}H NMR (400 MHz, D_2O-DCl): δ 5.85 (d, *J*=15.6, 12H), 5.74 (s, 12H), 4.48 (d, *J*=15.6, 12H), 4.08 (bt, *J*=5.6 Hz, 4H), 3.55 (bt, *J*=5.6 Hz, 4H), 2.84 (m, 4H), 0.82 (m, 4H), 0.5 (m, 4H). **^{13}C NMR** (100 MHz, D_2O- DCl): δ 156.4, 70.4, 47.4, 47.2, 46.9, 27.4, 22.7.

5@1: Crystals were obtained by using *i*-PrOH as the co-solvent.

^{1}H NMR (400 MHz, D_2O- DCl) δ 5.77 (d, *J*=15.6 Hz, 12H), 5.55 (s, 12H), 4.29 (d, *J*=15.6 Hz, 12H), 3.90 (bt, *J*=5.2, 4H), 3.36 (bt, *J*=5.2, 4H), 3.04 (bt, *J*=6.8, 4H), 1.16 (m, 4H), 0.48 (m, 4H), 0.4 (m, 2H). **^{13}C NMR** (100 MHz, D_2O-DCl): δ 156.1, 70.4, 51.4, 48.4, 47.4, 47.2, 30.2, 26.9, 26.2. **MS** (TOF-MS-ES$^+$) *m/z* (%): For $C_{47}H_{60}N_{32}O_{12}$ 1265.5 (100%), 1266.5 (54%), 1267.5 (21%); Found: $[MH]^+$ 1265.5 (100%), 1266.5 (52%), 1267.5 (18%); $[MH_2]^{+2}$ 633.2 (100%), 633.7 (60%), 634.2 (22%). **HR-MS** (TOF-MS-ES$^+$): calcd. for $C_{47}H_{61}N_{32}O_{12}$ 1265.5147, found 1265.5144.

6@1: Crystals were obtained by using *i*-PrOH as the co-solvent.

^{1}H NMR (400 MHz, D_2O-DCl) δ 5.66 (d, *J*=15.6 Hz, 12H), 5.43 (s, 12H), 4.18 (d, *J*=15.6 Hz, 12H), 3.78 (t, *J*=5.2, 4H), 3.27 (t, *J*=5.2, 4H), 3.06 (t, *J*=6.0, 4H), 1.33 (m, 4H), 0.54 (m, 4H), 0.35 (m, 4H). **^{13}C NMR** (100 MHz, D_2O-DCl): δ 155.9, 70.3, 51.4, 48.7, 47.4, 46.9, 30.3, 26.6, 25.6. **MS** (TOF-MS-ES$^+$) *m/z* (%): For $C_{48}H_{62}N_{32}O_{12}$ 1278.5 (100%), 1279.5 (53%), 1280.5 (16%); Found: $[MH]^+$ 1279.5 (100%), 1280.5 (64%), 1281.5 (26%);

Table A.5: Crystallographic data collection and structure refinement details of complexes of 1.

	2@1	**3@1**	**4@1**	**5@1**	**6@1**
Empirical formula	$C_{44}H_{52}N_{32}O_{12}$ *2HCl*4H_2O	$C_{46}H_{50}N_{32}O_{12}$ *2Cl*EG*8H_2O	$C_{46}H_{54}N_{32}O_{12}$ *2HCl*2EtOH	$C_{48}H_{52}N_{32}O_{12}$ *2Cl*7H_2O*2HO	$C_{48}H_{52}N_{32}O_{12}$ *2Cl*12H_2O
Formula weight	1366.16	1582.34	1556.39	1458.83	1566.39
Crystal system	Orthorhombic	Monoclinic	Orthorhombic	Monoclinic	Monoclinic
Space group	*Pbca*	$P2_1/c$	*P-1*	$P2_1/c$	$P2_1/c$
a (Å)	12.2810(6)	11.725(2)	11.3464(18)	24.246(5)	12.7070(3)
b (Å)	20.6410(10)	19.451(4)	12.0175(180)	13.201(3)	13.3900(3)
c (Å)	21.9930(10)	14.706(3)	13.007(2)	20.956(4)	20.3620(4)
α (°)	90	90	108.858(2)	90	90
β (°)	90	97.94(2)	97.857(3)	110.86(3)	103.1260(14)
γ (°)	90	90	103.185(3)	90	90
V (Å ^{3})	5575.1(54)	3321.7(11)	1590.9(4)	6268(3)	3374.1(13)
Z	4	2	1	4	2
$D_{calc.}$ (g×cm^{-3})	1.628	1.582	1.6258	1.546	1.542
T (K)	293(2)	153(2)	293(2)	293(2)	293(2)
F (000)	2848	1656	818	3060	1648
Crystal size (mm)	0.30×0.22×0.18	0.48×0.30×0.12	0.30×0.22×0.18	0.28×0.20×0.14	0.62×0.38×0.18
θ max (°)	25.05	25.03	26.35	24.75	24.93
Unique reflections	4398	5856	6403	10186	5776
$I > 2\sigma(I)$	5941	4105	5941	4456	4090
R-factor (all data)	0.0435	0.0741	0.0581	0.2001	0.0879
R-factor ($I > 2\sigma(I)$)	0.0389	0.0462	0.0550	0.1029	0.063
Goodness of fit	1.059	0.911	1.056	1.397	1.107

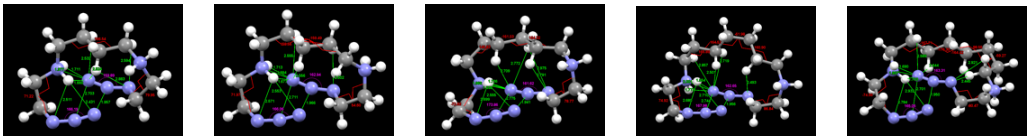


Figure A.2: Calculated structures of guests **2-6** (Geometry color codes: green, short contacts (< 3ÃĚ); magenta, bond angles; red, dihedral angles). Chloride ions were omitted for clarity.

Table A.6: Linearity of guest functional groups assesed by the $\theta_{\alpha\beta\gamma}$ bond angle, for PM6-DH+ geometry optimizations in implicit solvent (COSMO model) performed for the guest alone and in complex with the CB[6] host system.

Guest	Optimized $\theta_{\alpha\beta\gamma}$	
	free	in complex with host
5	169.1	168.4
	169.0	169.5
11	165.3	164.8
	165.2	165.4
12	175.6	176.0
	175.5	176.0
13	179.6	176.7
	178.6	174.0

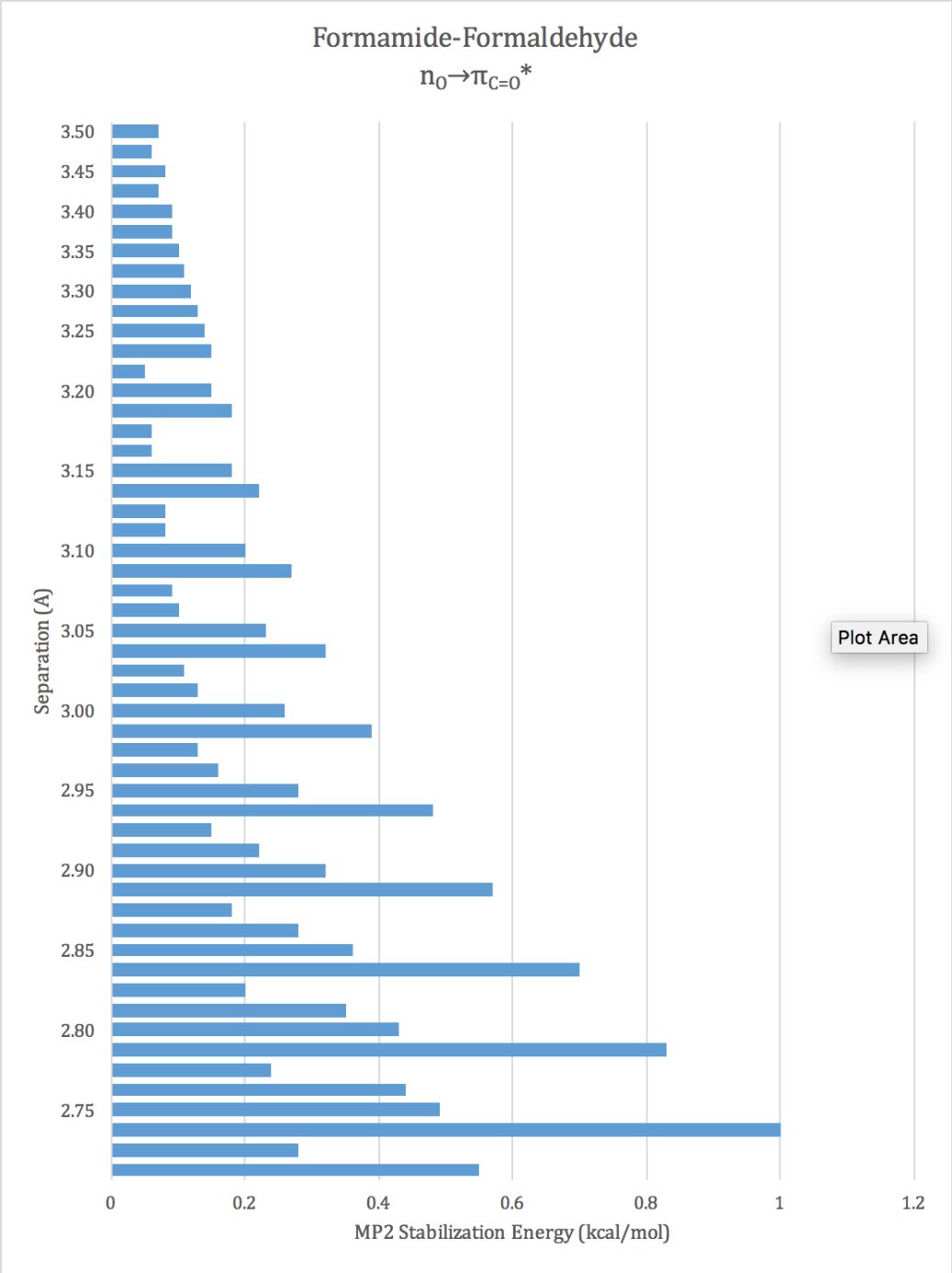


Figure A.3: MP2 stabilization energies for formaldehyde-formamide $n_O \rightarrow \pi_{A=B}^*$ delocalizations as a function of separation distance. Structures were obtained from the supporting information from [122].

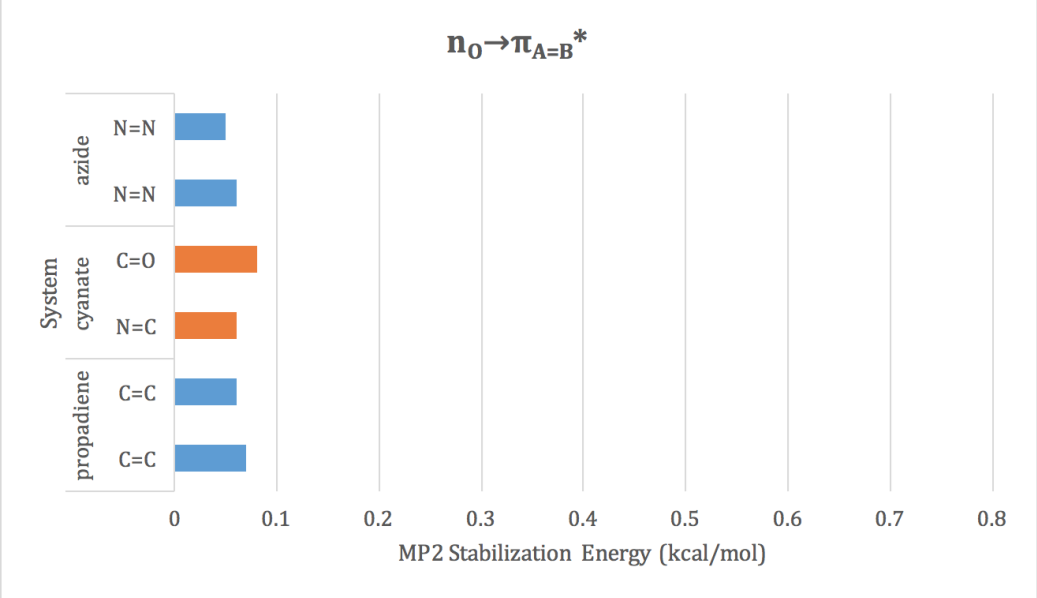


Figure A.4: MP2 stabilization energies for host-guest $n_O \rightarrow \pi_{A=B}^*$ delocalizations. No such delocalizations were recorded for isothiocyanate. (Note: stabilization energies less than 0.05 kcal/mol are not reported by the NBO method).

Bibliography

[1] Accuracy within MOPAC2012. http://openmopac.net/Manual/PM6_accuracy.html.

[2] Schrödinger release 2014-2: Maestro, version 10.6, 2014.

[3] NIST Computational Chemistry Comparison and Benchmark Database, 2015.

[4] Crc handbook of chemistry and physics, 2015-2016.

[5] ALLEN, F. H., AND MOTHERWELL, W. D. S. Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry. *Acta Crystallographica Section B Structural Science 58*, 3 (jun 2002), 407–422.

[6] AMADEI, A., CERUSO, M. A., AND DI NOLA, A. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Structure, Function, and Genetics 36*, 4 (sep 1999), 419–424.

[7] ANANDAKRISHNAN, R., AGUILAR, B., AND ONUFRIEV, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids research 40*, Web Server issue (jul 2012), W537–41.

[8] ANFINSEN, C. B. Principles that govern the folding of protein chains. *Science (New York, N.Y.) 181*, 4096 (jul 1973), 223–30.

[9] ANISIMOV, V. M., LAMOUREUX, G., VOROBYOV, I. V., HUANG, N., ROUX, B., AND MACKERELL, A. D. Determination of Electrostatic Parameters for a Polarizable Force Field Based on the Classical Drude Oscillator. *Journal of Chemical Theory and Computation 1*, 1 (jan 2005), 153–168.

[10] APPLEQUIST, J., CARL, J. R., AND FUNG, K.-K. Atom Dipole Interaction Model for Molecular Polarizability. Application to Polyatomic Molecules and Determination of Atom Polarizabilities. *Journal of the American Chemical Society 94*, 9 (may 1972), 2952–2960.

[11] ASSAF, K. I., AND NAU, W. M. Cucurbiturils: from synthesis to high-affinity binding and catalysis. *Chem. Soc. Rev. 44*, 2 (2015), 394–418.

[12] BAKER, C. M. Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science 5*, 2 (2015), 241–254.

[13] BALSERA, M. A., WRIGGERS, W., OONO, Y., AND SCHULTEN, K. Principal Component Analysis and Long Time Protein Dynamics. *The Journal of Physical Chemistry 100*, 7 (jan 1996), 2567–2572.

[14] BAYLY, C. I., CIEPLAK, P., CORNELL, W. D., AND KOLLMAN, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *The Journal of Physical Chemistry 97*, 40 (oct 1993), 10269–10280.

[15] BECKE, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *The Journal of Chemical Physics 98*, 7 (apr 1993), 5648–5652.

[16] BENZ, S., MACCHIONE, M., VEROLET, Q., MAREDA, J., SAKAI, N., AND MATILE, S. Anion Transport with Chalcogen Bonds. *Journal of the American Chemical Society 138*, 29 (jul 2016), 9093–9096.

[17] BERKA, K., LASKOWSKI, R., RILEY, K. E., HOBZA, P., AND VONDRÁŠEK, J. Representative Amino Acid Side Chain Interactions in Proteins. A Comparison of Highly Accurate Correlated ab Initio Quantum Chemical and Empirical Potential Procedures. *Journal of Chemical Theory and Computation 5*, 4 (apr 2009), 982–992.

[18] BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N., AND BOURNE, P. E. The protein data bank. *Nucleic acids research 28*, 1 (2000), 235–242.

[19] BIEDERMANN, F., NAU, W. M., AND SCHNEIDER, H.-J. The Hydrophobic Effect Revisited-Studies with Supramolecular Complexes Imply High-Energy Water as a Noncovalent Driving Force. *Angewandte Chemie International Edition 53*, 42 (oct 2014), 11158–11171.

[20] BISSANTZ, C., KUHN, B., AND STAHL, M. A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry 53*, 14 (jul 2010), 5061–84.

[21] BOYS, S. F., AND BERNARDI, F. D. The calculation of small molecular interactions by the differences of separate total energies. some procedures with reduced errors. *Molecular Physics 19*, 4 (1970), 553–566.

[22] BROOKS, B. R., BRUCCOLERI, R. E., OLAFSON, B. D., STATES, D. J., SWAMINATHAN, S., AND KARPLUS, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry 4*, 2 (1983), 187–217.

[23] BRUNK, E., AND ROTHLISBERGER, U. Mixed quantum mechanical/molecular mechanical molecular dynamics simulations of biological systems in ground and electronically excited states. *Chemical reviews 115*, 12 (2015), 6217–6263.

[24] BURNS, L. A., VÁZQUEZ-MAYAGOITIA, Á., SUMPTER, B. G., AND SHERRILL, C. D. Density-Functional Approaches to Noncovalent Interactions: A Comparison of Dispersion Corrections (DFT-D), Exchange-Hole Dipole Moment (XDM) Theory, and Specialized Functionals. *The Journal of Chemical Physics 134*, 8 (feb 2011), 084107.

[25] CASE, D., BERRYMAN, J., BETZ, R., CERUTTI, D., CHEATHAM III, T., DARDEN, T., DUKE, R., GIESE, T., GOHLKE, H., GOETZ, A., ET AL. Amber 2015, university of california: San francisco, 2015.

[26] CHAI, J.-D., AND HEAD-GORDON, M. Long-Range Corrected Hybrid Density Functionals With Damped Atom-Atom Dispersion corrections. *Physical Chemistry Chemical Physics 10*, 44 (nov 2008), 6615–6620.

[27] CHIPOT, C. Frontiers in free-energy calculations of biological systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science 4*, 1 (2014), 71–89.

[28] CHOUDHARY, A., NEWBERRY, R. W., AND RAINES, R. T. $n \rightarrow \pi^*$ Interactions Engender Chirality in Carbonyl Groups. *Organic Letters 16*, 13 (jul 2014), 3421–3423.

[29] CHOUDHARY, A., AND RAINES, R. T. $n \rightarrow \pi^*$ Interactions in the Molecules of Life. *Proceedings of the 31st European Peptide Symposium* (2010).

[30] CIEPLAK, P., CALDWELL, J., AND KOLLMAN, P. Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation: Aqueous Solution Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and Amide Hydrogen Bonding and Chloroform. *Journal of Computational Chemistry 22*, 10 (jul 2001), 1048–1057.

[31] CIEPLAK, P., DUPRADEAU, F.-Y., DUAN, Y., AND WANG, J. Polarization Effects in Molecular Mechanical Force Fields. *Journal of Physics: Condensed Matter 21*, 33 (aug 2009), 333102.

[32] ČÍŽEK, J. On the correlation problem in atomic and molecular systems. calculation of wavefunction components in ursell-type expansion using quantum-field theoretical methods. *The Journal of Chemical Physics 45*, 11 (1966), 4256–4266.

[33] CLARAGE, J. B., ROMO, T., ANDREWS, B. K., PETTITT, B. M., AND PHILLIPS, G. N. A sampling problem in molecular dynamics simulations of macromolecules. *Proceedings of the National Academy of Sciences 92*, 8 (apr 1995), 3288–3292.

[34] COLE, J. C., LOMMERSE, J. P. M., ROWLAND, R. S., TAYLOR, R., AND ALLEN, F. H. Use of the Cambridge Structural Database to Study Non-Covalent Interactions: Towards a Knowledge Base of Intermolecular Interactions. In *Structure-Based Drug Design*. Springer Netherlands, Dordrecht, 1998, pp. 113–124.

[35] CORNELL, W. D., CIEPLAK, P., BAYLY, C. I., GOULD, I. R., MERZ, K. M., FERGUSON, D. M., SPELLMEYER, D. C., FOX, T., CALDWELL, J. W., AND KOLLMAN, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society 117*, 19 (may 1995), 5179–5197.

[36] CORNELL, W. D., CIEPLAK, P., BAYLY, C. I., AND KOLLMANN, P. A. Application of RESP Charges to Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation. *Journal of the American Chemical Society 115*, 21 (oct 1993), 9620–9631.

[37] CRAMER, C. J. *Essentials of computational chemistry: theories and models*. John Wiley & Sons, 2013.

[38] DAVIS, M. E., AND MCCAMMON, J. A. Electrostatics in biomolecular structure and dynamics. *Chemical Reviews 90*, 3 (1990), 509–521.

[39] DELORBE, J. E., CLEMENTS, J. H., TERESK, M. G., BENFIELD, A. P., PLAKE, H. R., MILLSPAUGH, L. E., AND MARTIN, S. F. Thermodynamic and structural effects of conformational constraints in protein-ligand interactions. Entropic paradoxy associated with ligand preorganization. *Journal of the American Chemical Society 131*, 46 (nov 2009), 16758–70.

[40] DIEDERICH, F., STANG, P. J., AND TYKWINSKI, R. R. Cucurbit[n]urils. In *Modern Supramolecular Chemistry*, W.-H. Huang, S. Liu, and L. Isaacs, Eds. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2008, ch. 4, pp. 113–142.

[41] DILABIO, G. A., JOHNSON, E. R., AND OTERO-DE-LA ROZA, A. Performance of Conventional and Dispersion-Corrected Density-Functional Theory Methods for Hydrogen Bonding Interaction Energies. *Physical Chemistry Chemical Physics 15*, 31 (jun 2013), 12821–12828.

[42] DOUGHERTY, D. A. Cation-π interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science (New York, N.Y.) 271*, 5246 (jan 1996), 163–8.

[43] DRUDE, P., RIBORG, C., AND MILLIKAN, R. A. *The Theory of Optics. Translated from the German by CR Mann and RA Millikan*. London; New York, 1902.

[44] DUNNING, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *The Journal of Chemical Physics 90*, 2 (jan 1989), 1007.

[45] DUPRADEAU, F.-Y. F.-Y., PIGACHE, A., ZAFFRAN, T., SAVINEAU, C., LELONG, R., GRIVEL, N., LELONG, D., ROSANSKI, W., AND CIEPLAK, P. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Physical chemistry chemical physics : PCCP 12*, 28 (2010), 7821–39.

[46] FANFRLÍK, J., BRONOWSKA, A. K., ŘEZÁČ, J., PRENOSIL, O., KONVALINKA, J., AND 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. *The Journal of Physical Chemistry. B 114*, 39 (oct 2010), 12666–78.

[47] FENLEY, A. T., HENRIKSEN, N. M., MUDDANA, H. S., AND GILSON, M. K. Bridging calorimetry and simulation through precise calculations of cucurbituril-guest binding enthalpies. *Journal of Chemical Theory and Computation 10*, 9 (2014), 4069–4078.

[48] FERENCZY, G. G., AND REYNOLDS, C. A. Modeling Polarization Through Induced Atomic Charges. *Journal of Physical Chemistry A 105* (2001), 11470–11479.

[49] FERRIN, T. E., HUANG, C. C., JARVIS, L. E., AND LANGRIDGE, R. The MIDAS Display System. *Journal of Molecular Graphics 6*, 1 (mar 1988), 13–27.

[50] FLYVBJERG, H., AND PETERSEN, H. G. Error estimates on averages of correlated data. *The Journal of Chemical Physics 91*, 1 (jul 1989), 461.

[51] FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., SCALMANI, G., BARONE, V., MENNUCCI, B., PETERSSON, G. A., NAKATSUJI, H., CARICATO, M., LI, X., HRATCHIAN, H. P., IZMAYLOV, A. F., BLOINO, J., ZHENG, G., SONNENBERG, J. L., HADA, M., EHARA, M., TOYOTA, K., FUKUDA, R., HASEGAWA, J., ISHIDA, M., NAKAJIMA, T., HONDA, Y., KITAO, O., NAKAI, H., VREVEN, T., MONTGOMERY JR., J. A., PERALTA, J. E., OGLIARO, F., BEARPARK, M., HEYD, J. J., BROTHERS, E., KUDIN, K. N., STAROVEROV, V. N., KOBAYASHI, R., NORMAND, J., RAGHAVACHARI, K., RENDELL, A., BURANT, J. C., IYENGAR, S. S., TOMASI, J., COSSI, M., REGA, N., MILLAM, J. M., KLENE, M., KNOX, J. E., CROSS, J. B., BAKKEN, V., ADAMO, C., JARAMILLO, J., GOMPERTS, R., STRATMANN, R. E., YAZYEV, O., AUSTIN, A. J., CAMMI, R., POMELLI, C., OCHTERSKI, J. W., MARTIN, R. L., MOROKUMA, K., ZAKRZEWSKI, V. G., VOTH, G. A., SALVADOR, P., DANNENBERG, J. J., DAPPRICH, S., DANIELS, A. D., FARKAS, Ö., FORESMAN, J. B., ORTIZ, J. V., CIOSLOWSKI, J., AND FOX, D. J. Gaussian 09 Revision C.01. Gaussian Inc. Wallingford CT 2009.

[52] FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., SCALMANI, G., BARONE, V., MENNUCCI, B., PETERSSON, G. A., NAKATSUJI, H., CARICATO, M., LI, X., HRATCHIAN, H. P., IZMAYLOV, A. F., BLOINO, J., ZHENG, G., SONNENBERG, J. L., HADA, M.,

EHARA, M., TOYOTA, K., FUKUDA, R., HASEGAWA, J., ISHIDA, M., NAKAJIMA, T., HONDA, Y., KITAO, O., NAKAI, H., VREVEN, T., MONTGOMERY JR., J. A., PERALTA, J. E., OGLIARO, F., BEARPARK, M., HEYD, J. J., BROTHERS, E., KUDIN, K. N., STAROVEROV, V. N., KOBAYASHI, R., NORMAND, J., RAGHAVACHARI, K., RENDELL, A., BURANT, J. C., IYENGAR, S. S., TOMASI, J., COSSI, M., REGA, N., MILLAM, J. M., KLENE, M., KNOX, J. E., CROSS, J. B., BAKKEN, V., ADAMO, C., JARAMILLO, J., GOMPERTS, R., STRATMANN, R. E., YAZYEV, O., AUSTIN, A. J., CAMMI, R., POMELLI, C., OCHTERSKI, J. W., MARTIN, R. L., MOROKUMA, K., ZAKRZEWSKI, V. G., VOTH, G. A., SALVADOR, P., DANNENBERG, J. J., DAPPRICH, S., DANIELS, A. D., FARKAS, Ö., FORESMAN, J. B., ORTIZ, J. V., CIOSLOWSKI, J., AND FOX, D. J. Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT 2009.

[53] GAINZA, P., ROBERTS, K. E., GEORGIEV, I., LILIEN, R. H., KEEDY, D. A., CHEN, C.-Y., REZA, F., ANDERSON, A. C., RICHARDSON, D. C., RICHARDSON, J. S., AND DONALD, B. R. OSPREY: Protein Design With Ensembles, Flexibility, and Provable Algorithms. *Methods in Enzymology 523* (jan 2013), 87–107.

[54] GLENDENING, E., BADENHOOP, J., REED, A., CARPENTER, J., BOHMANN, J., MORALES, C., AND WEINHOLD, F. NBO 3.0, 1998.

[55] GOKEL, G., AND SCHALL, O. *Comprehensive supramolecular chemistry*.

[56] GORDON, J. C., MYERS, J. B., FOLTA, T., SHOJA, V., HEATH, L. S., AND ONUFRIEV, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic acids research 33*, Web Server issue (jul 2005), W368–71.

[57] GRÁFOVÁ, L., PITOŇÁK, M., ŘEZÁČ, J., AND HOBZA, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *Journal of Chemical Theory and Computation 6*, 8 (aug 2010), 2365–2376.

[58] GRESH, N., CISNEROS, G. A., DARDEN, T. A., AND PIQUEMAL, J.-P. Anisotropic, Polarizable Molecular Mechanics Studies of Inter- and Intramolecular Interactions and Ligand-Macromolecule Complexes. A Bottom-Up Strategy. *Journal of Chemical Theory and Computation 3*, 6 (nov 2007), 1960–1986.

[59] GRIMME, S. Semiempirical Hybrid Density Functional With Perturbative Second-Order Correlation. *The Journal of Chemical Physics 124*, 3 (jan 2006), 034108.

[60] GRIMME, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chemistry European Journal 18*, 32 (aug 2012), 9955–9964.

[61] GRIMME, S., ANTONY, J., EHRLICH, S., AND KRIEG, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction

(DFT-D) for the 94 Elements H-Pu. *The Journal of Chemical Physics 132*, 15 (apr 2010), 154104.

[62] GRIMME, S., EHRLICH, S., AND GOERIGK, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *Journal of computational chemistry 32*, 7 (may 2011), 1456–1465.

[63] HAGLER, A. T., HULER, E., AND LIFSON, S. Energy Functions for Peptides and Proteins. I. Derivation of a Consistent Force Field Including the Hydrogen Bond From Amide Crystals. *Journal of the American Chemical Society 96*, 17 (aug 1974), 5319–5327.

[64] HARDER, E., ANISIMOV, V. M., VOROBYOV, I. V., LOPES, P. E. M., NOSKOV, S. Y., MACKERELL, A. D., AND ROUX, B. Atomic Level Anisotropy in the Electrostatic Modeling of Lone Pairs for a Polarizable Force Field Based on the Classical Drude Oscillator. *Journal of Chemical Theory and Computation 2*, 6 (2006), 1587–1597.

[65] HAYWARD, S., AND GROOT, B. L. Normal Modes and Essential Dynamics. 2008, pp. 89–106.

[66] HENRIKSEN, N. M., FENLEY, A. T., AND GILSON, M. K. Computational Calorimetry: High-Precision Calculation of Host-Guest Binding Thermodynamics. *Journal of Chemical Theory and Computation 11*, 9 (sep 2015), 4377–94.

[67] HESSELMANN, A., AND KORONA, T. Intermolecular symmetry-adapted perturbation theory study of large organic complexes. *The Journal of Chemical Physics 141*, 9 (sep 2014), 094107.

[68] HOBZA, P. The calculation of intermolecular interaction energies. *Annual Reports Section" C"(Physical Chemistry) 107* (2011), 148–168.

[69] HOHENSTEIN, E. G., AND SHERRILL, C. D. Efficient evaluation of triple excitations in symmetry-adapted perturbation theory via second-order Møller-Plesset perturbation theory natural orbitals. *The Journal of Chemical Physics 133*, 10 (2010), 104107.

[70] HOHENSTEIN, E. G., AND SHERRILL, C. D. Wavefunction Methods for Noncovalent Interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science 2*, 2 (mar 2012), 304–326.

[71] HOMEYER, N., HORN, A. H. C., LANIG, H., AND STICHT, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *Journal of molecular modeling 12*, 3 (feb 2006), 281–9.

[72] HONIG, B., AND NICHOLLS, A. Classical electrostatics in biology and chemistry. *Science (New York, N.Y.) 268*, 5214 (may 1995), 1144–9.

[73] HOPKINS, C. W., LE GRAND, S., WALKER, R. C., AND ROITBERG, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *Journal of Chemical Theory and Computation 11*, 4 (apr 2015), 1864–1874.

[74] HOSTAŠ, J., ŘEZÁČ, J., AND HOBZA, P. On the Performance of the Semiempirical Quantum Mechanical PM6 and PM7 Methods for Noncovalent Interactions. *Chemical Physics Letters 568-569*, null (may 2013), 161–166.

[75] HUANG, L., AND ROUX, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on Ab Initio Target Data. *Journal of Chemical Theory and Computation 9*, 8 (aug 2013), 3543–3556.

[76] ILATOVSKIY, A. V., ABAGYAN, R., AND KUFAREVA, I. Quantum Mechanics Approaches to Drug Research in the Era of Structural Chemogenomics. *International Journal of Quantum Chemistry 113*, 12 (jun 2013), 1669–1675.

[77] ISAACS, L. Stimuli Responsive Systems Constructed Using Cucurbit[*n*]uril-type molecular containers. *Accounts of Chemical Research*, 7, 2052–2062.

[78] JEFFREY, G. A. *An introduction to hydrogen bonding*, vol. 12.

[79] JEZIORSKI, B., MOSZYNSKI, R., AND SZALEWICZ, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chemical Reviews 94*, 7 (nov 1994), 1887–1930.

[80] JOHNSON, E. R., KEINAN, S., MORI-SÁNCHEZ, P., CONTRERAS-GARCÍA, J., COHEN, A. J., AND YANG, W. Revealing Noncovalent Interactions. *Journal of the American Chemical Society 132*, 18 (may 2010), 6498–6506.

[81] JONES, E., OLIPHANT, T., PETERSON, P., AND OTHERS. SciPy: Open source scientific tools for Python.

[82] JORGENSEN, W. L., AND TIRADO-RIVES, J. The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *Journal of the American Chemical Society 110*, 6 (mar 1988), 1657–1666.

[83] JUŘECKA, P., SPONER, J., CERNÝ, J., AND HOBZA, P. Benchmark Database of Accurate (MP2 and CCSD(T) Complete Basis Set Limit) Interaction Energies of Small Model Complexes, DNA Base Pairs, and Amino Acid Pairs. *Physical Chemistry Chemical Physics 8*, 17 (may 2006), 1985–1993.

[84] KAMER, K. J., CHOUDHARY, A., AND RAINES, R. T. Intimate Interactions with Carbonyl Groups: Dipole-Dipole or $n \rightarrow \pi^*$? *The Journal of Organic Chemistry 78*, 5 (mar 2013), 2099–2103.

[85] KAMINSKI, G. A., STERN, H. A., BERNE, B. J., AND FRIESNER, R. A. Development of an Accurate and Robust Polarizable Molecular Mechanics Force Field from ab Initio Quantum Chemistry. *The Journal of Physical Chemistry A 108*, 4 (jan 2004), 621–627.

[86] KARSHIKOFF, A. *Non-covalent interactions in proteins*. World Scientific, 2006.

[87] KENDALL, R. A., DUNNING, T. H., AND HARRISON, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *The Journal of Chemical Physics 96*, 9 (may 1992), 6796.

[88] KIM, Y., KIM, H., KO, Y., SELVAPALAM, N., REKHARSKY, M., INOUE, Y., AND KIM, K. Complexation of Aliphatic Ammonium Ions with a Water-Soluble Cucurbit[6]uril Derivative in Pure Water: Isothermal Calorimetric, NMR, and X-ray Crystallographic Study. *Chemistry - A European Journal 15*, 25 (jun 2009), 6143–6151.

[89] KING, G., AND WARSHEL, A. Investigation of the Free Energy Functions for Electron Transfer Reactions. *The Journal of Chemical Physics 93*, 12 (1990), 8682.

[90] KLAMT, A. The COSMO and COSMO-RS solvation models. *Wiley Interdisciplinary Reviews: Computational Molecular Science 1*, 5 (sep 2011), 699–709.

[91] KNOWLES, R. R., AND JACOBSEN, E. N. Attractive noncovalent interactions in asymmetric catalysis: Links between enzymes and small molecule catalysts. *Proceedings of the National Academy of Sciences 107*, 48 (nov 2010), 20678–20685.

[92] KO, Y., KIM, H., KIM, Y., AND KIM, K. U-Shaped Conformation of Alkyl Chains Bound to a Synthetic Host. *Angewandte Chemie 120*, 22 (may 2008), 4174–4177.

[93] KORTH, M. Third-Generation Hydrogen-Bonding Corrections for Semiempirical QM Methods and Force Fields. *Journal of Chemical Theory and Computation 6*, 12 (nov 2010), 3808–3816.

[94] KORTH, M., AND GRIMME, S. "Mindless" DFT Benchmarking. *Journal of Chemical Theory and Computation 5*, 4 (apr 2009), 993–1003.

[95] KORTH, M., PITON, M., PITOŇÁK, M., ŘEZÁČ, J., AND HOBZA, P. A Transferable H-Bonding Correction for Semiempirical Quantum-Chemical Methods. *Journal of Chemical Theory and Computation 6*, 1 (jan 2010), 344–352.

[96] KRASIA, T., KHODABAKHSH, S., TUNCEL, D., AND STEINKE, J. H. Cucurbituril: a versatile "bead" for polyrotaxane synthesis. In *Macromolecular Nanostructured Materials*. Springer, 2004, pp. 41–59.

[97] LAGONA, J., MUKHOPADHYAY, P., CHAKRABARTI, S., AND ISAACS, L. The Cucurbit[n]uril Family. *Angewandte Chemie International Edition 44*, 31 (aug 2005), 4844–4870.

[98] LAMOUREUX, G., AND ROUX, B. Modeling Induced Polarization With Classical Drude Oscillators: Theory and Molecular Dynamics Simulation Algorithm. *Journal of Chemical Physics 119*, 6 (jul 2003), 3025–3039.

[99] LEACH, A. R. *Molecular modelling: principles and applications*. Pearson education, 2001.

[100] LEE, C., YANG, W., AND PARR, R. G. Development of the Colle-Salvetti Correlation-Energy Formula Into a Functional of the Electron Density. *Physical Review B 37*, 2 (jan 1988), 785–789.

[101] LEHN, J.-M. Supramolecular Chemistry–Scope and Perspectives Molecules, Supermolecules, and Molecular Devices (Nobel Lecture). *Angewandte Chemie International Edition in English 27*, 1 (jan 1988), 89–112.

[102] LEVITT, M., AND LIFSON, S. Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. *Journal of Molecular Biology 46*, 2 (1969), 269–279.

[103] LI, A., MUDDANA, H. S., AND GILSON, M. K. Quantum Mechanical Calculation of Noncovalent Interactions: A Large-Scale Evaluation of PMx, DFT, and SAPT Approaches. *Journal of Chemical Theory and Computation 10*, 4 (apr 2014), 1563–1575.

[104] LIFSON, S. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *The Journal of Chemical Physics 49*, 11 (sep 1968), 5116–5129.

[105] LODISH, H., BERK, A., ZIPURSKY, S. L., MATSUDAIRA, P., BALTIMORE, D., AND DARNELL, J. *Noncovalent bonds*. WH Freeman, 2000.

[106] MACKERELL, A. D. Empirical Force Fields for Biological Macromolecules: Overview and issues. *Journal of Computational Chemistry 25*, 13 (oct 2004), 1584–604.

[107] MACKERELL, A. D., BASHFORD, D., BELLOTT, M., DUNBRACK, R. L., EVANSECK, J. D., FIELD, M. J., FISCHER, S., GAO, J., GUO, H., HA, S., JOSEPH-MCCARTHY, D., KUCHNIR, L., KUCZERA, K., LAU, F. T., MATTOS, C., MICHNICK, S., NGO, T., NGUYEN, D. T., PRODHOM, B., REIHER, W. E., ROUX, B., SCHLENKRICH, M., SMITH, J. C., STOTE, R., STRAUB, J., WATANABE, M., WIÓRKIEWICZ-KUCZERA, J., YIN, D., AND KARPLUS, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of proteins. *The Journal of Physical Chemistry B 102*, 18 (apr 1998), 3586–616.

[108] MACKERELL, A. D., WIORKIEWICZ-KUCZERA, J., AND KARPLUS, M. An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids. *Journal of the American Chemical Society 117*, 48 (dec 1995), 11946–11975.

[109] MAMMEN, M., CHOI, S.-K., AND WHITESIDES, G. M. Polyvalent Interactions in Biological Systems: Implications for Design and Use of Multivalent Ligands and Inhibitors. *Angewandte Chemie International Edition 37*, 20 (nov 1998), 2754–2794.

[110] MAYNE, C. G., SAAM, J., SCHULTEN, K., TAJKHORSHID, E., AND GUMBART, J. C. Rapid Parameterization of Small Molecules Using the Force Field Toolkit. *Journal of Computational Chemistry 34*, 32 (dec 2013), 2757–2770.

[111] MCDANIEL, J. G., AND SCHMIDT, J. R. Physically-Motivated Force Fields From Symmetry-Adapted Perturbation Theory. *The Journal of Physical Chemistry. A 117*, 10 (mar 2013), 2053–2066.

[112] MCNEMAR, C., SNOW, M. E., WINDSOR, W. T., PRONGAY, A., MUI, P., ZHANG, R., DURKIN, J., LE, H. V., AND WEBER, P. C. Thermodynamic and structural analysis of phosphotyrosine polypeptide binding to Grb2-SH2. *Biochemistry 36*, 33 (aug 1997), 10006–14.

[113] MEYER, E. A., CASTELLANO, R. K., AND DIEDERICH, F. Interactions with Aromatic Rings in Chemical and Biological Recognition. *Angewandte Chemie International Edition 42*, 11 (mar 2003), 1210–1250.

[114] MOCK, W. L. Cucurbituril. Springer Berlin Heidelberg, 1995, pp. 1–24.

[115] MØLLER, C., AND PLESSET, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Physical Review 46*, 7 (oct 1934), 618–622.

[116] MORTIER, W. J., VAN GENECHTEN, K., AND GASTEIGER, J. Electronegativity Equalization: Application and Parametrization. *Journal of the American Chemical Society 107*, 4 (feb 1985), 829–835.

[117] MUDDANA, H. S., AND GILSON, M. K. Calculation of Host-Guest Binding Affinities Using a Quantum-Mechanical Energy Model. *Journal of Chemical Theory and Computation 8*, 6 (jun 2012), 2023–2033.

[118] MUDDANA, H. S., AND GILSON, M. K. Prediction of SAMPL3 Host-Guest Binding Affinities: Evaluating the Accuracy of Generalized Force-fields. *Journal of Computer-Aided Molecular Design 26*, 5 (may 2012), 517–25.

[119] MUDDANA, H. S., VARNADO, C. D., BIELAWSKI, C. W., URBACH, A. R., ISAACS, L., GEBALLE, M. T., AND GILSON, M. K. Blind Prediction of Host-Guest Binding Affinities: A New SAMPL3 Challenge. *Journal of Computer-Aided Molecular Design 26*, 5 (may 2012), 475–87.

[120] MÜLLER-DETHLEFS, K., AND HOBZA, P. Noncovalent Interactions:Âă A Challenge for Experiment and Theory. *Chemical Reviews 100*, 1 (jan 2000), 143–168.

[121] MYERS, J., GROTHAUS, G., NARAYANAN, S., AND ONUFRIEV, A. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins: Structure, Function, and Bioinformatics 63*, 4 (feb 2006), 928–938.

[122] NEWBERRY, R. W., BARTLETT, G. J., VANVELLER, B., WOOLFSON, D. N., AND RAINES, R. T. Signatures of $n \rightarrow \pi^*$ interactions in proteins. *Protein Science 23*, 3 (mar 2014), 284–288.

[123] OH, K.-I., LEE, J.-H., JOO, C., HAN, H., AND CHO, M. β-Azidoalanine as an IR Probe: Application to Amyloid Aβ(16-22) Aggregation. *The Journal of Physical Chemistry B 112*, 33 (aug 2008), 10352–10357.

[124] OOSTENBRINK, C., VILLA, A., MARK, A. E., AND VAN GUNSTEREN, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force-Field Parameter Sets 53A5 and 53A6. *Journal of Computational Chemistry 25*, 13 (oct 2004), 1656–76.

[125] PALMO, K., MANNFORS, B., MIRKIN, N. G., AND KRIMM, S. Potential Energy Functions: From Consistent Force Fields to Spectroscopically Determined Polarizable Force Fields. *Biopolymers 68*, 3 (mar 2003), 383–94.

[126] PARKER, T. M., BURNS, L. A., PARRISH, R. M., RYNO, A. G., AND SHERRILL, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *Journal of Chemical Physics 140*, 9 (2014).

[127] PARVARI, G., REANY, O., AND KEINAN, E. Applicable Properties of Cucurbiturils. *Israel Journal of Chemistry 51*, 5-6 (may 2011), 646–663.

[128] PATON, R. S., AND GOODMAN, J. M. Hydrogen Bonding and Pi-Stacking: How Reliable Are Force Fields? a Critical Evaluation of Force Field Descriptions of Nonbonded Interactions. *Journal of Chemical Information and Modeling 49*, 4 (apr 2009), 944–55.

[129] PAULINI, R., MÜLLER, K., AND DIEDERICH, F. Orthogonal Multipolar Interactions in Structural Chemistry and Biology. *Angewandte Chemie International Edition 44*, 12 (mar 2005), 1788–1805.

[130] PÉREZ, A., MARCHÁN, I., SVOZIL, D., SPONER, J., CHEATHAM, T. E., LAUGHTON, C. A., AND OROZCO, M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of Alpha/Gamma Conformers. *Biophysical Journal 92*, 11 (jun 2007), 3817–29.

[131] POLITZER, P., MURRAY, J. S., AND CLARK, T. Halogen bonding and other σ-hole interactions: a perspective. *Physical Chemistry Chemical Physics 15*, 27 (2013), 11178.

[132] PONDER, J. W., AND CASE, D. A. Force Fields for Protein Simulations. *Advances in Protein Chemistry 66* (jan 2003), 27–85.

[133] RAHA, K., AND 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. *Journal of Medicinal Chemistry 48*, 14 (jul 2005), 4558–4575.

[134] RAMACHANDRAN, K., DEEPA, G., AND NAMBOORI, K. *Computational chemistry and molecular modeling: principles and applications*. Springer Science & Business Media, 2008.

[135] RAPPÉ, A. K., AND GODDARD III, W. A. Charge Equilibration for Molecular Dynamics Simulations. *The Journal of Physical Chemistry 95*, 8340 (apr 1991), 3358–3363.

[136] REBEK, J. Molecular recognition and biophysical organic chemistry. *Accounts of Chemical Research 23*, 12 (dec 1990), 399–404.

[137] REED, A. E., WEINSTOCK, R. B., AND WEINHOLD, F. Natural population analysis. *The Journal of Chemical Physics 83*, 2 (1985), 735.

[138] REINHOUDT, D., Ed. *Comprehensive supramolecular chemistry*.

[139] REN, P., AND PONDER, J. W. Consistent Treatment of Inter- and Intramolecular Polarization in Molecular Mechanics Calculations. *Journal of Computational Chemistry 23*, 16 (dec 2002), 1497–506.

[140] ŘEZÁČ, J., FANFRLIÌĄK, J., SALAHUB, D., AND HOBZA, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *Journal of Chemical Theory and Computation 5*, 7 (jul 2009), 1749–1760.

[141] ŘEZÁČ, J., AND HOBZA, P. A Halogen-Bonding Correction for the Semiempirical PM6 Method. *Chemical Physics Letters 506*, 4-6 (apr 2011), 286–289.

[142] ŘEZÁČ, J., AND HOBZA, P. Extrapolation and Scaling of the DFT-SAPT Interaction Energies toward the Basis Set Limit. *Journal of Chemical Theory and Computation 7*, 3 (mar 2011), 685–689.

[143] ŘEZÁČ, J., AND HOBZA, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *Journal of Chemical Theory and Computation 8*, 1 (jan 2012), 141–151.

[144] ŘEZÁČ, J., AND HOBZA, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the "Gold Standard," CCSD(T) at the Complete Basis Set Limit? *Journal of Chemical Theory and Computation 9*, 5 (may 2013), 2151–2155.

[145] ŘEZÁČ, J., JUREČKA, P., RILEY, K. E., ČERNÝ, J., VALDES, H., PLUHÁČKOVÁ, K., BERKA, K., ŘEZÁČ, T., PITOŇÁK, M., VONDRÁŠEK, J., AND HOBZA, P. Quantum Chemical Benchmark Energy and Geometry Database for Molecular Clusters and Complex Molecular Systems (www.begdb.com): A Users Manual and Examples. *Collection of Czechoslovak Chemical Communications 73*, 10 (2008), 1261–1270.

[146] ŘEZÁČ, J., RILEY, K. E., AND HOBZA, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *Journal of Chemical Theory and Computation 7*, 8 (aug 2011), 2427–2438.

[147] ŘEZÁČ, J., RILEY, K. E., AND HOBZA, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *Journal of Chemical Theory and Computation 8*, 11 (nov 2012), 4285–4292.

[148] RICK, S. W., AND STUART, S. J. Potentials and Algorithms for Incorporating Polarizability in Computer Simulations. *Reviews in Computational Chemistry 18* (2002), 89–146.

[149] RICK, S. W., STUART, S. J., AND BERNE, B. J. Dynamical Fluctuating Charge Force Fields: Application to Liquid Water. *The Journal of Chemical Physics 101*, 7 (oct 1994), 6141–6159.

[150] SCHNEIDER, H.-J., AND YATSIMIRSKY, A. K. *Principles and methods in supramolecular chemistry*. Wiley, 2000.

[151] SCHYMAN, P., AND JORGENSEN, W. L. Exploring Adsorption of Water and Ions on Carbon Surfaces using a Polarizable Force Field. *The Journal of Physical Chemistry Letters 4*, 3 (feb 2013), 468–474.

[152] SEDLAK, R., JANOWSKI, T., PITOŇÁK, M., REZÁČ, J., PULAY, P., AND HOBZA, P. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *Journal of Chemical Theory and Computation 9*, 8 (jul 2013), 3364–3374.

[153] SENN, H. M., AND THIEL, W. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition 48*, 7 (2009), 1198–1229.

[154] SHARP, K., JEAN-CHARLES, A., AND HONIG, B. A local dielectric constant model for solvation free energies which accounts for solute polarizability. *The Journal of Physical Chemistry 96*, 9 (1992), 3822–3828.

[155] SHERRILL, C. D., SUMPTER, B. G., SINNOKROT, M. O., MARSHALL, M. S., HOHENSTEIN, E. G., WALKER, R. C., AND GOULD, I. R. Assessment of Standard Force Field Models Against High-Quality Ab Initio Potential Curves for Prototypes of π-π , CH/π , and SH/π Interactions. *Journal of Computational Chemistry 30*, 14 (nov 2009), 2187–93.

[156] SHI, Y., XIA, Z., ZHANG, J., BEST, R., WU, C., PONDER, J. W., AND REN, P. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *Journal of Chemical Theory and Computation 9*, 9 (aug 2013), 4046–4063.

[157] SHI, Y., ZHU, C. Z., MARTIN, S. F., AND REN, P. Probing the effect of conformational constraint on phosphorylated ligand binding to an SH2 domain using polarizable force field simulations. *The journal of physical chemistry. B 116*, 5 (feb 2012), 1716–27.

[158] SHIRTS, M. R., AND CHODERA, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics 129*, 12 (2008).

[159] SIMMONETT, A. C., PICKARD, F. C., SHAO, Y., CHEATHAM, T. E., AND BROOKS, B. R. Efficient treatment of induced dipoles. *The Journal of Chemical Physics 143*, 7 (2015), 074115.

[160] SINHA, M. K., REANY, O., YEFET, M., BOTOSHANSKY, M., AND KEINAN, E. Bistable Cucurbituril Rotaxanes Without Stoppers. *Chemistry - A European Journal 18*, 18 (apr 2012), 5589–5605.

[161] STEINBRECHER, T., LATZER, J., AND CASE, D. A. Revised AMBER parameters for bioorganic phosphates. *Journal of chemical theory and computation 8*, 11 (nov 2012), 4405–4412.

[162] STEINER, T. The Hydrogen Bond in the Solid State. *Angewandte Chemie International Edition 41*, 1 (jan 2002), 48–76.

[163] STERN, H. A., KAMINSKI, G. A., BANKS, J. L., ZHOU, R., BERNE, B. J., AND FRIESNER, R. A. Fluctuating Charge, Polarizable Dipole, and Combined Models: Parameterization from ab Initio Quantum Chemistry. *The Journal of Physical Chemistry B 103*, 22 (jun 1999), 4730–4737.

[164] STEWART, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *Journal of Molecular Modeling 13*, 12 (dec 2007), 1173–1213.

[165] STEWART, J. J. P. MOPAC2012, 2012.

[166] STEWART, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *Journal of Molecular Modeling 19*, 1 (jan 2013), 1–32.

[167] STRAATSMA, T. P., AND MCCAMMON, J. A. Molecular Dynamics Simulations with Interaction Potentials Including Polarization Development of a Noniterative Method and Application to Water. *Molecular Simulation 5*, February 2015 (1990), 181–192.

[168] TAN, Y.-H., TAN, C., WANG, J., AND LUO, R. Continuum polarizable force field within the poisson–boltzmann framework. *The Journal of Physical Chemistry B 112*, 25 (2008), 7675–7688.

[169] TANFORD, C. The hydrophobic effect and the organization of living matter. *Science (New York, N.Y.) 200*, 4345 (jun 1978), 1012–8.

[170] TANFORD, C. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes 2d Ed.* J. Wiley., 1980.

[171] THOLE, B. Molecular Polarizabilities Calculated with a Modified Dipole Interaction. *Chemical Physics 59* (1981), 341–350.

[172] TOMASI, J., AND PERSICO, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chemical Reviews 94*, 7 (nov 1994), 2027–2094.

[173] TRUCHON, J.-F., NICHOLLS, A., ROUX, B., IFTIMIE, R. I., AND BAYLY, C. I. Integrated continuum dielectric approaches to treat molecular polarizability and the condensed phase: Refractive index and implicit solvation. *Journal of Chemical Theory and Computation 5*, 7 (2009), 1785–1802.

[174] TURNEY, J. M., SIMMONETT, A. C., PARRISH, R. M., HOHENSTEIN, E. G., EVANGELISTA, F. A., FERMANN, J. T., MINTZ, B. J., BURNS, L. A., WILKE, J. J., ABRAMS, M. L., RUSS, N. J., LEININGER, M. L., JANSSEN, C. L., SEIDL, E. T., ALLEN, W. D., SCHAEFER, H. F., KING, R. A., VALEEV, E. F., SHERRILL, C. D., AND CRAWFORD, T. D. Psi4: An Open-Source Ab Initio Electronic Structure Program. *Wiley Interdisciplinary Reviews: Computational Molecular Science 2*, 4 (jul 2012), 556–565.

[175] VAN DER WALT, S., COLBERT, S. C., AND VAROQUAUX, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering 13*, 2 (mar 2011), 22–30.

[176] VANOMMESLAEGHE, K., AND MACKERELL, A. D. CHARMM Additive and Polarizable Force Fields for Biophysics and Computer-Aided Drug design. *Biochimica et Biophysica Acta 1850*, 5 (may 2015), 861–71.

[177] VANQUELEF, E., SIMON, S., MARQUANT, G., GARCIA, E., KLIMERAK, G., DELEPINE, J. C., CIEPLAK, P., AND DUPRADEAU, F. Y. R.E.D. Server: A web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Research 39*, SUPPL. 2 (jul 2011), W511–7.

[178] VÖGTLE, F., AND ATWOOD, J. L. *Comprehensive supramolecular chemistry. 2. Molecular recognition: receptors for molecular guests*, vol. 2.

[179] VONDRÁSEK, J., BENDOVÁ, L., KLUSÁK, V., AND HOBZA, P. Unexpectedly Strong Energy Stabilization Inside the Hydrophobic Core of Small Protein Rubredoxin Mediated by Aromatic Residues: Correlated Ab Initio Quantum Chemical Calculations. *Journal of the American Chemical Society 127*, 8 (mar 2005), 2615–2619.

[180] WANG, J., CIEPLAK, P., LI, J., HOU, T., LUO, R., AND DUAN, Y. Development of Polarizable Models for Molecular Mechanical Calculations I: Parameterization of Atomic Polarizability. *Journal of Physical Chemistry B 115*, 12 (2011), 3091–3099.

[181] WANG, J., WANG, W., KOLLMAN, P. A., AND CASE, D. A. Antechamber: an Accessory Software Package for Molecular Mechanical Calculations.

[182] WANG, J., WOLF, R. M., CALDWELL, J. W., KOLLMAN, P. A., AND CASE, D. A. Development and Testing of a General Amber Force field. *Journal of Computational Chemistry 25*, 9 (jul 2004), 1157–74.

[183] WANG, L.-P., HEAD-GORDON, T., PONDER, J. W., REN, P., CHODERA, J. D., EASTMAN, P. K., MARTINEZ, T. J., AND PANDE, V. S. Systematic Improvement of a Classical Molecular Model of Water. *The Journal of Physical Chemistry B 117*, 34 (aug 2013), 9956–72.

[184] WANG, W., DONINI, O., REYES, C. M., AND KOLLMAN, P. A. Biomolecular Simulations: Recent Developments in Force Fields, Simulations of Enzyme Catalysis, Protein-Ligand, Protein-Protein, and Protein-Nucleic Acid Noncovalent Interactions. *Annual Review of Biophysics and Biomolecular Structure 30* (jan 2001), 211–243.

[185] WANG, Z.-X., ZHANG, W., WU, C., LEI, H., CIEPLAK, P., AND DUAN, Y. Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Field for Simulations of Proteins and peptides. *Journal of Computational Chemistry 27*, 6 (apr 2006), 781–90.

[186] WHITESIDES, G. M., AND GRZYBOWSKI, B. Self-assembly at all scales. *Science (New York, N.Y.) 295*, 5564 (mar 2002), 2418–21.

[187] WILSON, A. K., WOON, D. E., PETERSON, K. A., AND DUNNING, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. IX. the Atoms Gallium Through Krypton. *The Journal of Chemical Physics 110*, 16 (apr 1999), 7667.

[188] WINN, P. J., FERENCZY, G. G., AND REYNOLDS, C. A. Towards Improved Force Fields: III. Polarization Through Modified Atomic Charges. *Journal of Computational Chemistry 20*, 7 (may 1999), 704–712.

[189] WOON, D. E., AND DUNNING, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. the Atoms Aluminum Through Argon. *The Journal of Chemical Physics 98*, 2 (jan 1993), 1358.

[190] WOON, D. E., AND DUNNING, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. IV. Calculation of Static Electrical Response Properties. *The Journal of Chemical Physics 100*, 4 (feb 1994), 2975.

[191] YILMAZER, N. D., AND KORTH, M. Comparison of Molecular Mechanics, Semi-Empirical Quantum Mechanical, and Density Functional Theory Methods for Scoring Protein-Ligand Interactions. *The Journal of Physical Chemistry. B 117*, 27 (jul 2013), 8075–84.

[192] YU, H., HANSSON, T., AND VAN GUNSTEREN, W. F. Development of a Simple, Self-Consistent Polarizable Model for Liquid Water. *Journal of Chemical Physics 118*, 1 (dec 2003), 221–234.

[193] ZHANG, K.-D., AJAMI, D., GAVETTE, J. V., AND REBEK, J. Alkyl Groups Fold to Fit within a Water-Soluble Cavitand. *Journal of the American Chemical Society 136*, 14 (apr 2014), 5264–5266.

[194] ZHANG, K.-D., AJAMI, D., GAVETTE, J. V., AND REBEK, J. Complexation of alkyl groups and ghrelin in a deep, water-soluble cavitand. *Chemical Communications 50*, 38 (2014), 4895.

[195] ZHAO, Y., AND TRUHLAR, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theoretical Chemistry Accounts 120*, 1-3 (jul 2008), 215–241.

[196] ZHAO, Y. H., ABRAHAM, M. H., AND ZISSIMOS, A. M. Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds. *The Journal of Organic Chemistry 68*, 19 (sep 2003), 7368–7373.

[197] ZHU, C., BYRD, R. H., LU, P., AND NOCEDAL, J. Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization. *ACM Transactions on Mathematical Software 23*, 4 (dec 1997), 550–560.