THEORETICAL INVESTIGATIONS OF PI-PI AND SULFUR-PI INTERACTIONS AND THEIR ROLES IN BIOMOLECLUAR SYSTEMS

A Thesis
Presented to
The Academic Faculty

by

Anthony P. Tauer

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Chemistry

Georgia Institute of Technology

December 2005

THEORETICAL INVESTIGATIONS OF PI-PI AND SULFUR-PI INTERACTIONS AND THEIR ROLES IN BIOMOLECLUAR SYSTEMS

Approved by:

C. David Sherrill, Advisor
School of Chemistry and Biochemistry
Georgia Institute of Technology

Jen-Luc Bredas
School of Chemistry and Biochemistry
Georgia Institute of Technology

Rigoberto Hernandez
School of Chemistry and Biochemistry
Georgia Institute of Technology

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The study of noncovalent interactions between aromatic rings and various functional groups is a very popular topic in current computational chemistry. The research presented in this thesis takes steps to bridge the gap between theoretical prototypes and real-world systems.

The non-additive contributions to the interaction energy in stacked aromatic systems are measured by expanding the prototype benzene dimer into trimeric and tetrameric systems. We show that the three- and four-body interaction terms generally do not contribute significantly to the overall interaction energy, and that the two-body terms are essentially the same as in the isolated dimer.

The sulfur–π interaction is then studied by using the H_2S–benzene dimer as a prototype system for theoretical predictions. We obtain higly-accurate potential energy curves, as well as an interaction energy extrapolated to the complete basis set limit. Energy decomposition analysis using symmetry-adapted perturbation theory shows that the S–π interaction is primarily electrostatic in nature.

These theoretical results are then compared to an analysis of real S–π contacts found by searching protein structures in the Brookhaven Protein DataBank. We find that the most frequently seen configuration does not correspond to the theoretically predicted equilibrium for H_2S–benzene, but instead to a configuration that suggests an alkyl–π interaction involving the carbon adjacent to the sulfur atom. We believe our findings indicate that environmental effects within proteins are altering the energetics of the S–π interaction so that other functional groups are preferred for interacting with the aromatic ring.

CHAPTER I

INTRODUCTION

1.1 Theoretical Study of Noncovalent Interactions

Noncovalent interactions involving aromatic systems are a key factor in many areas of biochemistry and molecular engineering. In particular, π–π interactions play a major role in protein folding, base-pair stacking in DNA, the mechanics of drug intercalation into DNA, and molecular self-assembly. Unfortunately, these interactions are typically very weak (< 5 kcal mol^{-1}), which makes it difficult to study them experimentally. This is less of a problem for theoretical chemists, however, and a large body of research has been conducted on many different interactions, including π–π, cation–π, alkyl–π, amino–π, oxygen–π, and sulfur–π.[1–9]

1.2 Thesis Objectives

The overall theme of the work in this thesis is the bridge between theoretical prototypes and real systems. The first chapter focuses on the π–π interaction and how it is influenced by environment in larger π–systems such as the stacked base-pairs of DNA or the crystal structure of polymers like polystyrene. An assortment of benzene trimers and tetramers are analyzed and then compared to the prototypical benzene dimer in order to ascertain the magnitude of any non-additive effects present in the larger systems.

The rest of the thesis focuses on the hydrogen sulfide–benzene dimer for use

as a prototype of S–π interactions. First, a detailed theoretical study of the H_2S–benzene dimer is conducted. High-level theoretical methods are used in order to obtain very accurate potential energy curves, including an equilibirum interaction energy extrapolated to the complete-basis-set (CBS) limit. The different basis sets used are examined, and it is found that the Dunning augmented correlation-consistent basis sets (aug-cc-pVXZ), with their full complement of diffuse functions, are better suited to the study of weak interactions than the Pople 6-31+G* family of basis sets, which lack diffuse polarization functions. Symmetry-adapted perturbation theory (SAPT) is used to decompose the interaction energy into electrostatic, exchange-repulsion, dispersion, and induction elements in order to more fully understand the energetics of the interaction.

The results of this study are then used as a basis for comparison in a data-mining study of protein crystal structures from the Protein Data Bank (PDB). Contacts between sulfur-bearing residues (cysteine and methionine) and aromatic residues (phenylalanine, tyrosine, and tryptophan) are searched for, and their geometric parameters are recorded and tabulated. 3-D histograms plotting the sulfur-to-ring-center distance vs. polar angle vs. frequency are created for various subsets of the residues in order to understand how the differences between the residues affect their ability to participate in S–π interactions. Unexpectedly, two other geometries are found in higher frequencies than the lowest-energy, hydrogen-bonding configuration. The possibility of groups competing for the H-bonding position is proposed to account for this.

CHAPTER II

OVERVIEW OF THEORETICAL METHODS

This section will focus primarily on the *conceptual* properties of the various theoretical methods used in the following studies. For detailed derivations and discussion of the mathematical properties of the methods, the author recommends Atilla Szabo and Neil S. Ostlund's *Modern Quantum Chemistry* (Dover, 1996)[10] and Frank Jensen's *Introduction to Computational Chemistry* (Wiley, 2003).[11]

2.1 The Schrodinger Equation

The ab initio methods of electronic structure theory attempt to solve the time-independent, non-relativistic Schrdinger equation

$$H\Psi = E\Psi$$

The Hamiltonian operator H can be partitioned into operators denoting the kinetic and potential energies of the nuclei and electrons.

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee} + \hat{V}_{NN}$$

The Born-Oppenheimer approximation says that because the nuclei are *much* more massive than the electrons and thus move much more slowly, the electrons can be considered to be moving in a field of fixed nuclei. Under this approximation, the nuclear kinetic energy T_N becomes zero, and the nuclear potential energy V_{NN} becomes a constant that can be calculated independently of the electronic energy.

The remaining three terms are called the *electronic Hamiltonian* since they depend directly on the electron coordinates (position and momentum) only.

Of these three terms, the first two only depend on the coordinates of a single electron; thus, they are usually considered together as a one-electron operator. The third term depends on the coordinates of two electrons and is called a two-electron operator. These two operators are the primary focus of Hartree-Fock theory.

2.2 Hartree-Fock Theory

Hartree-Fock theory (HF) is based on the approximation that each electron feels only an *average* electric field from the other electrons in the system. This is achieved by representing the wavefunction as a Slater determinant in which each electron is simultaneously associated with every orbital.

$$\Psi_{HF} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(x_1) & \chi_2(x_1) & \dots & \chi_n(x_1) \\ \chi_1(x_2) & \chi_2(x_2) & \dots & \chi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(x_n) & \chi_2(x_n) & \dots & \chi_n(x_n) \end{vmatrix}$$

The Hartree-Fock energy, then, is the expectation value of the electronic Hamiltonian within this wavefunction plus the nuclear potential energy.

$$E_{HF} = \left\langle \Psi_{HF} | \hat{H}_e | \Psi_{HF} \right\rangle + V_{NN}$$

The expectation value of $\hat{H}_e$ is given by two summations: one over simple one-electron intergrals, and the other over two-electron integrals. (For details of the derivations, see the references above.) The two-electron integrals are composed of *coulomb* integrals, J, and *exchange* integrals, K.

$$J_{ij} = \int dx_1 dx_2 \chi_i^*(x_1) \chi_j^*(x_2) \frac{1}{r_{ij}} \chi_i(x_1) \chi_j(x_2)$$

$$K_{ij} = \int dx_1 dx_2 \chi_i^*(x_1)\chi_j^*(x_2)\frac{1}{r_{ij}}\chi_j(x_1)\chi_i(x_2)$$

(These equations are given over spin orbitals $\chi_i(x_j)$; similar equations exist over spatial orbitals $\phi_i(x_j)$.)

The J integral is structured such that, when spin is integrated out, each orbital's spin function combines with itself, always giving 1. The K integral, however, is structured such that each orbital's spin function will combine with the *other* orbital's spin function. If the two electrons are of opposite spins, they will cancel each other out; only if the two electrons have parallel spins can K have a non-zero value. Because of this, we say that the HF method *correlates*, to a certain extent, the motions of electrons with parallel spin. Electrons with opposite spin, however, remain uncorellated in HF theory. This is the main failing of HF theory: even though the HF energy typically accounts for 99% of the true energy of a system, the missing *correlation energy* can be quite essential for the proper description of chemical phenomena. In order to "recover" this correlation energy that the HF method overlooks, other methods such as those discussed below are applied as refinements of the HF method.

2.3 Many-Body Perturbation Theory

The basic idea of Many-Body Perturbation Theory (MBPT) is to partition the *total* Hamiltonian of the system into two parts, the zeroth-order H_0 and a perturbation H'. The eigenvalues and eigenfuctions are known for H_0 with a reference wavefunction, $\Psi^{(0)}$. The perturbation H' is applied to the reference in order to generate "perturbed" wavefunctions of first-order ($\Psi^{(1)}$), second-order ($\Psi^{(2)}$), and so on. The expectation values of the pertubation operator between the reference and each perturbed wavefunction are added to the reference energy as first-, second-,

etc.-order energy corrections ($E^{(1)}$, $E^{(2)}$, etc.).

The specific flavor of MBPT used in this thesis is Møller-Plesset Perturbation Theory, or MPPT. In MPPT, the Hartree-Fock wavefunction is used as the reference, and the Hamiltonian is defined such that the sum of the zeroth-order energy and the first-order energy correction is the HF energy:

$$E_{MP1} = E^{(0)} + E^{(1)} = E_{HF}$$

Thus, the second-order MPPT correction constitutes the first improvement on the HF energy. Third-order and higher corrections are typically much smaller than the second-order correction, and the computational cost of obtaining these corrections scales exponentially. Thus, the most commonly used form of MPPT is that which stops at the second-order correction; this is called Second-Order MPPT, or MP2.

The main advantage of MPPT is that it is *size-extensive*, that is, the energy of a two-part system computed with the fragments at infinite separation will be equal to the sum of the individually computed energies for the two fragments. The HF method is not size-consistent.

2.4 Coupled-Cluster Theory

In coupled-cluster theory, the wavefunction is expressed as an exponential product of the reference, which is typically the HF wavefunction.

$$\Psi_{CC} = e^{\hat{T}}\Psi_{HF}$$

The exponential can be expanded as:

$$e^{\hat{T}} = 1 + \hat{T} + \frac{\hat{T}^2}{2!} + \frac{\hat{T}^3}{3!} + \dots$$

where the operator $\hat{T}$ is defined as:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \ldots + \hat{T}_N$$

for an N-electron system. The $\hat{T}_1$ operator generates singly-excited wavefunctions from the HF reference, $\hat{T}_2$ generates doubles, and so on.

In order to be computationally feasible, the exponential expansion and the operator $\hat{T}$ must be truncated. For the coupled-cluster singles and doubles method (CCSD), $\hat{T}$ is defined as simply $\hat{T}_1 + \hat{T}_2$. When this is put into the cluster expansion, terms such as $\hat{T}_1^2$ and $\hat{T}_2^3$ are produced. These are called "disconnected" excitations because they produce higher-order excitations as products of lower-order ones. For example, $\hat{T}_2^2$ produces a quadruple excitation by combining two doubles. Thus, any given level of CC theory will include contributions from higher-order excitations through these disconnected terms.

Adding $\hat{T}_3$ to the definition of $\hat{T}$ gives the CCSDT (CCSD + triples) method, which, while being very accurate, is also very computationally expensive. Instead, the CCSD(T) method is typically used, where the triple excitations are calculated by a perturbative method rather than in the cluster expansion. This method is commonly called the "gold standard of quantum chemistry" for its high accuracy coupled with size extensivity.

2.5 Symmetry-Adapted Perturbation Theory

Similarly to MBPT, Symmetry-Adapted Perturbation Theory (SAPT) improves on the HF wavefunction by means of a perturbation to the Hamiltonian. The SAPT Hamiltonian is:

$$\hat{H} = \hat{F} + \hat{V} + \hat{W}$$

where $\hat{F}$ is the Fock operator, which acts as a sum of individual Fock operators for each monomer; $\hat{W}$ is the intramonomer correlation operator, similar to an MP2 perturbation on each monomer; and $\hat{V}$ is the intermolecular interaction operator. Each of the physical components of the interaction energy (electrostatic, exchange-repulsion, induction, and dispersion) may be written in terms of different orders of perturbation for the $\hat{V}$ and $\hat{W}$ operators.

SAPT gives interaction energies for weakly bound dimers that are typically within 1-2% of the corresponding MP2 energies,[12–16] although at a much higher computational cost. The benefit of SAPT is the energy decomposition, which provides a great deal of insight into the energetics of an interaction.

CHAPTER III

BEYOND THE BENZENE DIMER: AN INVESTIGATION OF THE ADDITIVITY OF π–π INTERACTIONS

[Previously published in *J. Phys. Chem. A*, **2005**]

3.1 Introduction

Noncovalent interactions are fundamental to supramolecular chemistry, drug design, protein folding, crystal engineering, and other areas of molecular science.[17] In particular, π–π interactions between aromatic rings are ubiquitous in biochemistry and they govern the properties of many organic materials. Aromatic side-chains in proteins are often found in pairs due to the favorable energetics of the π–π interaction,[18,19] and certain drugs utilize π–π interactions to intercalate into DNA.[20] The fundamental physics of individual π–π interactions has been a subject of several high-level quantum mechanical studies,[1–5] but demonstrable convergence of the results even for the prototype benzene dimer has been achieved only recently[4] due to the extreme sensitivity of the results to electron correlation and basis set effects.

In many instances, an aromatic ring may be involved in more than one π–π interaction at a time, such as the stacking of nucleic acid bases in the double-helical

structure of DNA. In proteins as well, aromatic side-chains can be found in clusters; for example, the carp parvalbumin protein (P3CPV) exhibits a cluster of 7 phenylalanine residues. Burley and Petsko observed that 80% of the aromatic pairs they identified in a protein data bank (PDB) search were involved in "pair networks" as opposed to being isolated pairs.[18] Additionally, self-assembled stacks of aromatic macrocycles have been studied as possible molecular wires.[21] It is therefore critical to understand whether the properties of π–π interactions, as understood from prototype studies of benzene dimers, change significantly when they occur in clusters due to polarization or other many-body effects.

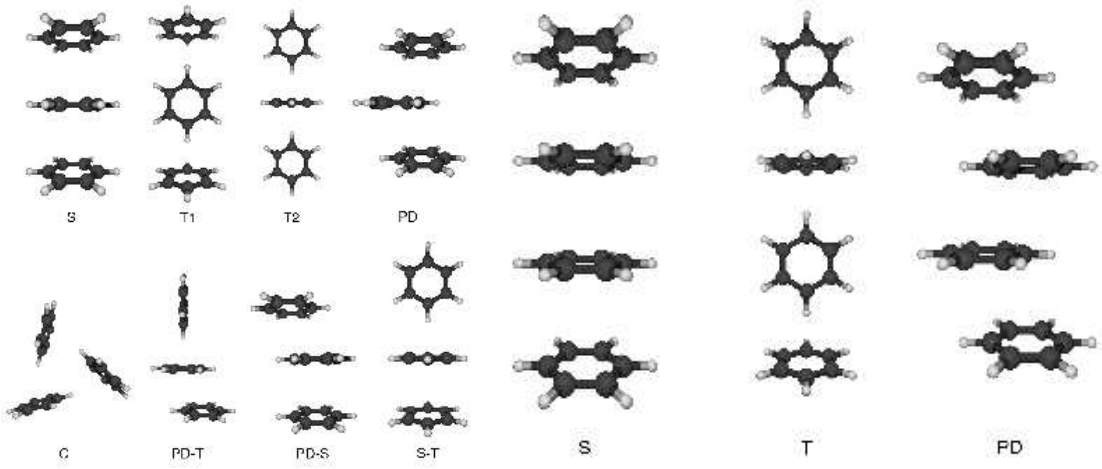
Some work along these lines was performed by Engkvist et al.,[22] who used simple potentials derived from CCSD(T) energies for benzene dimer to find and analyze local minimum structures on the trimer and tetramer potential energy surfaces. While their objective was to explore the potential energy surfaces and shed light on benzene cluster experiments, they did note that the two observed linear trimers ("H" and "double-T", which we call T1 and T2, respectively; see Fig. 3.1) had an interaction energy about twice that of the T-shaped dimer, and that the cyclic trimer (C, Fig. 3.1) had a total energy about three times that of the dimer. More recent ab initio results have been reported by Ye et al.,[23] who performed density functional theory (DFT) and second-order Møller-Plesset perturbation theory (MP2) computations for small benzene clusters in a parallel-displaced (PD) configuration as a model of π-stacks in polystyrene. In accord with other studies of weak interactions (see, e.g., ref 24), these authors found DFT to be unreliable for π-stacking. Their MP2 results indicated that the interaction energy for five benzenes (-7.09 kcal mol^{-1}) was somewhat larger than one might expect by thinking of the pentamer simply as four benzene dimers (-6.24 kcal mol^{-1} at the same level

of theory). This implies that something other than nearest-neighbor two-body interactions (i.e., benzene dimers) is making a significant contribution to the total interaction.

To better understand and model clusters of aromatic systems, it is important to understand the nature and magnitude of these other contributions, and to determine the relative magnitude of the different kinds of contributions (two-body vs three-body, nearest-neighbor vs non-nearest-neighbor, etc.). Here, we consider these different contributions in benzene trimers and tetramers consisting of various combinations of the prototypical configurations of the benzene dimer: the sandwich (S), T-shaped (T), and parallel-displaced (PD) configurations (see Fig. 3.1). These configurations are chosen as interesting prototypes, but it is not our objective to survey and identify the lowest-energy configurations of the benzene trimer and tetramer. Nevertheless, we also consider the cyclic configuration of the trimer, which according to experiment should be the most stable.[25] In addition, our inclusion of diffuse functions, found to be critical in previous work[4] but neglected in the MP2 computations of Ye et al.,[23] allows us to examine their role in the additivity of these interactions.

3.2 Theoretical Methods

Due to the large size of these systems, we were unable to apply the very high levels of theory we previously applied to the benzene dimer.[4] However, we have observed that MP2 in conjunction with small basis sets tends to exhibit a fortuitous cancellation of errors: small basis sets underestimate binding, while MP2 overestimates binding. We found that a modified aug-cc-pVDZ basis, which we will designate cc-pVDZ+, provides interaction energies within a few tenths of 1 kcal mol^{-1} of



(a) Trimers (b) Tetramers

Figure 3.1: Eight benzene trimer configurations and three benzene tetramer configurations considered in this study.

our previous estimates of the complete basis set coupled-cluster [CCSD(T)] limit for the geometries considered. The cc-pVDZ+ basis is the usual cc-pVDZ basis plus the diffuse s and p functions on carbon from the aug-cc-pVDZ basis. At the MP2/cc-pVDZ+ level of theory, using the geometries given below, we predict dimer interaction energies of -1.87 (sandwich), -2.84 (parallel-displaced), and -2.35 kcal mol^{-1} (T-shaped), while our previous estimates of the CCSD(T)/complete-basis-set values[4] were -1.81, -2.78, and -2.74 kcal mol^{-1}, respectively.

To compute the three- and four-body interaction terms between the monomers, we used a modified version of the Boys-Bernardi counterpoise correction[26] developed by Hankins, Moskowitz, and Stillinger,[27] which defines the many-body interactions in terms of the lower-order interaction energies. For a trimeric system, the total energy would be:

$$E_{tot} = \sum_i E(i) + \sum_{ij} \Delta^2 E(ij) + \Delta^3 E(123)$$

where

$$\Delta^2 E(ij) = E(ij) - E(i) - E(j)$$

$$\Delta^3 E(123) = E(123) - \sum_i E(i) - \sum_{ij} \Delta^2 E(ij)$$

and all computations are performed using the full basis of the trimer. The scheme can be extended for tetramers (denoting the four-body terms as $\Delta^4 E$) or larger clusters.

For simplicity, we use rigid monomers with parameters recommended by Gauss and Stanton[28] [r_e(C-C) = 1.4079 Å and r_e(C-H) = 1.0943 Å]; our previous work[4] indicates that there is almost no relaxation of monomer geometries when the dimers are fully optimized. We also used intermonomer parameters previously determined[4] at the MP2/aug-cc-pVDZ level of theory for the dimers [R_S = 3.8 Å, R_T = 5.0 Å, R^1_{PD} = 3.4 Å, and R^2_{PD} = 1.6 Å]. Tests of the sandwich trimer show that optimizing the intermonomer distances results in only a 0.05 Å increase from the dimer distance of 3.8 Å, a 0.03 kcal mol^{-1} change in the total energy, and changes on the order of 0.01 kcal mol^{-1} in the various many-body terms. With the assumption that all systems will exhibit the same magnitude of changes upon similar optimization, such optimization does not appear to be necessary for the purposes of this study. For the cyclic or C-trimer configuration, which experiment suggests is the lowest-energy configuration,[25] we were unable to find any geometric parameters in the literature. However, we found that the MP2/cc-pVDZ+ equilibrium geometry for this configuration (subject to C3h symmetry) has a 4.8 Å intermonomer (center-to-center) separation with each monomer tilted 12° away from perpendicular.

3.3 Results and Discussion

Theoretical results for the trimers are summarized in Table 3.1. The reported values $\Delta^2 E$ and $\Delta^3 E$ are the sum of individual two- and three-body interaction energies, respectively, for the given trimer. A few general trends are readily apparent from the table. One is that the nearest-neighbor two-body energies [$\Delta^2 E(1)$ and $\Delta^2 E(23)$] are in every case slightly larger than the corresponding benzene dimer energy. This is a result of the ghost functions from the additional monomers stabilizing the "dimer" systems when considered in the full basis of the trimer/tetramer. A second trend is that in all systems besides the C-trimer (which only has nearest-neighbor two-body interactions), the long-distance two-body interactions [$\Delta^2 E(13)$] are generally small but stabilizing contributions to the overall interaction. On the other hand, the three-body interaction terms ($\Delta^3 E$) are mostly small but destabilizing. For the C-trimer, the three-body term is definitely significant – more than 0.3 kcal mol^{-1} – which might be expected because the C-trimer is a true three-body system, with each monomer having a close interaction with both of the other monomers. Because the three-body and long-distance two-body terms are small, one might expect that the binding energies of these trimers might be reasonably well estimated simply from the sum of (nearest-neighbor) benzene dimer energies at these geometries, a quantity we denote . As shown in Table 3.1, this simple sum-of-dimers estimate is rather good, within 0.3 kcal mol^{-1} of the explicitly computed values for all but the C trimer, where the difference is 0.6 kcal mol^{-1}.

In the tetramers, the results for which are summarized in Table 3.2, we see similar trends in regards to the two-body interactions: nearest-neighbor interactions are slightly more stabilizing than those in the isolated dimer, and long-distance interactions are, individually, relatively small. For the three-body interactions, the

Table 3.1: Total and Many-Body Interaction Energies (kcal mol^{-1}) of Various Benzene Trimers at the MP2/cc-pVDZ+ Level of Theory

	S	PD	T1	T2	C	S/PD	S/T	PD/T
$\Delta^2E(12)$	-1.93	-2.91	-2.37	-2.38	-2.52	-1.95	-1.95	-2.90
$\Delta^2E(13)$	-0.01	-0.05	0.01	-0.03	-2.52	-0.04	-0.12	-0.14
$\Delta^2E(23)$	-1.93	-2.91	-2.37	-2.38	-2.52	-2.90	-2.39	-2.39
Δ^2E	-3.87	-5.88	-4.72	-4.80	-7.55	-4.88	-4.46	-5.42
Δ^3E	0.034	0.000	0.078	0.064	-0.33	0.023	-0.026	0.001
E_{tot}	**-3.83**	**-5.88**	**-4.64**	**-4.73**	**-7.88**	**-4.86**	**-4.49**	**-5.42**
E_{dimer} [a]	-3.74	-5.68	-4.70	-4.70	-7.32	-4.71	-4.22	-5.19

[a] E_{dimer} is the predicted interaction energy based on a simple sum of (nearest-neighbor) benzene dimer energies. The MP2/cc-pVDZ+ interaction energies of benzene dimer at these geometries are -1.87 (S), -2.84 (PD), -2.35 (T), and -2.44 kcal mol^{-1} (C).

two all-nearest-neighbor terms $\Delta^3E(123)$ and $\Delta^3E(234)$ correspond very closely to the three-body term for the trimer, while the other two terms are essentially zero, such that the tetramer Δ^3E is essentially the sum of the two Δ^3E's from the trimers (123) and (234). The four-body terms are negligible for all cases, being no more than a hundredth of 1 kcal mol^{-1}. Although the new types of interactions (four-body and non-nearest-neighbor three-body terms) are negligible, the larger number of long-distance two-body terms and all-nearest-neighbor three-body terms leads to larger deviations from the simple sum-of-dimers estimate than was observed for the trimers (except for the T-tetramer, which shows a fortuitous agreement with the sum-of-dimers estimate). The aggregate effects of long-distance two-body terms and all-nearest-neighbor three-body terms will become more significant on an absolute basis for larger clusters and would need to be included if accurate total binding energies are required. Fortunately, however, it should be possible to obtain good estimates of these effects simply from trimers. Overall, we observe deviations from the

Table 3.2: Total and Many-Body Interaction Energies (kcal mol^{-1}) of Various Benzene Tetramers at the MP2/cc-pVDZ+ Level of Theory

	S	PD	T
$\Delta^2E(12)$	-1.94	-2.93	-2.37
$\Delta^2E(13)$	-0.01	-0.06	0.01
$\Delta^2E(14)$	0.01	0.01	-0.01
$\Delta^2E(23)$	-1.98	-2.97	-2.39
$\Delta^2E(24)$	-0.01	-0.06	-0.03
$\Delta^2E(34)$	-1.94	-2.93	-2.39
Δ^2E	-5.87	-8.94	-7.17
$\Delta^3E(123)$	0.035	0.000	0.077
$\Delta^3E(124)$	0.005	0.002	-0.006
$\Delta^3E(134)$	0.005	0.002	-0.005
$\Delta^3E(234)$	0.035	0.000	0.062
Δ^3E	0.079	0.004	0.127
Δ^4E	-0.0002	-0.0012	-0.0050
E_{tot}	**-5.80**	**-8.94**	**-7.05**
E_{dimer}[a]	-5.61	-8.52	-7.05

[a] See footnote *a* on Table 3.1.

sum-of-dimers estimate of about 0.4 kcal mol^{-1} or less for the tetramer stacks considered. This is considerably smaller than the 0.85 kcal mol^{-1} deviation noted for the slightly larger PD pentamer system (with a somewhat different geometry) considered by Ye et al.[23] Given the similarity between the two- and three-body terms obtained for the trimers and tetramers, we can reasonably assume that they remain similar for the pentamer, allowing us to obtain a simple estimate of the interaction energy that would be obtained by adding one more benzene to our PD tetramer. This estimate yields -11.96 kcal mol^{-1}, giving a deviation of 0.6 kcal mol^{-1} from our sum-of-dimers estimate. The remaining 0.25 kcal mol^{-1} difference between our

Table 3.3: Total and Many-Body Interaction Energies (kcal mol^{-1}) of Various Benzene Trimers at the CCSD(T)/cc-pVDZ+ Level of Theory

	S	PD	T1	C
$\Delta^2E(12)$	-0.48	-0.92	-1.62	-1.61
$\Delta^2E(13)$	0.02	-0.01	0.02	-1.61
$\Delta^2E(23)$	-0.48	-0.92	-1.62	-1.61
Δ^2E	-0.94	-1.85	-3.22	-4.84
Δ^3E	0.038	0.014	0.072	-0.25
E_{tot}	**-0.90**	**-1.84**	**-3.14**	**-5.09**
E_{dimer}[a]	-0.86	-1.72	-3.20	-4.62

[a] E_{dimer} is the predicted interaction energy based on a simple sum of (nearest-neighbor) benzene dimer energies. The CCSD(T)/cc-pVDZ+ interaction energies of benzene dimer at these geometries are -0.43 (S), -0.86 (PD), -1.60 (T), and -1.54 kcal mol^{-1} (C).

estimate of this deviation and that of Ye et al. may be ascribed to the different geometries and basis sets employed. We also note, however, that the lack of diffuse functions in the MP2 computations of Ye et al. leads to considerably smaller total interaction energies, making the discrepancy from the sum-of-dimers estimate larger on a percentage basis. Overall, the differences between our ab initio interaction energies and the simple sum-of-dimers estimates are 1-6% for the trimers (7% for the C-trimer), 0-5% for the tetramers, and 5% for the pentamer (estimated) versus 12% from the work of Ye et al.

It is important to determine whether the near-addivity of the interaction energies persists when higher-level treatments of electron correlation are employed. Therefore, we performed CCSD(T)/cc-pVDZ+ calculations on four of the trimers,

the results of which are summarized in Table 3.3. While the total interaction energies and the nearest-neighbor two-body terms vary greatly from the MP2 energies in Table 3.1 (consistent with our previous work[4]), the magnitudes of the three-body terms are very similar to those computed via MP2, demonstrating that these three-body terms do not depend greatly on the computational method employed. On a percentage basis, the deviations from the sum-of-dimers estimates are 2-7% for the S, PD, and T1 configurations, and a somewhat larger 9% for the C trimer. It is interesting to note that the total energies for the T1 and C systems here are quite similar to those reported by Engkvist et al.,[22] who, as noted above, used CCSD(T) results to calibrate their potential.

3.4 Conclusions

In conclusion, we have demonstrated that the interaction energies in larger benzene clusters are fairly close to what one might expect based simply on the sum of interaction energies for isolated benzene dimers, with an error of less than 10% for all systems considered. Two considerations keep this simple picture from being perfectly accurate:

1. Nearest-neighbor two-body interactions are stabilized by up to one tenth of 1 kcal mol^{-1} when computed in the basis set of the full system as opposed to the dimer basis.

2. Long-distance two-body interactions, as well as nearest-neighbor three-body terms, have an aggregate effect which will become increasingly important for the total binding energy of larger clusters (although these effects are readily estimated from trimers).

Fortunately, we find that four-body terms and three-body terms that include any non-nearest-neighbor monomer pairs are insignificant for the configurations considered and can be safely neglected.

Because the nearest-neighbor three-body terms are fairly insensitive to the electronic structure method, it seems worthwhile to use a less expensive method to determine these terms, while very accurate methods may be used to determine the dominating two-body terms. In this light, the recent multi-center model of Hopkins and Tschumper,[29] which employs high-level computations only on dimers and low-level computations on the entire cluster, is very promising.

CHAPTER IV

ESTIMATES OF THE AB INITIO LIMIT FOR SULFUR–π INTERACTIONS: THE H_2S–BENZENE DIMER

[Previously published in *J. Phys. Chem. A*, **2005**, *109*, 191.]

4.1 Introduction

Non-covalent interactions involving the aromatic side chains of certain amino acids are some of the most important factors in determining the dynamics of protein folding. The experimental and computational aspects of π–π, cation–π, alkyl–π, and amino–π interactions have been a subject of much recent interest.[30] One type of interaction that has not received as much attention computationally is the sulfur–π interaction, partly because it is not as common as the others in natural systems and partly because the presence of the sulfur atom increases the computational expense.

Morgan et al.[31] first proposed the hypothesis that strong and favorable S–π interactions exist after identifying chains of alternating "sulfur and π-bonded atoms" in the crystal structures of eight different proteins. This finding suggested that S–π stacking might play a significant role in stabilizing the folded conformations of these proteins. Database searches performed by Morgan et al.[32] and Reid et al.[33] on the Brookhaven Protein Data Bank,[34] and by Zauhar et al.[7] on the Cambridge

Crystallographic Database,[35] all confirmed that S–π interactions occurred more frequently than expected from the random association of amino acids.

Viguera and Serrano[36] directly investigated the contribution of S–π interactions to the stability of α-helices by calculating the helical content of a model protein from NMR and circular dichroism spectra. The AGADIR[37] algorithm, which calculates the helical content of peptides, was then parameterized in order to reproduce the experimental results; the optimized parameters gave interaction free energies of -2.0 kcal mol^{-1} for phenylalanine-cysteine interactions and -0.65 kcal mol^{-1} for phenylalanine-methionine.

Cheney et al.[38] performed a quantum mechanical study on the methanethiol-benzene system as a model of cysteine-aromatic interactions. They optimized various initial configurations using Hartree-Fock theory (HF) with the 3-21G* Pople basis set and subsequently performed single-point calculations using second-order Møller-Plesset perturbation theory (MP2) and the 6-31G* basis set. The optimum configuration was found at a distance of 4.4 Å between the sulfur and the center of the benzene ring and an angle of 56° between the line joining these two points and the plane of the benzene ring. The interaction energy for this geometry was computed as -3.0 kcal mol^{-1}. A more recent study by Duan et al.[8] utilized much larger basis sets, up to 6-311+G(2d,p). Using three different starting geometries, they first optimized the methanethiol-benzene dimer at the MP2/6-31G** level of theory and then performed single-point calculations using the larger basis sets in order to construct potential energy surfaces. Their results show that the equilibrium for the lowest-energy conformation (with the sulfur over the center of the ring) is at 3.73 Å separation, which gave an interaction energy of -3.71 kcal mol^{-1}. Using their results from a previous study of the methane–benzene dimer, they were able

to isolate the contribution of only the SH–π interaction, which they said "should be greater than 2.6 kcal mol^{-1}." To our knowledge, these MP2/6-311+G(2d,p) calculations are the highest level of theory previously applied to a S–π complex. However, our previous experience with weak interactions in the benzene dimer suggests that this level of theory might be far from convergence.[4]

A molecular mechanics study of site-directed mutagenesis in staphylococcal nuclease by Yamaotsu et al.[39] reported a quite shocking result: they found that an M32L substitution (substituting leucine for the methionine at position 32) resulted in a structure that was 1.6 kcal mol^{-1} *more* stable than the wild type peptide, which is unusual because peptide mutations normally result in less stable protein structures. The mutant protein was subsequently synthesized by Spencer and Stites,[40] who reported a *decrease* in stability of 0.8 kcal mol^{-1} compared to the wild type, a much more conventional result. These results inspired Pranata[41] to perform a theoretical study on the dimethyl sulfide-benzene system using both quantum mechanical and molecular mechanical methods. Although his MM results did not agree with Yamaotsu's results using the same force field, they were in good agreement with his MP2/6-31G* QM results, which all showed that the M32L substitution was destabilizing.

Here we present high-level quantum mechanical predictions for the simplest possible prototype ofS–π interactions, the H_2S–benzene dimer. Not only is this system a prototype of S–π interactions in biological contexts, but H_2S is a typical oil-gas component, and its interaction with benzene is important in modeling vapor-liquid equilibria relevant to oil and gas processing.[42]

At present, very few high-quality potential energy curves are available for prototype noncovalent interactions. However, such results are crucial for calibrating

new methods aimed at modeling these interactions reliably and efficiently. Coupled cluster theory through perturbative triple substitutions, CCSD(T),[43] is often referred to as the "gold standard of quantum chemistry" and is very reliable for such studies. In validating his new density functional theory plus dispersion model, Grimme[44] has observed that "very accurate CCSD(T) data are still missing" for complexes of benzene with small molecules. Here we use CCSD(T) with very large basis sets, up through augmented correlation consistent polarized valence quadruple-zeta[45] (aug-cc-pVQZ). Note that this augmented basis set, which includes a set of diffuse functions for every angular momentum present in the basis, is much larger than the cc-pVQZ basis set. The potential energy curves thus obtained should be of "subchemical" accuracy — within a few tenths of 1 kcal mol^{-1}. Our results for the equilibrium geometry of the complex will be compared to recent microwave experiments by Arunan et al.[46] In addition, the reliability of less complete levels of theory for S–π interactions will be evaluated in light of our benchmark results. These comparisons will be valuable in determining appropriate levels of theory for studies of larger models of S–π interactions.

4.2 Theoretical Methods

Energy computations using second-order Møller-Plesset perturbation theory (MP2), coupled-cluster theory with single and double substitutions (CCSD), and coupled-cluster including perturbative triple substitutions [CCSD(T)] were performed using various basis sets.[47] Rigid monomer geometries were used, according to best values in the literature: r_e(C-C) = 1.3915 Å and r_e(C-H) = 1.0800 Å for benzene,[28] and r_e(S-H) = 1.3356 Å and θ_e(H-S-H) = 92.12° for hydrogen sulfide.[48] The monomers were oriented with the sulfur atom directly over the center of the

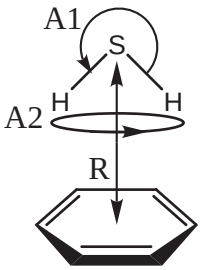


Figure 4.1: Geometry specification for the H_2S–benzene dimer. $A1$ is in the C_{2v} plane of the complex, $A2$ is centered on the C_{2v} axis, and the intermonomer distance R is measured from the center of the benzene ring to the sulfur atom.

benzene ring, such that the C_{2v} axis of H_2S matches the C_{6h} axis of benzene (Figure 4.1). Potential energy curves (PECs) for the "swing" angle, $A1$, and the "twist" angle, $A2$, were obtained at the CCSD(T)/aug-cc-pVDZ level of theory in order to determine the optimum values of these parameters for later computations. The intermonomer distance R was held fixed at 3.9 Å for these computations.

PECs over the intermonomer distance R were then obtained with the MP2, CCSD, and CCSD(T) methods in conjunction with the 6-31+G*, aug-cc-pVDZ, and aug-cc-pVTZ basis sets. MP2 curves were also obtained with the very large aug-cc-pVQZ basis set (932 functions). Taking advantage of the relative insensitivity to basis set of the difference between CCSD(T) and MP2 energies, we estimate the CCSD(T)/aug-cc-pVQZ energies as follows:

$$E_{\rm int}^{\rm CCSD(T)/aug\text{-}cc\text{-}pVQZ} = E_{\rm int}^{\rm MP2/aug\text{-}cc\text{-}pVQZ} + \delta_{\rm MP2}^{\rm CCSD(T)},$$

where

$$\delta_{\rm MP2}^{\rm CCSD(T)} = E_{\rm int}^{\rm CCSD(T)/aug\text{-}cc\text{-}pVTZ} - E_{\rm int}^{\rm MP2/aug\text{-}cc\text{-}pVTZ}$$

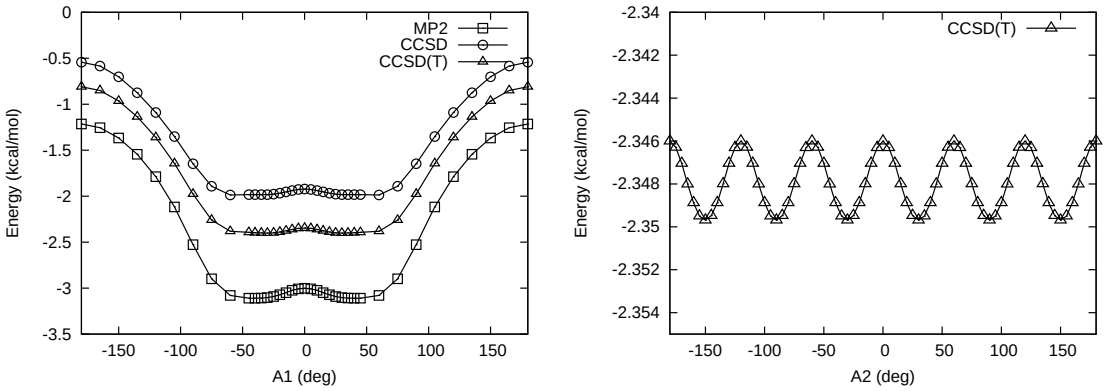
is calculated from the interaction energies computed with a smaller basis set, in this case, aug-cc-pVTZ.

With the availability of these high-quality results, we decided to assess the

reliability of some smaller basis sets which have commonly been used for such calculations. Specifically, we obtained PECs for the 6-31++G** basis (for comparison to aug-cc-pVDZ) and the 6-311+G(2d,p) basis (used by Duan,[8] for comparison to aug-cc-pVTZ). We also obtained PECs for three modifications of the aug-cc-pVDZ basis: (1) aug(sp/sp)-cc-pVDZ, with the diffuse d-functions on carbon and sulfur removed; (2) aug(spd/s)-cc-pVDZ, with the diffuse p-function on hydrogen removed; and (3) aug(sp/s)-cc-pVDZ with both the d and p diffuse functions removed. The aug(sp/s)-cc-pVDZ basis has the same number and types of contracted functions as 6-31++G** with the only difference being in the number of primitive functions used, thus allowing us to directly compare the inherent quality of the Pople and Dunning basis sets for predictions of energies in van der Waals complexes.

The counterpoise (CP) correction method of Boys and Bernardi[26] was used to account for the basis set superposition error in all computations, since our previous results have shown that CP-corrected energies converge more quickly to the complete basis set limit for π–π interactions.[4] Core orbitals were constrained to remain doubly occupied in all correlated calculations. Calculations were performed in MOLPRO[49] running on an IBM SP2 supercomputer.

Symmetry-adapted perturbation theory (SAPT)[13,50] was employed to decompose the energy into physically meaningful components, including electrostatic, induction, dispersion, and exchange energies. The specifics of this method have been described in detail elsewhere.[51] The SAPT calculations reported here used the correlation level technically designated as SAPT2, and they were carried out using the aug-cc-pVDZ basis set at the CCSD(T)/aug-cc-pVQZ geometry. SAPT calculations were performed using the SAPT2002 program.[52]

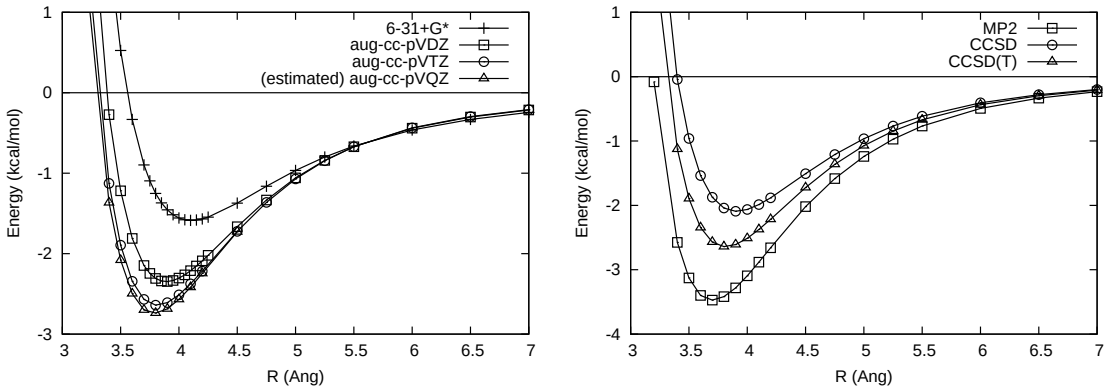


(a) PES for A1 (b) PES for A2

Figure 4.2: Potential energy curves over the two configuration angles, aug-cc-pVDZ basis.

4.3 Results and Discussion

The CCSD(T)/aug-cc-pVDZ curves showing the interaction energy as a function of the angles $A1$ and $A2$ are shown in Figure 4.2. The curve for $A1$ shows a shallow minimum around 30° from the starting geometry; this angle would have one of the hydrogens pointed almost directly down toward the center of the ring. However, the energy at this point is only 0.06 kcal mol^{-1} below the initial E_{int} of -2.35 kcal mol^{-1} at 0°. This difference is so small that the curve can be considered essentially flat near 0°. At 180°, the sulfur lone pairs are pointed down at the ring and the hydrogens are pointed away; the lone pair electrons interact much less favorably with the negatively charged π-cloud of the benzene, and the CCSD(T)/aug-cc-pVDZ interaction energy becomes only -0.81 kcal mol^{-1}. The corresponding curve for $A2$ is even flatter, showing very shallow minima (< 0.01 kcal mol^{-1}) at angles that place the H_2S hydrogens between the ring carbons. Because of this flatness in the potential energy surface of both parameters, and because setting A1 and A2 both equal to 0° gives the system C_{2v} symmetry, we decided to use this geometry



(a) Basis set (b) Correlation method

Figure 4.3: Effects of choice of basis set and correlation method for the H_2S–benzene dimer. All curves in (a) use the CCSD(T) method; all curves in (b) use the aug-cc-pVTZ basis set.

in order to reduce the cost of the computations.

The interaction energies as a function of intermonomer distance are shown in Figure 4.3. Figure 3(a) shows the effect of basis set size on the CCSD(T) results; the values obtained for R_{eq} and E_{int} are summarized in Table 4.1. The general trends in R_{eq} and E_{int} are readily observable: R_{eq} decreases and the magnitude of E_{int} increases (E_{int} becomes more negative) as the size of the basis increases. As the basis set becomes larger, the changes to E_{int} become smaller: between 6-31+G* and aug-cc-pVDZ, E_{int} increases by 0.8 kcal mol^{-1}, while it increases by only 0.3 kcal mol^{-1} between aug-cc-pVDZ and aug-cc-pVTZ, and only 0.1 kcal mol^{-1} between aug-cc-pVTZ and aug-cc-pVQZ. This is as expected, since the correlation consistent basis sets were designed around the principle of systematically converging the correlation energy correction with increasing basis size.[53]

This convergence can be estimated by correcting for the two main types of basis set error. The first is basis set superposition error, or BSSE, which arises

Table 4.1: Intermonomer Distance (Å) and Interaction Energy (kcal mol^{-1}) at Equilibrium for Various Levels of Theory.[a]

basis set	method	R_{eq}	E_{int}
6-31+G*	MP2	4.00	-1.92
	CCSD	4.15	-1.42
	CCSD(T)	4.10	-1.58
aug-cc-pVDZ	MP2	3.80	-3.06
	CCSD	3.95	-1.94
	CCSD(T)	3.90	-2.34
aug-cc-pVTZ	MP2	3.70	-3.47
	CCSD	3.90	-2.09
	CCSD(T)	3.80	-2.64
aug-cc-pVQZ	MP2	3.70	-3.60
	CCSD(T)	(3.80)[b]	(-2.74)[b]
CBS	CCSD(T)		-2.81[c]

[a] All energies include counterpoise corrections.
[b] CCSD(T)/aug-cc-pVQZ results are estimated as described in the text.
[c] Complete basis set extrapolation at the CCSD(T)/aug-cc-pVQZ geometry.

because each monomer in the complex can artificially lower its energy by "borrowing" basis functions from the other monomer, so that the attraction between the two monomers is overestimated; the recommended procedure for eliminating BSSE is the counterpoise correction,[54] which we have applied to all of our results. The second main basis set error is the basis set incompleteness error, or BSIE, which is a consequence of the incomplete description of the electronic Coulomb cusp. In an examination of hydrogen-bonded systems, Halkier and co-workers[55] developed a two-point extrapolation scheme to correct for the BSIE which has the following simple closed form:

$$E_{corr,lim} = \frac{X^3}{X^3-(X-1)^3}E_{corr,X} - \frac{(X-1)^3}{X^3-(X-1)^3}E_{corr,X-1},$$

where $E_{corr,X}$ is the correlation energy obtained with the correlation consistent basis set with cardinal number X (aug-cc-pVXZ). For the various hydrogen-bonded

systems they studied, it was found that a "3–4" MP2 extrapolation (i.e., using the MP2/aug-cc-pVTZ and MP2/aug-cc-pVQZ correlation energies) always gave results within 0.05 kcal mol^{-1} of the MP2-R12 basis set limit. Using the same "3–4" extrapolation here for the CCSD(T) correlation energies, and taking the CP-corrected SCF/aug-cc-pVQZ energy as our reference, we obtained an extrapolated, complete-basis-set (CBS) CCSD(T) limit $E_{\rm int}$ of -2.81 kcal mol^{-1}. This is an improvement of only 0.07 kcal mol^{-1} over our CCSD(T)/aug-cc-pVQZ results. Based on Halkier's results, and the good reliability of CCSD(T) for such problems, it seems certain that this result is within a few tenths of 1 kcal mol^{-1} of the true value.

$R_{\rm eq}$ and $E_{\rm int}$ show consistent trends with regards to correlation method, as well. Figure 3(b) compares the MP2, CCSD, and CCSD(T) potential energy curves with the aug-cc-pVTZ basis set. MP2 binds more strongly than CCSD(T) ($R_{\rm eq}$ is shorter, $E_{\rm int}$ is more negative), which binds more strongly than CCSD. This finding is consistent with the results of Hopkins and Tschumper,[56] who found the same trend in their study of various π-bonded dimers. They also concluded that the effects of triple excitations, included here via the (T) term in CCSD(T), is required in order to determine $E_{\rm int}$ to chemical accuracy. From the figure, we see that the difference between CCSD(T) and MP2, $\delta_{\rm MP2}^{\rm CCSD(T)}$, is largest at short distances and dies off to zero at large distances. This coupled-cluster correction, which was added to the MP2/aug-cc-pVQZ results to estimate the CCSD(T)/aug-cc-pVQZ level of theory, was found to be quite insensitive to the basis set. If we compute this coupled-cluster correction in the smaller aug-cc-pVDZ basis set instead, the largest discrepancy from the aug-cc-pVTZ values is only $0.03 - 0.04$ kcal mol^{-1} at small R. This suggests that the errors in $\delta_{\rm MP2}^{\rm CCSD(T)}$ computed with the aug-cc-pVTZ basis set are smaller than this.

Our best theoretical results compare very well with the experimental geometry of Arunan et al.[46] Those authors reported geometrical parameters of $A1 = 28.5°$ and $R_{eq} = 3.818$ Å; our CCSD(T) calculations showed minima at $A1$=30° (aug-cc-pVDZ basis) and R_{eq} =3.80 Å (aug-cc-pVQZ basis). The deviations from Arunan's results are well within the resolution of our curves, ±5° for $A1$ and ±0.1 Å for R_{eq}. Unfortunately, we could not find any reports in the literature of experimental interaction energies for this dimer. We can, however, compare our results to the theoretical results of Duan et al.,[8] who determined that the SH–π interaction in methyl sulfide should be $\sim$ 2.6 kcal mol^{-1} at the MP2/6-311+G(2d,p) level of theory. Their lower-level theoretical binding energy for the methylated model is roughly the same as our higher-level estimated CBS CCSD(T) limit of −2.81 kcal mol^{-1}. The binding energy of H_2S–benzene is also found to be very similar to that of H_2O–benzene, estimated by Tsuzuki et al.[57] as −3.17 kcal mol^{-1} using computational techniques similar to those employed here.

An interesting result from the comparison of basis set effects is the large difference between the 6-31+G* and aug-cc-pVDZ binding energies of 0.76 kcal mol^{-1}. Both are double-zeta basis sets with polarization and diffuse functions, with the exception that 6-31+G* does not include diffuse and polarization functions for hydrogen. To investigate this discrepancy, we performed computations with the 6-31++G** basis, which does include these functions. We also obtained results with the triple-zeta 6-311+G(2d,p) basis set, used by Duan et al.,[8] for comparison to the triple-zeta aug-cc-pVTZ basis. These results are summarized in Figure 4.4 and Table 4.2. It is readily apparent that the extra hydrogen functions provided by the 6-31++G** basis are not particularly important, as they only increased the magnitude of E_{int} by 0.054 kcal mol^{-1}; there is still a large discrepancy (0.713 kcal mol^{-1})

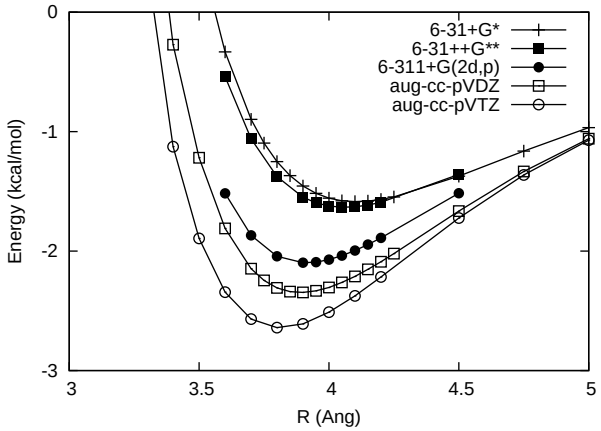


Figure 4.4: Comparison of Pople vs. unmodified Dunning basis sets.

Table 4.2: Intermonomer Distance (Å) and Interaction Energy (kcal mol^{-1}) at Equilibrium; Comparison between Pople and Dunning Basis Sets, CCSD(T) Method.

basis set	R_{eq}	E_{int}
6-31+G*	4.10	-1.58
6-31++G**	4.10	-1.63
aug-cc-pVDZ	3.90	-2.34
6-311+G(2d,p)	4.10	-2.02
aug-cc-pVTZ	3.80	-2.64

between the Pople 6-31++G** and Dunning aug-cc-pVDZ double-zeta basis sets. The only other difference between the 6-31++G** and aug-cc-pVDZ basis sets is that 6-31++G** only includes diffuse functions for the core and valence function sets — (1s1p/1s) — while aug-cc-pVDZ also includes diffuse functions for the polarization sets — (1s1p1d)/(1s1p). This led us to wonder whether these diffuse (1d/1p) functions could account for such a large difference, nearly a full kilocalorie per mole?

To investigate this possibility, we employed modified versions of the aug-cc-pVDZ basis set, as described in the Methods section. The results are displayed in Figure 4.5. Removing the diffuse d-functions on carbon and sulfur reduced the binding

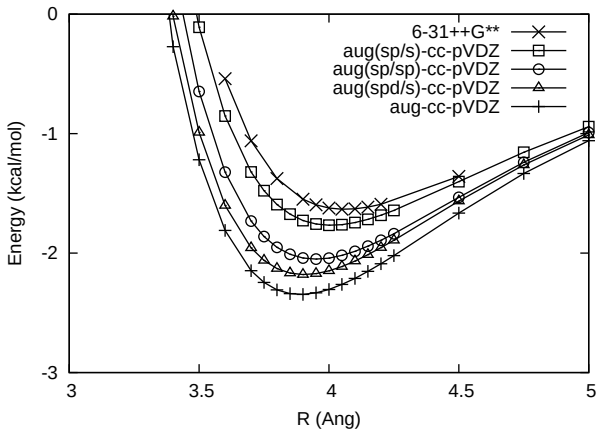


Figure 4.5: Comparison of Pople vs. modified Dunning basis sets.

energy by ~ 0.30 kcal mol^{-1}, almost half of the total difference between the basis sets. Removing the diffuse p-functions on hydrogen had half as great of an effect, reducing E_{int} by ~ 0.16 kcal mol^{-1}. Removing both sets of functions at the same time reduced E_{int} by ~ 0.62 kcal mol^{-1}, leaving a difference of only ~ 0.1 kcal mol^{-1} between 6-31++G** and aug(sp/s)-cc-pVDZ. It therefore appears that both the Pople and Dunning basis sets are very similar in fundamental quality, even though they do not use the same number of primitive Gaussians for the contractions of valence orbitals. On the other hand, the extra diffuse functions present in the augmented Dunning basis set make a fairly large contribution to lowering the interaction energy.

A similar discrepancy also appears to exist between the two triple-zeta basis sets, with a difference in E_{int} at equilibrium of 0.62 kcal mol^{-1}. The difference in the number of basis functions in these two basis sets is greater than the difference in the number of functions in the double-zeta sets: compared to aug-cc-pVDZ, aug-cc-pVTZ includes an additional set of (1d1f/1p1d) polarization functions, as well as and additional set of (1f/1d) diffuse functions. Besides the number of valence functions, 6-311+G(2d,p) only differs from 6-31++G** by an additional

Table 4.3: SAPT2/aug-cc-pVDZ Results for Contributions to the Interaction Energy (kcal mol^{-1}) at CCSD(T)/aug-cc-pVQZ Equilibrium Geometry.

	$A1 = 0°$	180°
E_{elst}	-2.37	0.01
E_{exch}	4.19	1.03
E_{ind}	-0.81	-0.17
E_{disp}	-4.16	-2.14
E_{int}(SAPT2)	-3.15	-1.27
E_{int}(MP2)	-3.06	-1.21

(1d) polarization function on heavy atoms and the lack of a diffuse (1s) function on hydrogen. The overall difference between 6-311+G(2d,p) and aug-cc-pVTZ is then composed of (1f/1p1d) polarization and (1d1f/1s1p1d) diffuse functions. Even though the difference in the number of functions is greater than that between the double-zeta basis sets, the magnitude of the difference in energies is slightly smaller; this is consistent with the systematic convergence of energies using the correlation consistent basis sets. Overall, the higher angular momentum diffuse functions in the correlation consistent basis sets, especially the diffuse d functions, contribute significantly to the overall interaction energy and should remain important in other van der Waals complexes.

The SAPT-derived components of the binding energy are summarized in Table 4.3. Although we were only able to perform the SAPT analysis at the SAPT2/aug-cc-pVDZ level of theory, which gives total binding energies very similar to those from counterpoise-corrected MP2/aug-cc-pVDZ, this level of theory features a favorable cancellation of basis set and correlation errors and yields a binding energy similar to that of CCSD(T)/aug-cc-pVQZ. To simplify the analysis, for present purposes we have designated the exchange-dispersion and exchange-induction terms as dispersion and induction, respectively. Additionally, the term $\delta E^{HF}_{int,resp}$, which includes third- and higher-order HF induction and exchange induction contributions, has been

designated as induction. From the table, we see that electrostatic terms make a fairly strong attractive contribution, -2.37 kcal mol^{-1}, arising primarily from the interaction between the partial positive charge on the H_2S hydrogens and the partial negative charge in the benzene π–cloud. The exchange energy is repulsive (4.19 kcal mol^{-1}), and has nearly twice the magnitude of the electrostatic energy. The induction energy is a product of the interaction between each monomer and the static electric field of the other; here it contributes a modest attractive component (-0.80 kcal mol^{-1}) to the binding energy. The dispersion energy is by far the largest attractive component (-4.16 kcal mol^{-1}), with nearly twice the magnitude of the electrostatic energy. It is interesting that the magnitude of the dispersion energy is nearly equivalent to the exchange energy, which roughly holds for substituted benzene dimer systems also.[51]

We also performed an SAPT decomposition at the inverted, sulfur-down geometry, $A1 = 180°$. In this geometry, instead of the electron-deficient hydrogen atoms, the sulfur lone pairs are directed toward the benzene ring. As one might expect, this causes the electrostatic component of the interaction to decrease and even become slightly repulsive. The other three energy components also decrease in magnitude because the electron density from the sulfur lone pairs does not extend as far from the sulfur as the electron density associated with the hydrogens in H_2S. This might be anticipated from simple VSEPR considerations, which would suggest that the very small H-S-H bond angle of 92° would imply a large angle between the sulfur lone pairs. We note that the exchange-repulsion is reduced in magnitude much more than the dispersion interaction, so that the sum of exchange-repulsion and dispersion is now somewhat attractive (-1.11 kcal mol^{-1}) rather than

almost zero as in the hydrogens-down $A1 = 0°$ configuration. However, the reduction in the electrostatic term outweighs this effect, so that overall, the sulfur-down configuration is 1.88 kcal mol^{-1} less favorable than the hydrogens-down configuration at the SAPT2/aug-cc-pVDZ level of theory [1.54 kcal mol^{-1} less favorable for CCSD(T)/aug-cc-pVDZ]. Based on these considerations, the "S–π" interaction, at least in this model system, is best thought of as being primarily an electrostatic attraction between the H_2S hydrogens and the aromatic π–cloud.

4.4 Conclusions

In this study, we examined the H_2S–benzene dimer as the simplest model of S–π interactions. Calculations using several basis sets and different levels of electron correlation were performed to obtain potential energy curves for the intermonomer geometric parameters $A1$, $A2$, and R. Estimates of the CCSD(T)/aug-cc-pVQZ potential energy curves presented here for the C_{2v} configuration represent a great leap forward in the reliability of theoretical data for this system, and they should be suitable as benchmarks for the calibration of new theoretical methods for noncovalent interactions. The results at our highest levels of theory, $A1 = 30°$ for CCSD(T)/aug-cc-pVDZ, $R_{\mathrm{eq}} = 3.80$ Å and $E_{\mathrm{int}} = -2.74$ kcal mol^{-1} for CCSD(T)/aug-cc-pVQZ, are in good agreement with previous experimental and lower-level theoretical results. Complete basis set extrapolations yield a CCSD(T) interaction energy of -2.81 kcal mol^{-1}, which is very similar to our aug-cc-pVQZ result and suggests that errors due to basis set incompleteness are very small.

Analysis of the interaction using symmetry-adapted perturbation theory, together with the potential energy curve for rotation of the H_2S unit relative to the

benzene ring, suggests that the S–π interaction here is primarily an electrostatic attraction between the partial positive hydrogens in H_2S and the negatively-charged π electrons of benzene.

Comparison of different theoretical treatments showed that MP2 overbinds and CCSD underbinds with respect to CCSD(T), in accord with studies on other van der Waals systems. The extra (1d/1p) diffuse functions present in the aug-cc-pVDZ basis set improve the overall quality of results obtained with that basis over those obtained with the otherwise comparable 6-31++G** basis set by a significant amount. The extra functions in the aug-cc-pVTZ basis produce a similar but smaller effect compared to the 6-311+G(2d,p) basis. It is therefore recommended that the more complete aug-cc-pVXZ basis sets be employed when possible in future computational studies of this and similar van der Waals systems.

CHAPTER V

ANALYSIS OF S–π CONTACTS IN PROTEIN STRUCTURES AND COMPARISON TO THEORETICAL PREDICTIONS

5.1 Introduction

In the previous chapter, we studied the H_2S–benzene dimer in a hydrogen-bonded geometry with high-level computational methods in order to ascertain the energetics of this interaction. Because we used such high-level methods as well as large basis sets, any residual error from basis set incompleteness or from unrecovered correlation energy should be very small, and our results should match very well with gas-phase experimental data. We compared our results to those of Arunan et al.,[46] who used microwave spectrometry to study the geometry of gas-phase H_2S–benzene clusters. Our results matched theirs very closely — as close as possible given the resolution of our potential energy surfaces.

The long-range goal of studying the H_2S–benzene system was to use it as a prototype for modeling S–π interactions in proteins. Unfortunately, there are many environmental effects introduced in shifting our focus from the simple model system to a pair of amino acids within a protein. Obviously, there are many more atoms involved in the system, and each of these can contribute to a number of environmental effects. While it would be possible to extend the theoretical side

of our previous comparison to perform theoretical calculations on systems that are more representative of real proteins, the high computational cost would limit the study to lower levels of theory and smaller basis sets, which would also limit the accuracy of the results. Instead, what we can do is to compare the theoretical results on the model system to experimental results on real systems, i.e. protein crystal structures.

Two large and well-known repositories of these are the Brookhaven Protein DataBank (PDB)[34] and the Cambridge Crystallographic Database (CCD);[35] the "data-mining" of these collections has nearly become a scientific field in itself. In the previous chapter, we mentioned two studies which performed this data-mining in relation to the S–π interaction: Zauhar et al. extracted data on Met–Phe contacts from the CCD,[7] while Duan et al. used the PDB to compile data on Cys–Phe contacts.[8] Zauhar reported that Met–Phe contacts strongly prefer a geometry where the sulfur is in the plane of the phenyl ring, with the sulfur lone pairs pointed towards the ring. Duan reported the same result for Cys–Phe contacts, which seemed odd considering that this configuration has a much less favorable interaction energy than the H-bonded geometry.

In this study, we extract data on *all* possible S–π contacts (including both Cys and Met residues for sulfur, as well as Phe, Tyr, and Trp for aromatic rings) from a large dataset of PDB protein structures. We then analyze these data in light of theoretical calculations on three geometries of the H_2S–benzene dimer: the optimal H-bond geometry (reported in the previous chapter), an "inverted" geometry, and an "in-plane" geometry (both reported here).

5.2 Methods

The dataset was constructed by culling all structures from the PDB which contained protein only and were resolved to 1.5 Å or less; this produced a set of 946 high-resolution protein structures for analysis.

A custom Perl script was written to process these files and return data about sulfur-π contacts within each protein, as per the following steps:

1. For each file, the coordinates of all sulfur atoms and ring carbons (as well as the ring nitrogen for tryptophan) are pulled directly from the ATOM info and stored in hashes sorted by residue type. Cysteine sulfurs participating in a disulfide bond (as designated on the SSBOND lines) are sorted into a separate hash named CDI. The five-membered and six-membered rings of each tryptophan are treated separately as TRP5 and TRP6 for the calculations. The master hash thus contain three types of sulfurs (Met, Cys, and cystine) and four types of rings (Phe, Tyr, and both the five and six rings of Trp).

2. For each ring, the geometric center is computed from the coordinates of its ring atoms. The program then loops over all possible sulfur-ring combinations and calculates the sulfur-to-ring-center distance R for each pair.

3. For each ring, the vector normal to the ring plane is computed. The program again loops over all sulfur-ring combinations, computes the S–π vector, and then computes the angle (θ) between the normal vector and the S–π vector for each combination. The normal vector of the ring is used as the reference for $\theta = 0°$; see Figure 5.1.

4. To prevent double-counting of tryptophan contacts, the set of TRP5 contacts

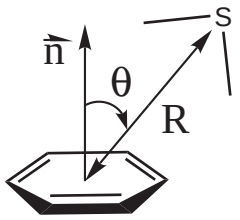


Figure 5.1: Definition of R, θ, and $\vec{n}$ for the H_2S–benzene Dimer

is checked against the set of TRP6 contacts; if contacts with the same sulfur occur in both sets, the one with the longer distance is deleted.

5. For each contact that has a distance less than a certain threshold (10 Å in this study), the source protein ID, sulfur residue ID, ring residue ID, distance, and angle are output to a text file for further processing in a spreadsheet program.

6. Total occurrences of each residue are counted over the entire dataset and output at the end of the operation.

(All of the actual calculations performed within this program are encapsulated within subroutines, making it a simple matter to modify the main code to obtain data on any other contact of interest. The overall source code is reproduced in Appendix A.)

The S–π contact data returned from the processing of these files was then analyzed in Microsoft Excel. All contacts were counted into an array of "bins", each of which was defined by a 0.5 Å increment of R and a 10° increment of θ. The spatial volume defined by each of these bins is different - the size of the spherical wedge defined by each increment of θ increases as θ increases; also, the size of the spherical shell defined by each increment of R increases as R increases. Because of this, the final counts of all bins were divided by the spatial volume of that bin (in Å^3) in order to produce an evenly-scaled dataset of relative frequencies. [Neither of

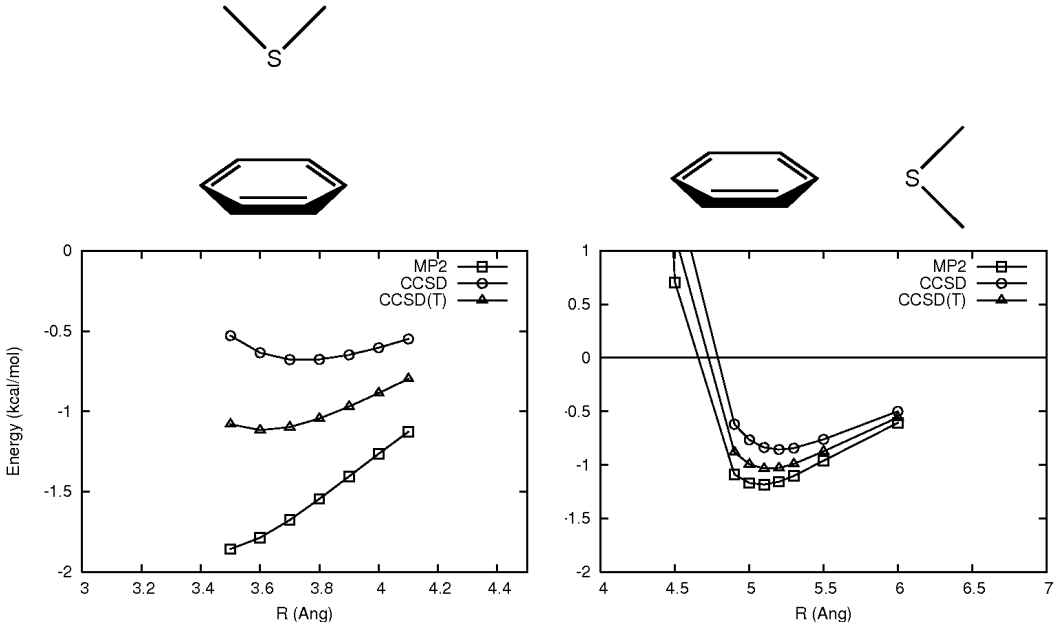
the previously mentioned data-mining studies took this into account, but we believe it to be the proper way to interpret this data.]

3-D histograms plotting R vs θ vs Relative Frequency were then produced for the full dataset and various subsets thereof, such as Cys sulfurs only, Phe rings only, or Met-Tyr contacts only. No contacts were found below $R = 3.0$ Å, and beyond $R = 8.0$ Å the frequency of contacts simply increased steadily as a function of increasing volume; thus we used these two points as our limits on R.

Additional PECs were computed corresponding to the inverted and in-plane geometries of the H_2S–benzene dimer as shown in Figure 5.2. MP2, CCSD, and CCSD(T) levels of theory were used with the aug-cc-pVTZ basis set. Our previous study showed that using the aug-cc-pVQZ basis yielded only a minimal improvement over aug-cc-pVTZ, and we felt that this small increase in accuracy did not justify the use of the larger basis in this case. All calculations are counterpoise-corrected and use the same rigid monomers defined in the previous chapter.

5.3 Results and Discussion

The results of our additional theoretical calculations are shown in Figure 5.2. The predicted equilibrium for the inverted geometry lies at $R_{eq} = \sim 3.6$ Å and $E_{int} = -1.12$ kcal mol^{-1}. This R is very close to our previously predicted R_{eq} for the H-bonded geometry of 3.80 kcal mol^{-1}, which is somewhat unfortunate for our PDB analysis. Since most protein crystal structures have a resolution much larger than 0.2 Å (the smallest in our dataset is 0.5 Å), and since PDB structures generally do not include hydrogen data, it is difficult to differentiate between these two interaction geometries in our analysis. When this is the case, we will refer to these two geometries collectively as the *perpendicular* configurations.



(a) Inverted geometry (b) In-plane geometry

Figure 5.2: Structure and PES for inverted and in-plane geometry of H_2S–benzene, CCSD(T)/aug-cc-pVTZ level of theory.

The predicted equilibrium for the in-plane geometry is at $R_{eq} = 5.1$ Å and $E_{int} = -1.03$ kcal mol^{-1}. This E_{int} is less than half of that predicted for the H-bond (-2.81 kcal mol^{-1}), a result that agrees quite nicely with the results of Duan et al.[8] Thus, in the absence of other considerations, we would expect to see about a 2:1 ratio between the optimum and in-plane geometries for contacts involving cysteines. On the other hand, we expect those involving methionine to be evenly distributed between the two configurations, as the E_{int} for the inverted and in-plane geometries are very similar.

These expectations did not hold in light of our PDB analysis, however. The overall histogram is shown in Figure 5.3 (see section 5.5, where one can see that the H-bond geometry is actually slightly less favored than the in-plane geometry. A similar trend can be seen when the results were broken down by the type of

sulfur residue: Met (Figure 5.4) and Cys (Figure 5.5) contacts both show a the same slight preference for the in-plane geometry. When divided by the type of ring residue, however, the trend can only be seen for Phe (Figure 5.6) and Tyr (5.7) residues; Trp residues (Figure 5.8) show extremely few contacts at all for the in-plane geometry. This seems to indicate that there are indeed some environmental effects that are changing the energetics of this S–π interaction.

Exactly what these environmental effects are, however, is not readily apparent from our analysis. One possibility is that the sulfur atom is having to compete with other functional groups for the spot over the ring center, and when the other groups have a more favorable interaction, the sulfur is displaced to the secondary minimum at the in-plane geometry. Possible candidates for this H–π–bonding competition are –OH, –NH, and –CH groups. Studies by Tsuzuki et al. on prototype dimer (water–benzene, ammonia–benzene, and methane–benzene) have shown that these interactions can have E_{int} up to -3.17, -2.22, and -1.45 kcal mol^{-1}, respectively (all calculated with CCSD(T) at the CBS limit).[3,6] The –OH–π interaction is stronger than -SH–π, while the other two are weaker. Still, as we have said before, environmental effects could alter these interactions as well. One such effect is polarization of the H-bonding atom (C, N, O, or S) due to ionic or strongly polar groups nearby in the protein structure. Since sulfur has a much larger atomic polarizability than C, N, or O (see Table 5.1), it would be more easily polarized and its interaction with aromatic rings would be more greatly affected.

Looking back at the histograms, one can see that there is more going on than the interplay between the perpendicular and in-plane geometries. In particular, there is an interesting cluster of large peaks near $R = 5.0$ Å and $\theta = 0°$. Based on the PECs in our previous study, the S–π interaction in the H-bond geometry

Table 5.1: Atomic polarizabilities of H-bonding elements

Element	Polarizability (Å^3)
C	1.8
N	1.1
O	0.793
S	2.9

at this distance would have an E_{int} of only -1.1 kcal mol^{-1}, about the same as the optimum E_{int} for both the inverted and in-plane geometries. Looking at the separate Met and Cys histograms, one sees a much larger set of peaks in this area for Met residues than for Cys; in fact, the peaks for Met in this region are larger than any of those in the perpendicular or in-plane regions! All three ring-residue histograms show similar groupings of peaks in this region.

Because of this, we believe that this grouping of contacts indicates a –CH–π interaction involving the terminal ϵ–C on Met or the β–C on Cys. From Tsuzuki's work,[6] the R_{eq} for methane–benzene is 3.8 Å, and the average length of a C–S bond from PDB data is 1.3 Å, making the total sulfur-to-ring-center distance 5.1 Å (assuming that the C–S bond is directly collinear with the ring center). Recent work in our group has shown that the methane–phenol and methane–indole dimers both show similar E_{int} and R_{eq} as the Tsuzuki data, with the indole interaction being somewhat stronger (by about 0.5 kcal mol^{-1}).[58]

In Table 5.2 we give the number of contacts found in each interesting region, along with the percentage of the total contacts for that residue that this number represents. From this we can make inferences about the relative probabilities of the different residues to enter into the different kinds of contacts. First, we see that while Met and Cys are equally as likely to enter into either perpendicular or in-plane contacts, Met residues are twice as likely to form alkyl–π contacts. This

Table 5.2: Numerical analysis of S–π contacts in PDB based on geometry and type of residue

	Cys	Met	Phe	Tyr	Trp	Total
Residues	3869	3964	8676	7735	3283	7833
Contacts	5577	5405	4397	3476	3109	10982
Perpendicular contacts[a]	64	80	75	28	41	144
Percent	1.15	1.48	1.70	0.80	1.32	1.31
Alkyl–π contacts[b]	264	456	286	187	247	720
Percent	4.73	8.44	6.50	5.38	7.94	6.56
In-plane contacts[c]	430	407	279	176	84	837
Percent	7.71	7.53	6.34	5.06	2.70	7.62

[a] Perpendicular contacts include both the H-bond and inverted geometries and are defined as $\{R \leq 4.0$ Å$\}$ and $\{A \leq 30°\}$
[b] Alkyl–π contacts are defined as $\{4.5$ Å $\leq R \leq 6.0$ Å$\}$ and $\{A \leq 30°\}$
[c] In-plane contacts are defined as $\{4.5$ Å $\leq R \leq 6.0$ Å$\}$ and $\{80° \leq A \leq 90°\}$

make sense since the ϵ–C of Met has a much higher degree of freedom than the β–C of Cys and can more easily enter into a close contact with an aromatic ring.

As for the different ring residues, we see that in all cases, Phe and Tyr are about equally likely to enter into the different contact geometries, with Tyr having slightly lower percentages (although this may not be statistically significant). Trp residues are the least likely to have in-plane contacts by a large margin, but they are the most likely to have alkyl–π contacts. The aversion to in-plane contacts is not well understood, although the stronger interaction mentioned for methane–indole above is probably the cause of the preference for alkyl–π contacts.

5.4 Conclusions

We have shown that in proteins, contrary to theoretical expectations, S–π contacts show approximately equal preferences for the perpendicular and in-plane

geometries, while also showing a strong preference for configurations that put the Met ϵ–C and the Cys β–C into a alkyl–π type contact. These preferences are exhibited by all residues except Trp, which shows an unusual aversion for in-plane contacts.

We proposed an explanation for the discrepancies between this analysis and our theoretical predictions for the H_2S–benzene dimer, namely, that other functional groups are competing with the sulfur for the H-bonding position. While this idea is quite reasonable in light of our data, it should be explored further in order to determine which other groups are most likely to displace the sulfur residues from the H-bond position. A double-contact search could be performed, where in-plane S–π contacts in which the ring was *also* in contact with a different (-ϕ, -CH, -NH, or -OH) side-chain would be sought out. Studying the environment around these contacts could then provide insight into how environmental effects alter the energetics of the different interactions.

5.5 Histograms of S–π Contacts

The data for all histograms in this section are scaled based on the actual volume (in Å^3) of the space defined by the R and θ bins for each sector. R is measured from the ring center to the sulfur atom; θ is measured from the normal vector of the ring. Each label n on the R axis designates a bin containing data for $\{n - 0.5 < R \leq n\}$; likewise, each label n on the θ axis designates a bin for $\{n - 10 < \theta \leq n\}$.

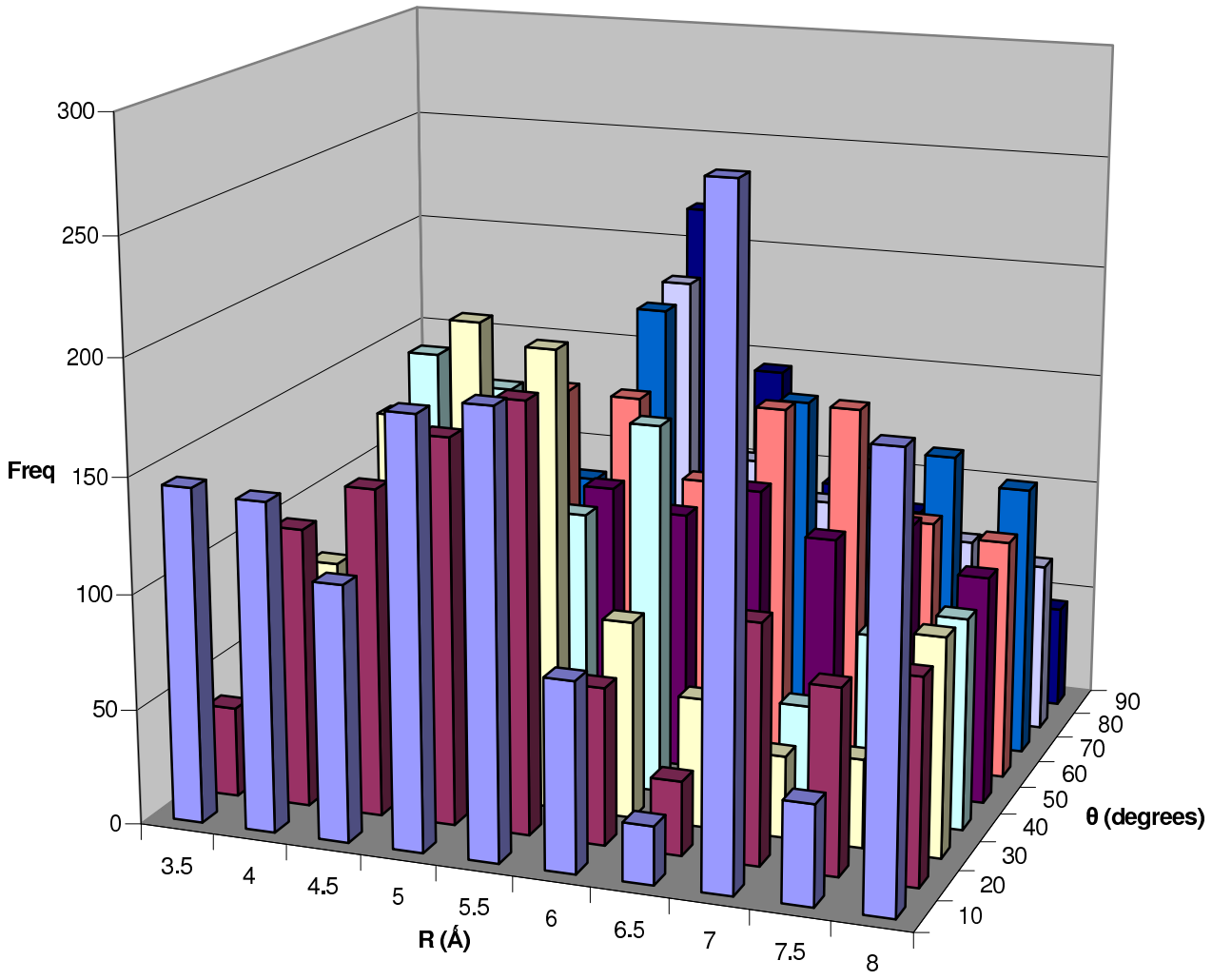


Figure 5.3: Overall S–π contacts.

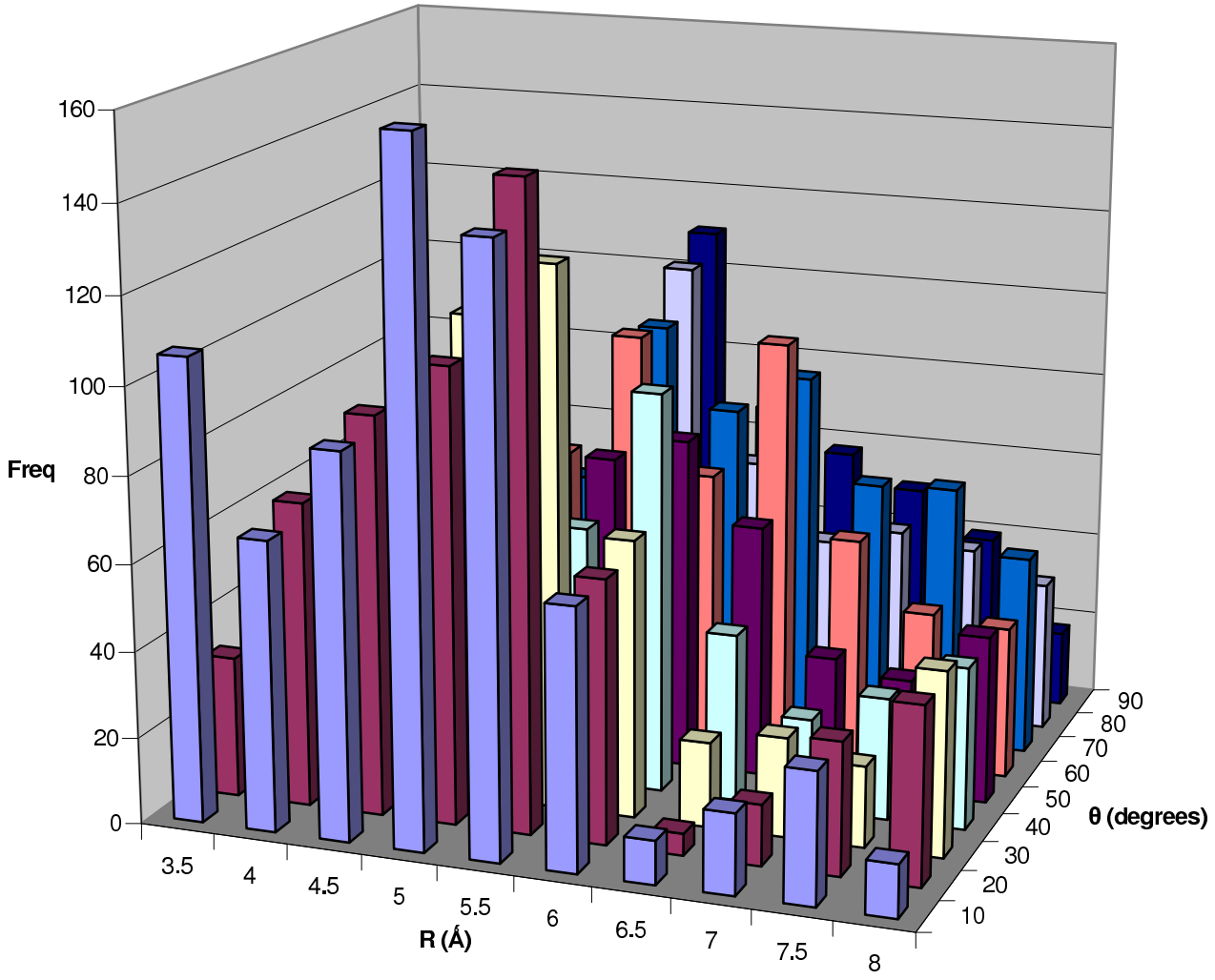


Figure 5.4: All S–π contacts involving methionine.

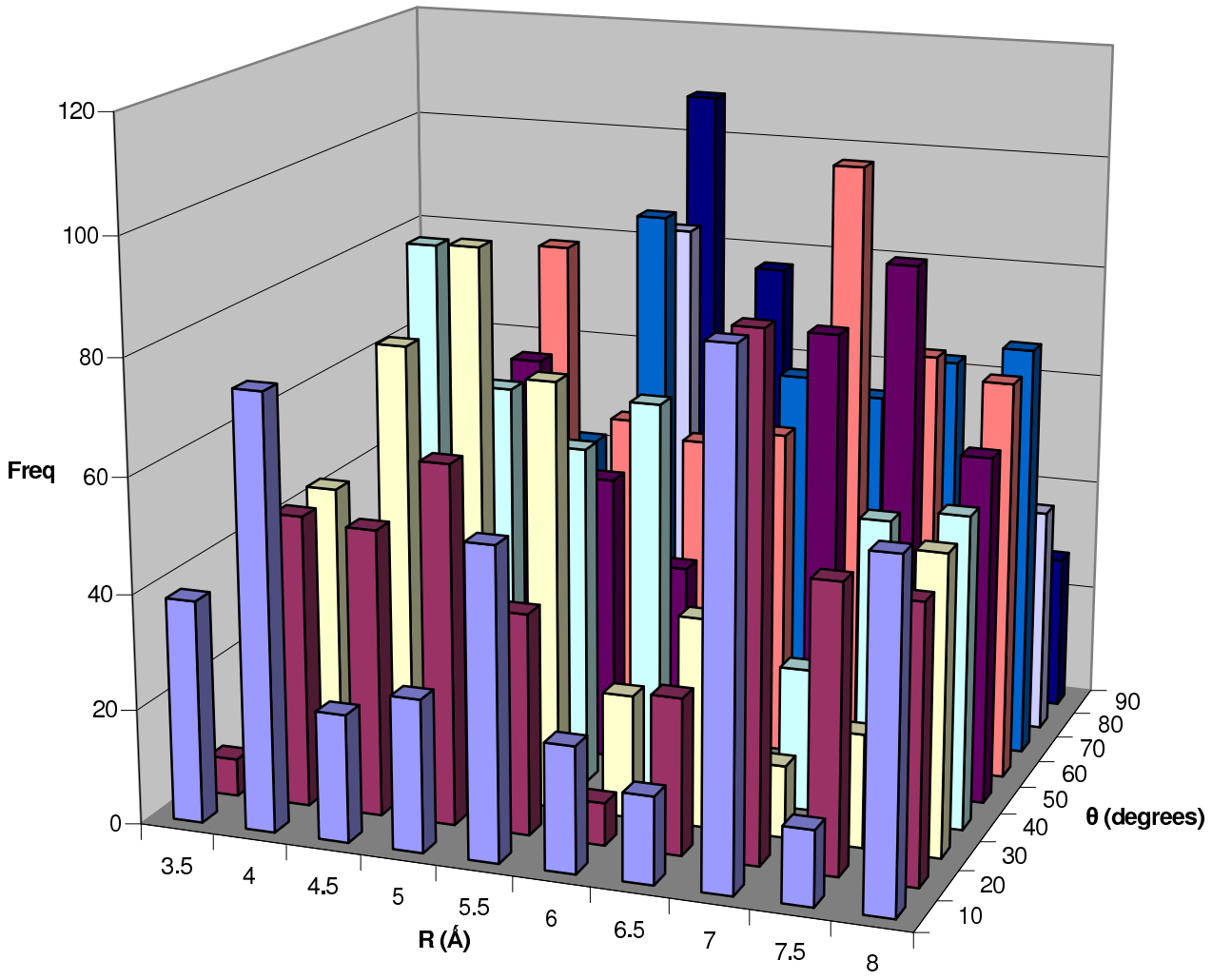


Figure 5.5: All S–π contacts involving cysteine.

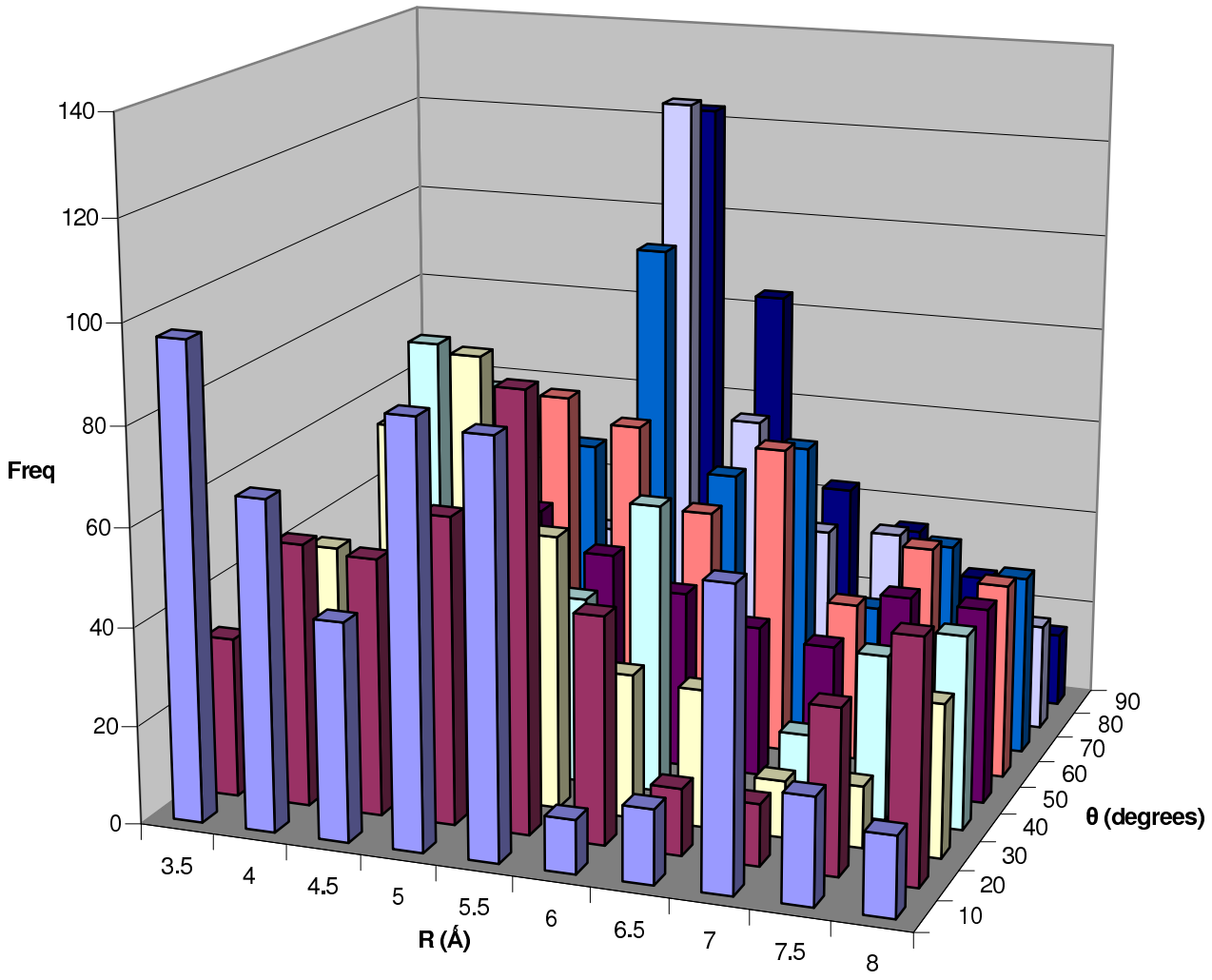


Figure 5.6: All S–π contacts involving phenylalanine.

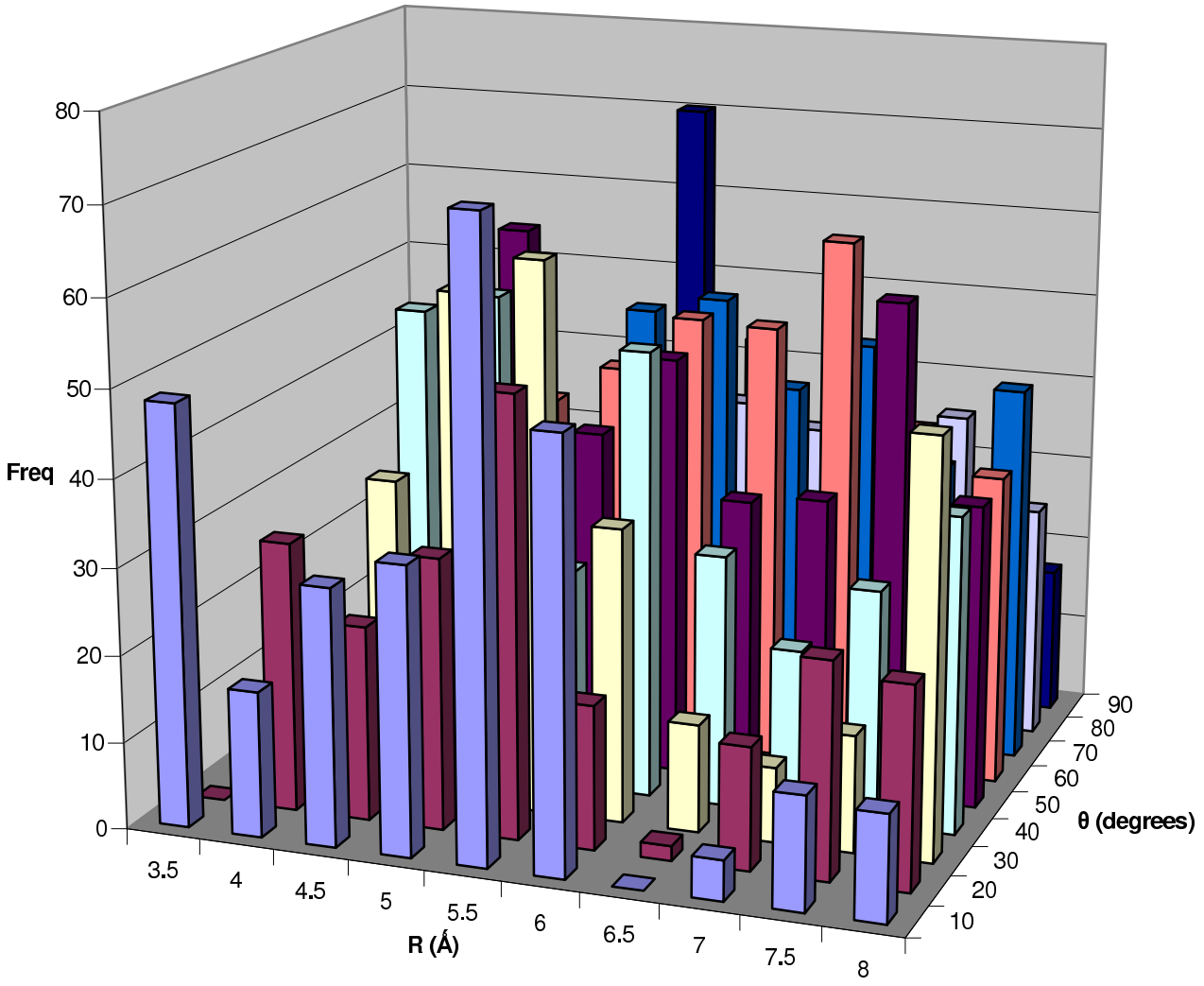


Figure 5.7: All S–π contacts involving tyrosine.

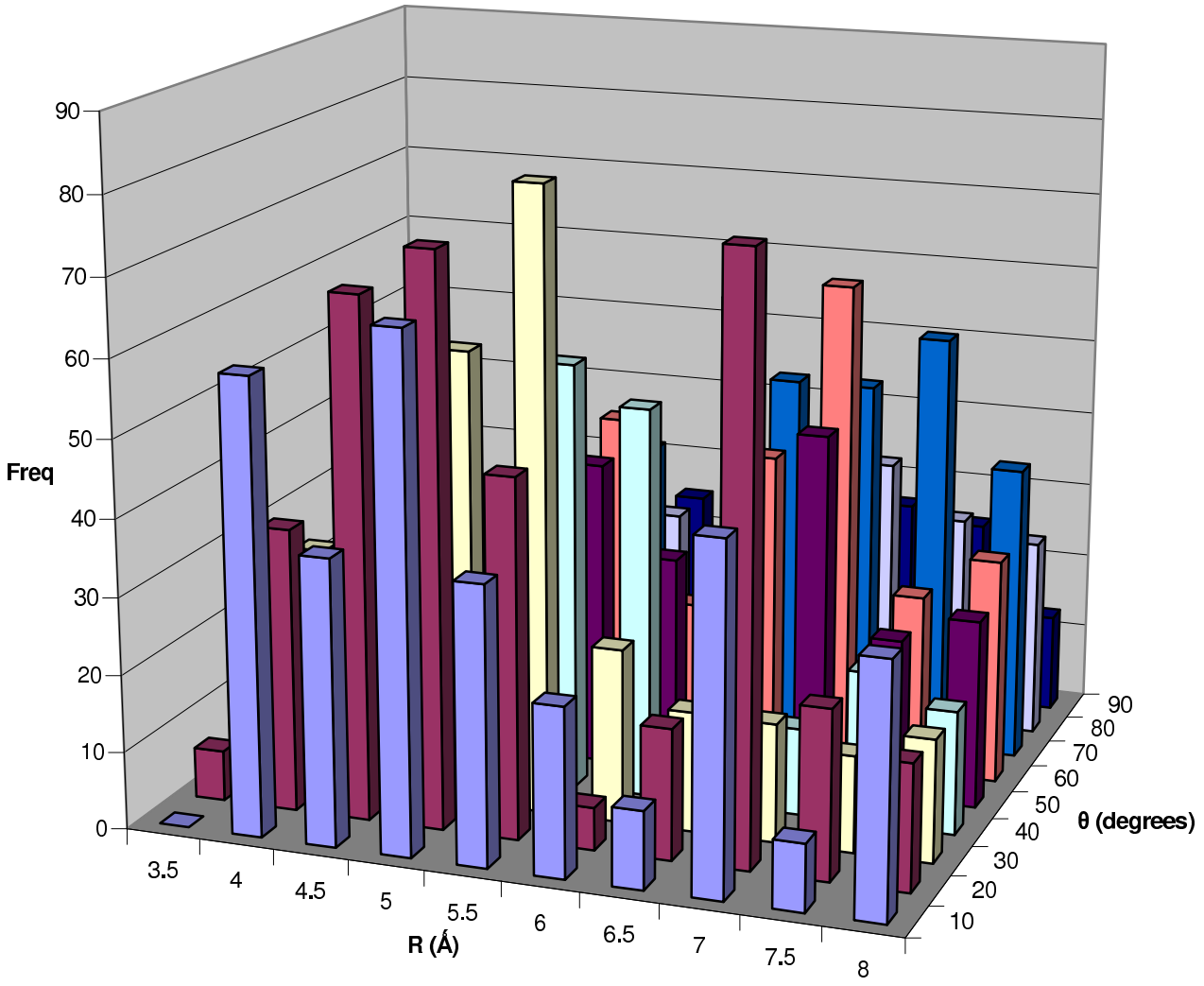


Figure 5.8: All S–π contacts involving tryptophan.

APPENDIX A

SOURCE CODE FOR PDB ANALYSIS PROGRAM

```
#!/usr/bin/perl -w

use strict;
use warnings;
use Benchmark;
#use BeginPerlBioinfo;
use Math::Trig;

my $t0 = new Benchmark;


#####
# Main
#####

# Threshold for determining "interesting" S-pi contacts, in
# angstroms
my $dist_thr = 10;

# Flag for differentiating cystines from cysteines
my $cystineflag = 1;

my $filecount = 0;
my $folder = "data";
my @files = ();

opendir (DIR, $folder) or die "Can't open folder: $!\n";
@files = readdir(DIR);
closedir (DIR);

open (OFP, ">output.txt") or die "Can't open output file:
$!\n";
```

```
  # Variables to count the total occurences of each residue
  my $cyscount = 0;
  my $cdicount = 0;
  my $metcount = 0;
  my $phecount = 0;
  my $tyrcount = 0;
  my $trp5count = 0;
41 my $trp6count = 0;

  foreach my $input (sort @files) {

      unless ($input =~ /.ent$/) { next; };
      my ($protein, $ext) = split /\./, $input;

      $input = $folder."/".$input unless $folder eq '.';
      $filecount++;
      print "Now processing file $filecount: $input...\n";
51
      #####
      # Variables
      #####

      # Hash to hold all atomic data parsed from .ent file
      my %biglist = (
         'ALA' => [],
         'VAL' => [],
         'LEU' => [],
61       'ILE' => [],
         'PRO' => [],
         'TRP' => [],
         'PHE' => [],
         'MET' => [],
         'GLY' => [],
         'SER' => [],
         'THR' => [],
         'TYR' => [],
         'CYS' => [],
71       'CDI' => [],
         'ASN' => [],
         'GLN' => [],
         'LYS' => [],
         'ARG' => [],
         'HIS' => [],
         'ASP' => [],
         'GLU' => [],
```

```
   );

# Inidividual arrays to hold coordinates of sulfurs and
# ring centers
my @cyssulfurs = ();
my @cdisulfurs = ();
my @metsulfurs = ();
my @phecenters = ();
my @tyrcenters = ();
my @trp5centers = ();
my @trp6centers = ();

# Hash to store S-pi distances
my %spidist = ( cys => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    cdi => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    met => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    );

# Hash to store matrices of vectors from ring centers to
# sulfurs
my %spivecs = ( cys => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    cdi => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    met => { phe => [],
        tyr => [],
        trp5 => [],
```

```
        trp6 => [],
    },
     );

# Hash to store arrays of normal vectors of pi-rings
my %pinorms = ( phe => [],
     tyr => [],
     trp5 => [],
     trp6 => [],
     );

# Hash to store reference vectors of pi rings
my %pirefs = ( phe => [],
    tyr => [],
    trp5 => [],
    trp6 => [],
    );

# Hash to store projections of S-pi vector onto pi-normal
my %spiproj = ( cys => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
    },
     cdi => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
    },

     met => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
    },
     );

# Hash for easy interconversion from 3-letter-code to
# data array
my %spiarrays = ( cys => \@cyssulfurs,
      cdi => \@cdisulfurs,
      met => \@metsulfurs,
      phe => \@phecenters,
      tyr => \@tyrcenters,
      trp5 => \@trp5centers,
```

```
        trp6 => \@trp6centers,
        );

# Hash for conversion of 3-letter code to full name
my %spinames = ( cys => "Cysteine",
     cdi => "Cystine",
     met => "Methionine",
     phe => "Phenylalanine",
     tyr => "Tyrosine",
     trp5 => "Tryptophan(5)",
     trp6 => "Tryptophan(6)",
     );

# Hash to store S-pi elevation angles
my %spiangs = ( cys => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    cdi => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    met => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    );

# Arrays to store S-pi orientational angles
my %spiphis = ( cys => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    cdi => { phe => [],
        tyr => [],
        trp5 => [],
        trp6 => [],
   },
    met => { phe => [],
        tyr => [],
        trp5 => [],
```

```
            trp6 => [],
        },
         );

  # Build hash of all atom data keyed by residue type
  get_res_list($input, %biglist);

  # Extract sulfur data from cysteines and methionines in
  # biglist
  foreach my $cysref (@{$biglist{'CYS'}}) {
if (defined(@$cysref[5]) && $$cysref[5][0] =~ 'SG') {
    push @cyssulfurs, @$cysref[5];
}
else { push @cyssulfurs, 'dummy' }
  }
  $cyscount += scalar(@cyssulfurs);

  foreach my $cdiref (@{$biglist{'CDI'}}) {
if (defined(@$cdiref[5]) && $$cdiref[5][0] =~ 'SG') {
    push @cdisulfurs, @$cdiref[5];
}
else { push @cyssulfurs, 'dummy' }
  }
  $cdicount += scalar(@cdisulfurs);

  foreach my $metref (@{$biglist{'MET'}}) {
if (defined(@$metref[6]) && $$metref[6][0] =~ 'SD') {
    push @metsulfurs, @$metref[6];
}
else { push @metsulfurs, 'dummy' }
  }
  $metcount += scalar(@metsulfurs);

  # Calculate ring center data for PHE, TRP, and TYR5&6
  foreach my $pheref (@{$biglist{'PHE'}}) {
if (defined(@$pheref[5]) && defined(@$pheref[6]) &&
    defined(@$pheref[7]) && defined(@$pheref[8]) &&
    defined(@$pheref[9]) && defined(@$pheref[10]) &&
    $$pheref[5][0] =~ 'CG' && $$pheref[6][0] =~ 'CD1' &&
    $$pheref[7][0] =~ 'CD2' && $$pheref[8][0] =~ 'CE1' &&
    $$pheref[9][0] =~ 'CE2' && $$pheref[10][0] =~ 'CZ')
{
    push @phecenters, [ get_center($$pheref[5],
           $$pheref[6],
           $$pheref[7],
```

```
            $$pheref[8],
            $$pheref[9],
            $$pheref[10]) ];
} else { push @phecenters, 'dummy' }
  }
  $phecount += scalar(@phecenters);

  foreach my $tyrref (@{$biglist{'TYR'}}) {
if (defined(@$tyrref[5]) && defined(@$tyrref[6]) && defined
   (@$tyrref[7]) &&
    defined(@$tyrref[8]) && defined(@$tyrref[9]) && defined
       (@$tyrref[10]) &&
    $$tyrref[5][0] =~ 'CG' && $$tyrref[6][0] =~ 'CD1' &&
       $$tyrref[7][0] =~ 'CD2' &&
    $$tyrref[8][0] =~ 'CE1' && $$tyrref[9][0] =~ 'CE2' &&
       $$tyrref[10][0] =~ 'CZ' ) {
    push @tyrcenters, [ get_center($$tyrref[5], $$tyrref
       [6], $$tyrref[7], $$tyrref[8], $$tyrref[9], $$tyrref
       [10]) ];
} else { push @tyrcenters, 'dummy' }
  }
  $tyrcount += scalar(@tyrcenters);

  foreach my $trpref (@{$biglist{'TRP'}}) {
if (defined(@$trpref[5]) && defined(@$trpref[6]) && defined
   (@$trpref[7]) &&
    defined(@$trpref[8]) && defined(@$trpref[9]) &&
    $$trpref[5][0] =~ 'CG' && $$trpref[6][0] =~ 'CD1' &&
       $$trpref[7][0] =~ 'CD2' &&
    $$trpref[8][0] =~ 'NE1' && $$trpref[9][0] =~ 'CE2'
    ) {
    push @trp5centers, [ get_center($$trpref[5], $$trpref
       [6], $$trpref[7], $$trpref[8], $$trpref[9]) ]
    } else { push @trp5centers, 'dummy' }

if (defined(@$trpref[7]) && defined(@$trpref[9]) && defined
   (@$trpref[10]) &&
    defined(@$trpref[11]) && defined(@$trpref[12]) &&
       defined(@$trpref[13]) &&
    $$trpref[7][0] =~ 'CD2' && $$trpref[9][0] =~ 'CE2' &&
       $$trpref[10][0] =~ 'CE3' &&
    $$trpref[11][0] =~ 'CZ2' && $$trpref[12][0] =~ 'CZ3' &&
        $$trpref[13][0] =~ 'CH2'
    ) {
```

```
      push @trp6centers, [ get_center($$trpref[7], $$trpref
         [9], $$trpref[10], $$trpref[11], $$trpref[12],
         $$trpref[13]) ]
291   } else { push @trp6centers, 'dummy' }
    }
    $trp5count += scalar(@trp5centers);
    $trp6count += scalar(@trp6centers);

    # Calculate distance matrices for each S-pi pair
    foreach my $sulfur (sort keys %spidist) {
  foreach my $ring (sort keys %{$spidist{$sulfur}}) {
      @{$spidist{$sulfur}{$ring}} = calc_dist_mat($spiarrays{
         $sulfur}, $spiarrays{$ring});
#       print_mat("$spinames{$sulfur}-$spinames{$ring}
 distances", $sulfur, substr($ring, 0, 3), @{$spidist{
 $sulfur}{$ring}});
301 }
    }

    # Prepare to calculate elevation angles for each pair
    # Calculate array of normal vectors of pi rings
    foreach my $ring (keys %{$spivecs{cys}}) {
  my $ring2 = uc(substr($ring, 0, 3));
  my $piarray = $spiarrays{$ring};
  for my $i (0..$#$piarray) {
      if ($$piarray[$i] eq 'dummy') {
311   $pinorms{$ring}[$i] = 'dummy';
    next;
      }
      if ($ring eq "phe" || $ring eq "tyr") {
    $pinorms{$ring}[$i] = [ normal_3points( $biglist{$ring2}[
       $i][6], $biglist{$ring2}[$i][8], $biglist{$ring2}[$i
       ][10] ) ];
      }
      elsif ($ring eq "trp5") {
    $pinorms{$ring}[$i] = [ normal_3points( $biglist{$ring2}[
       $i][5], $biglist{$ring2}[$i][7], $biglist{$ring2}[$i
       ][9] ) ];
      }
      elsif ($ring eq "trp6") {
321   $pinorms{$ring}[$i] = [ normal_3points( $biglist{$ring2}[
       $i][9], $biglist{$ring2}[$i][10], $biglist{$ring2}[$i
       ][13] ) ];
      }
  }
```

```
    }
    foreach my $sulfur (keys %spivecs) {
  foreach my $ring (keys %{$spivecs{$sulfur}}) {
      my $ring2 = uc(substr($ring, 0, 3));
      my $sarray = $spiarrays{$sulfur};
      my $piarray = $spiarrays{$ring};
      # Calculate matrix of vectors from center of rings to
        sulfurs
331   for my $i (0..$#$sarray) {
    for my $j (0..$#$piarray) {
      if ($$piarray[$j] eq 'dummy' || $$sarray[$i] eq '
        dummy') {
    $spivecs{$sulfur}{$ring}[$i][$j] = 'dummy';
    next;
      }
      my $x = $$sarray[$i][1] - $$piarray[$j][1];
      my $y = $$sarray[$i][2] - $$piarray[$j][2];
      my $z = $$sarray[$i][3] - $$piarray[$j][3];
      $spivecs{$sulfur}{$ring}[$i][$j] = [$x, $y, $z];
341   }
    }
  }
    }

    # Calculate elevation angles for each pair
    foreach my $sulfur (sort keys %spiangs) {
  foreach my $ring (sort keys %{$spiangs{$sulfur}}) {
      @{$spiangs{$sulfur}{$ring}} = calc_angle_mat($spivecs{
        $sulfur}{$ring}, $pinorms{$ring});
#     print_mat("$spinames{$sulfur}-$spinames{$ring}
 elevation angles", $sulfur, substr($ring, 0, 3), @{
 $spiangs{$sulfur}{$ring}});
351 }
    }

    # Prepare to calculate orientational angles for each pair
    foreach my $sulfur (keys %spivecs) {
  foreach my $ring (keys %{$spivecs{$sulfur}}) {
      my $ring2 = uc(substr($ring, 0, 3));
      my $sarray = $spiarrays{$sulfur};
      my $piarray = $spiarrays{$ring};
      for my $i (0..$#$sarray) {
361   for my $j (0..$#$piarray) {
      if ($$sarray[$i] eq 'dummy' || $$piarray[$j] eq '
        dummy') {
```

```
        $spiproj{$sulfur}{$ring}[$i][$j] = 'dummy';
        next;
          }
          my ($nx, $ny, $nz) = @{$pinorms{$ring}[$j]};
          my ($sx, $sy, $sz) = @{$spivecs{$sulfur}{$ring}[$i][
             $j]};
          my $theta = deg2rad($spiangs{$sulfur}{$ring}[$i][$j])
             ;

          # Calculate projection of S-pi vector into pi-plane
             by difference from projection onto pi-normal
371       my $slen = sqrt($sx*$sx + $sy *$sy + $sz*$sz);
          my $plen = $slen * cos($theta);
          my ($px, $py, $pz) = ($nx*$plen, $ny*$plen, $nz*$plen
             );
          $spiproj{$sulfur}{$ring}[$i][$j] = [$sx-$px, $sy-$py,
              $sz-$pz];
      }
        }
        # Calculate vector defining 'zero' angle for ring,
           referenced to backbone carbon
        for my $i (0..$#$piarray) {
      if ($$piarray[$i] eq 'dummy') {
          $pirefs{$ring}[$i] = 'dummy';
381       next;
      }
      my ($cx, $cy, $cz);
      if ($ring ne 'trp6') {
          ($cx, $cy, $cz) = ($biglist{$ring2}[$i][5][1],
             $biglist{$ring2}[$i][5][2], $biglist{$ring2}[$i
             ][5][3]);
      }
      else {
          ($cx, $cy, $cz) = ($biglist{$ring2}[$i][7][1],
             $biglist{$ring2}[$i][7][2], $biglist{$ring2}[$i
             ][7][3]);
      }
      $pirefs{$ring}[$i] = [ $cx-$$piarray[$i][1], $cy-
         $$piarray[$i][2], $cz-$$piarray[$i][3] ];
391     }
    }
      }

      # Calculate orientational angles for each pair
      foreach my $sulfur (sort keys %spiphis) {
```

```
    foreach my $ring (sort keys %{$spiphis{$sulfur}}) {
        @{$spiphis{$sulfur}{$ring}} = calc_angle_mat($spiproj{
           $sulfur}{$ring}, $pirefs{$ring});
#       print_mat("$spinames{$sulfur}-$spinames{$ring}
     orientational angles", $sulfur, substr($ring, 0, 3), @{
     $spiphis{$sulfur}{$ring}});
    }
401   }

      # Extract data on "interesting" contacts as determined by
          $dist_thr
      foreach my $sulfur (sort keys %spidist) {
    foreach my $ring (sort keys %{$spidist{$sulfur}}) {
        my $ofp2 = $sulfur.$ring.".txt";
        open (OFP2, ">>$ofp2")  or die "Can't open output file
           $ofp2: $!\n";
        for my $i (0..$#{$spidist{$sulfur}{$ring}}) {
      for my $j (0..$#{$spidist{$sulfur}{$ring}[$i]}) {
          if ($spidist{$sulfur}{$ring}[$i][$j] ne 'dummy' &&
             $spidist{$sulfur}{$ring}[$i][$j] < $dist_thr) {
411     printf OFP "%s  %-5s  %-5s  %5.4f  %7.3f  %7.3f\n",
           $protein, $sulfur.$i, $ring.$j, $spidist{$sulfur}{
           $ring}[$i][$j],
        $spiangs{$sulfur}{$ring}[$i][$j], $spiphis{$sulfur}{
           $ring}[$i][$j];
        printf OFP2 "%s  %-5s  %-5s  %5.4f  %7.3f  %7.3f\n",
           $protein, $sulfur.$i, $ring.$j, $spidist{$sulfur}{
           $ring}[$i][$j],
        $spiangs{$sulfur}{$ring}[$i][$j], $spiphis{$sulfur}{
           $ring}[$i][$j];

          }
      }
        }
        close (OFP2);
    }
421   }

  # End foreach $input (@files)
  }

  open (OFP2, ">count.txt") or die "Can't open output file
     count.txt: $!\n";
```

```
print OFP2 "Cys  $cyscount\nCdi  $cdicount\nMet  $metcount\
   nPhe  $phecount\nTyr  $tyrcount\nTrp5 $trp5count\nTrp6
   $trp6count\n";
close (OFP2);

close (OFP);

my $t1 = new Benchmark;
my $td = timediff($t1, $t0);
print "$filecount files processed in ", timestr($td), "\n";

exit;



#####
# Subroutines
#####

# get_res_list($filename, %hash)
#
# given a PDB filename, extracts atomic names and coordinates
#  and returns them in a hash keyed on residue type
#
# hash is formatted: %hash{residue type}[res number][atom
   number][0-3]
# 0: atom name
# 1-3: xyz coordinates

sub get_res_list {

    my ($input, %reshash) = @_;
    my %cystines = ();
    my $lastresnum = 0;
    my $lastresname = '';
    my $lastchain = '';
    my @res = ();

    open (IFP, "$input") or die "Can't open $input for
       reading: $!\n";

    # Extract list of cystines from SSBOND lines if requested
    if ($cystineflag) {
  my $found = 0;
```

```
while (<IFP>) {
    # If we reach EOF, there is no SSBOND section; reset
       file pointer
    if (eof(IFP)) {
  seek (IFP, 0,0);
  last;
    }
    # Skip lines until we find the SSBOND section
    /^SSBOND/ or next unless $found;
    $found = 1;
    # Make sure we leave the loop after we finish the
       SSBOND section
    last unless /^SSBOND/;

    my ($chain1, $res1, $chain2, $res2) = ($_ =~
       /^.{15}(.{1})(.{5}).{8}(.{1})(.{5})../);

    if (!defined($cystines{$chain1})) {
  $cystines{$chain1} = ();
    }

    push @{$cystines{$chain1}}, $res1;
    push @{$cystines{$chain2}}, $res2;
}
# Sort each element of %cystines for quicker searching
   later
foreach my $chain (keys %cystines) {
    @{$cystines{$chain}} = sort { $a <=> $b } @{$cystines{
       $chain}};
}
  }

  while (<IFP>) {
# Skip everything that doesn't define atoms
/^ATOM/ or next;

# Extract relevant info from each ATOM line
my ($name, $resname, $thischain, $thisresnum, $x, $y, $z) =
    ($_ =~ /^.{13}(.{3}).{1}(.{3}).{1}(.{1})(.{4})
   .{4}(.{8})(.{8})(.{8})../);


# See if we've begun a new residue; if so, push current
   @res to %reslist and reset @res
if ($thisresnum != $lastresnum) {
```

```
      if ($cystineflag) {
    # Check residue against cystine list; change $lastresname
      as appropriate
    # Elements of %cystines have been sorted; need only
      compare to first value, then pop that value when
      matched.
    if (defined($cystines{$lastchain}[0]) && $lastresnum ==
      $cystines{$lastchain}[0]) {
       $lastresname = 'CDI';
       shift @{$cystines{$lastchain}};
    }
      }

      push @{$reshash{$lastresname}}, [@res];
      @res = ();
  }

  # Add atom to current residue
  push @res, [$name, $x, $y, $z];

  $lastchain = $thischain;
  $lastresnum = $thisresnum;
  $lastresname = $resname;
    }
    # Dont forget the last residue!
    if ($cystineflag) {
  if (defined($cystines{$lastchain}[0]) && $lastresnum ==
     $cystines{$lastchain}[0]) {
      $lastresname = 'CDI';
      shift @{$cystines{$lastchain}};
  }
    }
    push @{$reshash{$lastresname}}, [@res];

    close (IFP);

    return;
}

# get_center($atomref*)
#
# given a list of atoms, returns the geometric center
#
# atom is an array like that returned in get_res_list,
#  with elements 0-3 defined as:
```

```
# 0: atom name
# 1-3: xyz coordinates

sub get_center {

    my @points = @_;

    my $sumx = 0;
    my $sumy = 0;
    my $sumz = 0;
    my @center = ();

    for my $i (0..$#points) {
  $sumx += $points[$i][1];
  $sumy += $points[$i][2];
  $sumz += $points[$i][3];
    }

    $center[0] = "X".scalar(@points);
    $center[1] = $sumx / scalar(@points);
    $center[2] = $sumy / scalar(@points);
    $center[3] = $sumz / scalar(@points);

    return @center;
}

# calc_dist_mat($atomarrayref, $atomarrayref)
#
# given two atom arrays (or matrices), return a matrix of the
#  distances between each pair

sub calc_dist_mat {

    my ($aref, $bref) = @_;
    my ($i, $j);
    my @distances = ();

    for my $i (0..$#$aref) {
  if ($$aref[$i] eq 'dummy') {
      for my $j (0..$#$bref) {
  $distances[$i][$j] = 'dummy';
      }
      next;
  }
```

```
    my ($ax, $ay, $az) = ($$aref[$i][1], $$aref[$i][2], $$aref[
       $i][3]);
    for my $j (0..$#$bref) {
591     if ($$bref[$j] eq 'dummy') {
      $distances[$i][$j] = 'dummy';
      next;
      }
      my ($bx, $by, $bz) = ($$bref[$j][1], $$bref[$j][2],
         $$bref[$j][3]);
      my $delx = $ax - $bx;
      my $dely = $ay - $by;
      my $delz = $az - $bz;
      my $dist = sqrt($delx*$delx + $dely*$dely + $delz*$delz
         );
      $distances[$i][$j] = $dist;
601 }
    }

    return @distances;
}

# calc_angle_mat($vectorarrayref, $vectorarrayref)
#
# given two vector arrays (not atoms; element 0 != atom name)
   ,
#  returns a matrix of the angles between each pair
611
sub calc_angle_mat {

    my ($aref, $bref) = @_;
    my ($i, $j);
    my @angles = ();

    for $i (0..$#$bref) {
  if ($$bref[$i] eq 'dummy') {
      for my $j (0..$#$aref) {
621   $angles[$j][$i] = 'dummy';
      }
      next;
  }
  my ($bx, $by, $bz) = ($$bref[$i][0], $$bref[$i][1], $$bref[
     $i][2]);
  for $j (0..$#$aref) {
      if ($$aref[$j] eq 'dummy' || ( ref($$aref[$j]) eq '
         ARRAY' && $$aref[$j][$i] eq 'dummy')) {
```

```
      $angles[$j][$i] = 'dummy';
      next;
        }
631     my ($ax, $ay, $az) = ($$aref[$j][0], $$aref[$j][1],
           $$aref[$j][2]);
        if (ref($$aref[$j]) eq 'ARRAY') {
      $ax = $$aref[$j][$i][0];
      $ay = $$aref[$j][$i][1];
      $az = $$aref[$j][$i][2];
        }
        my $dotprod = $ax*$bx + $ay*$by + $az*$bz;
        my $alen = sqrt($ax*$ax + $ay*$ay + $az*$az);
        my $blen = sqrt($bx*$bx + $by*$by + $bz*$bz);
        my $theta = acos($dotprod/($alen*$blen));
641     $angles[$j][$i] = rad2deg($theta);
    }
      }

      return @angles;
}

# print_mat($title, $xlabel, $ylabel, @matrix);
#
# given a matrix (array of arrays), title, and x and y labels
   ,
651 #  prints out the labeled matrix to the global output file

sub print_mat {

      my ($title, $x, $y, @mat) = @_;
      my ($i, $j);

      print OFP "\n$title\n";
      for $i (0..$#mat) {
    print OFP "\t\t$x$i" unless $mat[$i][0] eq 'dummy';
661   }
      print OFP "\n";

      for $j (0..$#{$mat[0]}) {
    next if $mat[0][$j] eq 'dummy';
    print OFP "$y$j\t";
    for $i (0..$#mat) {
        printf OFP "%12.6f\t", $mat[$i][$j] unless $mat[$i][$j]
            eq 'dummy';
    }
```

```
    print OFP "\n";
671   }

}

# normal_3points($atomref, $atomref, $atomref)
#
# given three atoms, returns the normal vector of the plane
#  defined by those atoms

sub normal_3points {
681
    my ($a, $b, $c) = @_;

    my @ab = ( ($$b[1]-$$a[1]), ($$b[2]-$$a[2]), ($$b[3]-$$a
       [3]) );
    my @ac = ( ($$c[1]-$$a[1]), ($$c[2]-$$a[2]), ($$c[3]-$$a
       [3]) );

    my @normal = ( ($ab[1]*$ac[2] - $ac[1]*$ab[2]), -($ab[0]*
       $ac[2] - $ac[0]*$ab[2]), ($ab[0]*$ac[1] - $ac[0]*$ab
       [1]) );
    my $length = sqrt($normal[0]*$normal[0] + $normal[1]*
       $normal[1] + $normal[2]*$normal[2]);
    $normal[0] /= $length;
    $normal[1] /= $length;
691 $normal[2] /= $length;

    return @normal;
}
```

BIBLIOGRAPHY

[1] Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Am. Chem. Soc.*, **1994**, *116*, 3500.

[2] Jaffe, R. L.; Smith, G. D. *J. Chem. Phys.*, **1996**, *105*, 2780.

[3] Tsuzuki, S.; Uchimaru, T.; Matsumura, K.; Mikami, M.; Tanabe, K. *Chem. Phys. Lett.*, **2000**, *319*, 547.

[4] Sinnokrot, M. O.; Valeev, E. F.; Sherrill, C. D. *J. Am. Chem. Soc.*, **2002**, *124*, 10887.

[5] Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.*, **2002**, *124*, 104.

[6] Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.*, **2000**, *122*, 3746.

[7] Zauhar, R. J.; Colbert, C. L.; Morgan, R. S.; Welsh, W. J. *Biopolymers*, **2000**, *53*, 233.

[8] Duan, G. L.; Smith Jr., V. H.; Weaver, D. F. *Mol. Phys.*, **2001**, *99*, 1689.

[9] Tauer, T. P.; Derrick, M. E.; Sherrill, C. D. *J. Phys. Chem. A*, **2005**, *109*, 191.

[10] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover: New York, 1996.

[11] Jensen, F. *Introduction to Computational Chemistry*. Wiley & Sons: West Sussex, England, 2003.

[12] Rybak, S.; Szalewicz, K.; Jeziorski, B. *J. Chem. Phys.*, **1989**, *91*, 4779.

[13] Williams, H. L.; Szalewicz, K.; Jeziorski, B.; Moszynski, R.; Rybak, S. *J. Chem. Phys.*, **1993**, *98*, 1279.

[14] Bukowski, R.; Jeziorski, B.; Szalewicz, K. *J. Chem. Phys.*, **1996**, *104*, 3306.

[15] Mas, E. M.; Szalewicz, K. *J. Chem. Phys.*, **1996**, *104*, 7606.

[16] Willimas, H. L.; Korona, T.; Bukowski, R.; Jeziorski, B.; Szalewicz, K. *Chem. Phys. Lett.*, **1996**, *262*, 431.

[17] Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives*. VCH: New York, 1995.

[18] Burley, S. K.; Petsko, G. A. *Science*, **1985**, *229*, 23.

[19] Hunter, C. A.; Singh, J.; Thornton, J. M. *J. Mol. Biol.*, **1991**, *218*, 837.

[20] Brana, M. F.; Cacho, M.; Gradillas, A.; de Pascual-Teresa, B.; Ramos, A. *Curr. Pharm. Design*, **2001**, *7*, 1745.

[21] van de Craats, A. M.; Warman, J. M.; Mullen, K.; Geerts, Y.; Brand, J. D. *Adv Mater*, **1998**, *10*, 36.

[22] Engkvist, O.; Hobza, P.; Selzle, H. L.; Schlag, E. W. *J. Chem. Phys.*, **1999**, *110*, 5758.

[23] Ye, X. Y.; Li, Z. H.; Wang, W. N.; Fan, K. N.; Xu, W.; Hua, Z. Y. *Chem. Phys. Lett.*, **2004**, *397*, 56.

[24] Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. *Chem. Phys. Lett.*, **2004**, *394*, 334.

[25] de Meijere, A.; Huisken, F. *J. Chem. Phys.*, **1990**, *92*, 5826.

[26] Boys, S. F.; Bernardi, F. *Mol. Phys.*, **1970**, *19*, 553.

[27] Hankins, D.; Moskowitz, J. W.; Stillinger, F. H. *J. Chem. Phys.*, **1970**, *53*, 4544.

[28] Gauss, J.; Stanton, J. F. *J. Phys. Chem. A*, **2000**, *104*, 2865.

[29] Hopkins, B. W.; Tschumper, G. S. *Chem. Phys. Lett.*, **2005**, *407*, 362.

[30] Meyer, E. A.; Castellano, R. K.; Diederich, F. *Angew. Chem. Int. Ed.*, **2003**, *42*, 1210.

[31] Morgan, R. S.; Tatsch, C. E.; Gushard, R. H.; Mcadon, J. M.; Warme, P. K. *Int. J. Pept. Prot. Res*, **1978**, *11*, 209.

[32] Morgan, R. S.; Mcadon, J. M. *Int. J. Pept. Prot. Res*, **1980**, *15*, 177.

[33] Reid, K. S. C.; Lindley, P. F.; Thornton, J. M. *FEBS Lett.*, **1985**, *190*, 209.

[34] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.*, **1977**, *112*, 535.

[35] Allen, F. H. *Acta Crystallogr. B*, **2002**, *58*, 380.

[36] Viguera, A. R.; Serrano, L. *Biochemistry*, **1995**, *34*, 8771.

[37] Munoz, V.; Serrano, L. *J. Mol. Biol.*, **1995**, *245*, 275.

[38] Cheney, B. V.; Schulz, M. W.; Cheney, J. *Biochim. Biophys. Acta*, **1989**, *996*, 116.

[39] Yamaotsu, N.; Moriguchi, I.; Kollman, P. A.; Hirono, S. *Biochim. Biophys. Acta*, **1993**, *1163*, 81.

[40] Spencer, D. S.; Stites, W. E. *J. Mol. Biol.*, **1996**, *257*, 497.

[41] Pranata, J. *Bioorg. Chem.*, **1997**, *25*, 213.

[42] Prikhod'ko, I. V.; Vinogradova, I. V. *Russ. J. Appl. Chem.*, **2002**, *75*, 1774.

[43] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.*, **1989**, *157*, 479.

[44] Grimme, S. *J. Comp. Chem.*, **2004**, *25*, 1463.

[45] Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.*, **1992**, *96*, 6796.

[46] Arunan, E.; Emilsson, T.; Gutowsky, H. S.; Fraser, G. T.; de Oliveira, G.; Dykstra, C. E. *J. Chem. Phys.*, **2002**, *117*, 9766.

[47] Basis sets were obtained from the Extensible Computational Chemistry Environment Basis Set Database, Version 12/03/03, as developed and distributed by the Molecular Science Computing Facility, Environmental and Molecular Sciences Laboratory which is part of the Pacific Northwest Laboratory, P.O. Box 999, Richland, Washington 99352, USA, and funded by the U.S. Department of Energy. The Pacific Northwest Laboratory is a multi-program laboratory operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC06-76RLO 1830. Contact David Feller or Karen Schuchardt for further information.

[48] Edwards, T. H.; Moncur, N. K.; Snyder, L. E. *J. Chem. Phys.*, **1967**, *46*, 2139.

[49] MOLPRO, a package of ab initio programs designed by H.-J.Werner and P. J. Knowles, version 2002.1, R. D. Amos, A. Bernhardsson, A. Berning, P. Celani, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, C. Hampel, G. Hetzer, P. J. Knowles, T. Korona, R. Lindh, A. W. Lloyd, S. J. McNicholas, F. R. Manby, W. Meyer, M. E. Mura, A. Nicklass, P. Palmieri, R. Pitzer, G. Rauhut, M. Schütz, U. Schumann, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, and H.-J. Werner. http://www.molpro.net.

[50] Jeziorski, B.; Moszynski, R.; Szalewicz, K. *Chem. Rev.*, **1994**, *94*, 1887.

[51] Sinnokrot, M. O.; Sherrill, C. D. *J. Am. Chem. Soc.*, **2004**, *126*, 7690.

[52] Bukowski, R.; Cencek, W.; Jankowski, P.; Jeziorski, B.; Jeziorska, M.; Kucharski, S. A.; Misquitta, A. J.; Moszynski, R.; Patkowski, K.; Rybak, S.; Szalewicz, K.; Williams, H. L. *SAPT2002: An Ab Inito Program for Many-Body Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Interaction Energies. Sequential and Parallel Versions*, 2003. http://www.physics.udel.edu/∼szalewic/SAPT/SAPT.html.

[53] Dunning, T. H. *J. Chem. Phys.*, **1989**, *90*, 1007.

[54] van Duijneveldt, F. B.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Lenthe, J. H. *Chem. Rev.*, **1994**, *94*, 1873.

[55] Halkier, A.; Klopper, W.; Helgaker, T.; Jørgenson, P.; Taylor, P. R. *J. Chem. Phys.*, **1999**, *111*, 9157.

[56] Hopkins, B. W.; Tschumper, G. S. *J. Phys. Chem. A*, **2004**, *108*, 2941.

[57] Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.*, **2000**, *122*, 11450.

[58] Figgs, M.; Sinnokrot, M. O.; Sherrill, C. D. unpublished results.