

AUTOMATION OF THE VIRTUAL WORKBENCH: A PROTOCOL FOR THE
ENTRY OF BIG DATA WITHIN A CHEMICAL DOMAIN

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Yen Hoang Bui

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Prof. Lyudmila V. Slipchenko
Department of Chemistry

Prof. Dor Ben-Amotz
Department of Chemistry

Prof. Sabre Kais
Department of Chemistry

Prof. Mark Daniel Ward
Department of Statistics

Approved by:

Prof. Christine A. Hrycyna
Head of the Chemistry Graduate Program

To Hank, my dog.

ACKNOWLEDGMENTS

First and foremost, I need to acknowledge the unyielding support of my parents Huy Bui and Ngat Le. Without their efforts towards discipline, it is likely I would have never learned how to read.

Second, I must acknowledge my colleagues, YB and Claudia for their perspectives shared during the course of my research, and former colleagues Ben, Pradeep, Carlos for their guidance during the course of the program.

Third, to my computer science undergrads, Addison, Hanjing, Jia, and Qifeng, I owe a great deal. In attempting to build a bridge for communication between chemistry and computer science, I have gained a completely different perspective on how to approach science. I cannot possibly express how proud I am of our work and will miss our daily discussions.

Lastly, I must acknowledge my Professor, Lyudmila Slipchenko.

Despite, not having any real experience with computational chemistry or programming - she gambled a great deal and took me on as a graduate student. I do, have, and will continue to remain indebted to her.

PREFACE

Herein lies the culmination of my implementations and contributions to the development of EFP - a theoretical model by which non-covalent interactions are described - under the advisory of Prof. Lyudmila Slipchenko; a brief overview of a thesis detailing an intersection of computational science, chemistry, and big data analysis.

I spent the majority of my time at Purdue, attempting to get a computer - an artificial being - to perform simulations using various chemical models. The purpose of this work is to document my attempt to break down the work-flow process of running a LIBEFP calculation in order to automate it through the development of iSpiEFP.

TABLE OF CONTENTS

	Page
LIST OF TABLES	viii
LIST OF FIGURES	ix
ABSTRACT	xii
1 BACKGROUND	1
1.1 Chemical Visualization And Representation	1
1.2 Computational Methods for Chemically-Relevant System Representation	2
1.3 EFP Method Development	3
2 VALIDATION	8
2.1 Prior Studies	8
2.2 EFP on the SSI Dataset	9
2.2.1 Basisset Selection Benchmarks	10
3 IMPLEMENTATION	24
3.1 BioEFP	25
3.2 Pairwise Energy Component Decompositions using the BioEFP Method	27
3.2.1 Theoretical and Technical Implementation	27
3.2.2 Factor-Xa Ligand Binding	31
3.2.3 Methods	32
3.2.4 Results and Discussion	36
3.2.5 Conclusions	43
3.3 EFPMD-MC	43
3.3.1 Monte-Carlo Sampling	44
3.3.2 Technical Implementation	45
3.3.3 S22 Dataset	47
3.3.4 Conclusions	52

	Page
4 SIMILARITY MEASURE	53
4.1 Fingerprinting	56
4.1.1 Distance-Matrix	56
4.1.2 PRDF	60
4.2 Lysine Monomer	64
4.3 Random Amino Acid RMSD Cosine-Distance Measures	66
4.4 Conclusions	68
5 TOOLS	70
5.1 EFPdB	70
5.2 iSpiEFP	81
5.3 Conclusions	92
6 SUMMARY	93
REFERENCES	94

LIST OF TABLES

Table	Page
2.1 EFP Schemes	13
3.1 Difference in binding energies between Cl- and Me- ligands, computed using different methods.	42
3.2 EFP Energies (kcal/mol) Obtained through Geometry Optimization or Monte-Carlo on S22 Equilibrium Geometries	48
3.3 EFP Energies (kcal/mol) Obtained through Geometry Optimization or Monte-Carlo on S22 Nonequilibrium Geometries	49
3.4 EFP vs CI Reference Energies (kcal/mol) on Water Dimer Local Minima .	51
4.1 LYS Similarity Average Measures	66
5.1 Sample EFP Electrostatic Parameters	72
5.2 Sample Polarization EFP Parameters	73
5.3 Sample Dispersion EFP Parameters	73
5.4 Sample Exchange-Repulsion EFP Parameters	74
5.4 Sample Exchange-Repulsion EFP Parameters	75
5.5 EFPdB Cartesian Average Amino Acid RMSD Values	77
5.6 EFPdB Cartesian Average Amino Acid Cosine Distance Values	78
5.7 EFPdB PRDF Average Amino Acid RMSD Values	79
5.8 EFPdB PRDF Average Amino Acid Cosine Distance Values	80

LIST OF FIGURES

Figure	Page
2.1 SSI Amino Acid Monomer Classification	12
2.2 EFP-SAPT MAD MSD Errors	15
2.3 EFP-CCSD MAD MSD Errors	20
2.4 EFP-CCSD Total Energy IOWA plots	23
3.1 P-EFP Technical Implementation	28
3.2 Structure of the 3ENS-Cl molecule (red) bound to factor Xa.	32
3.3 S1 binding pocket of the 3ENS-Cl ligand-protein system. The pocket is shown in a ball and stick representation, while the ligand is shown using thick sticks.	33
3.4 Electron density map of 3ENS ligands with Cl- and Me- substitutions. Adapted from [74].	34
3.5 Cut-and-cap strategy used in BioEFP modeling of ligand-S1 pocket. Each amino acid residue is fragmented across two sites, resulting in a sidechain fragment and a peptide fragment.	35
3.6 Contribution of electrostatic interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.	37
3.7 Contribution of polarization interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.	37
3.8 Contribution of dispersion interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.	38
3.9 Contribution of exchange-repulsion interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.	39
3.10 Contribution of total interaction energies to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.	40
3.11 Total interaction energies and energy component contributions of individual fragments to $\Delta(\Delta E)$ term in the small ligand-S1 pocket model.	40
3.12 Pairwise contributions to $\Delta(\Delta E)$ term in small ligand-S1 pocket model as a function of distances between ligand and fragments.	41

Figure	Page
3.13 Convergence of $\Delta(\Delta E_{elec})$ term in the ligand - protein complex as a function of distance.	42
3.14 Convergence of $\Delta(\Delta E_{total})$ term in the ligand - protein complex as a function of distance. A running sum of the $\Delta(\Delta E_{total})$ term is plotted (orange line).	43
3.15 Monte-Carlo implementation scheme within <i>efpmd</i>	45
3.16 Selected S22 Dimer Complexes	47
3.17 Initial Monte-Carlo Water Dimer Configurations	51
3.18 Geometry Optimized Monte-Carlo Water Dimer Configurations	51
3.19 Local Minima Obtained From Monte-Carlo Optimized Water Dimer Configurations	52
4.1 Example of Isomers	57
4.2 Optical Isomers	60
4.3 Water and Ammonia Dimer with Atom Labels	61
4.4 Sample H-H Intermolecular Conversation to Histogram In Data Structure Array	62
4.5 Conversion of Histograms to 'Molecular Fingerprint'	63
4.6 Comparison of the Molecular Fingerprints of Configurations of Ammonia and Water	64
4.7 Similarity Comparison of Molecular Representations of Lysine Residues; A) RMSD similarity comparison using Cartesian Representation; B) RMSD similarity comparison using PRDF Representation; C) Cosine-Distance similarity comparison using Cartesian Representation; D) Cosine-Distance similarity comparison using PRDF Representation;	65
4.8 RMSD of Random Amino Acid Residues Versus Valine3947	67
4.9 Cosine Distance of Random Amino Acid Residues Versus Valine3947	68
5.1 EFPdB Standard Amino Acid Cartesian Representation RMSD values.	76
5.2 EFPdB Standard Amino Acid PRDF Representation RMSD values.	81
5.3 iSpiEFP Desktop Application	82
5.4 iSpiEFP General Workflow	84
5.5 Lysine Molecule Fragmented Into EFP Fragments along covalent bonds.	85

Figure	Page
5.6 iSpiEFP enables fragment parameter similarity selection using RMSD and molecular fingerprint	87
5.7 iSpiEFP workflow components	88
5.8 iSpiEFP Software Overview	89

ABSTRACT

Bui, Yen H. Ph.D., Purdue University, May 2019. Automation of the Virtual Workbench: A Protocol For the Entry of Big Data Within A Chemical Domain. Major Professor: Lyudmila Slipchenko.

Here we describe recent technical implementations and modifications to the libefp package as well as applications of those implementations. Applications of the EFP method to biologically relevant systems are provided on a benchmark EFP-SAPT-CCSD study on the SSI dataset along with suggested basis set recommendations and a study on the pairwise EFP total energy decomposition on Factor Xa. We also report the technical overview of two computational tools we believe will lower the human barrier to utilizing the EFP method - iSpiEFP and EFPdB.

1. BACKGROUND

In the most general sense, utilizing computation involves developing simplistic models with which dynamics can be performed. This is done by implementing a set of mathematical formalisms that capture the nature of the system of interest in a way that is computer programmable. However, in reality the natural phenomena one is attempting to model is often too complex for a general or simplified representation. Thus, areas of development for computational science within the realm of chemistry are the actual mathematical theoretical description of a system, the algorithms for simulating those systems, and the analysis to derive statistical properties to investigate scientific hypothesis regarding the large amounts of complex data (Big Data) following each investigation.

This thesis provides a brief background on the EFP formalism used to model/represent molecular systems, conclusions drawn from EFP benchmarking studies on the SSI dataset obtained from the bFDB database validating EFP to describe biologically relevant systems, a chapter on recent modifications of the LIBEFP API to further utilize the LIBEFP method to describe non-covalent interactions in biologically relevant systems, methods for coordinate transformations, and the development of iSpiEFP - a generalized workflow manager capable of automating the workflow of running a libefp calculation.

1.1 Chemical Visualization And Representation

Generalized models have played a crucial role in understanding chemical phenomena and evaluating potential mechanisms that explain their behavior. They serve as vital bridges between abstract concepts within theory and experimental observations. Models serve to guide inquiry, make predictions, analyze data, and draw

inferences [1, 2]. Thus for chemists, interactions between matter have been able to be elucidated through the development of atomic and chemical structure - from the rutherford model to the lewis structure [3, 4] and then to more modern variants of molecular representations such as the 'ball and stick' models [5] and more sophisticated space-filling models [6, 7]. The underlying trend across all chemical models is that atoms in molecules are bonded together in a definite order and the manner in which they were organized determine their chemical and physical properties.

Ideally, a mathematical model as a chemical description is a highly representative system. However, due to general limitations on computational resources, models are simplified instances of the molecular system of interest. As such, the approximations made to simplify these molecular models are only as viable as their ability to explain or predict natural phenomena - detailing atomic configuration and composition within a molecule. Thankfully, through the advancements in high performance distributed computing platforms and more efficient algorithm implementation methods, more robust mathematical models are possible for a more accurate description of a molecular system.

1.2 Computational Methods for Chemically-Relevant System Representation

Thus, through the combination of molecular modeling and computer programming, it possible to replace rigid ball and stick molecular representations with representative interconnected 'soft' spherical atom centers for simulation and analysis. These molecular representations can be derived using classical Newtonian mechanics and/or wave-based methods. With either method, interactions between neighboring atoms are modeled using a potential function to describe the molecular potential energy as a sum of energy terms that vary with respect to geometric configuration.

As such, computational methods for the calculation of energetics of molecular systems at various levels of theory are continuously in development. In the specific

case of non-covalent interactions, all methods attempt to provide a more accurate representation of chemical interactions to provide a correct description of chemical phenomena relevant to biology [8–11]. Ideally, these representations are described using variants of quantum-based (QM) methods such as Moller-Plesset (MP) perturbation theory [12], coupled cluster (CC) theory [13,14], and Density Functional Theory (DFT) methods [15–17]. These are the methods in which the energetics of a molecular system can be obtained using rigorous QM-based formalism. However, due to the high computational cost associated with these wave-function based methods, energetics have been described in terms of bonded and non-bonded interactions between atoms using mechanical models (MM) that employ transferable geometric parameters collectively known as a 'force field'. MM force fields have been designed for different purposes and to represent different systems. Assisted Model Building and Energy Refinement (AMBER) [18] and Chemistry at HARvard Molecular Mechanics (CHARMM) [19,20], Groningen Molecular Simulation (GROMOS) [21,22] and Optimized Potential for Liquid Simulations (OPLS) [23] have been developed for biologically relevant molecules and macro-molecules. However, the accuracy of MM models are limited by the degree of transfer-ability of parameters from molecule to another. It is also possible to couple these methods, using hybrid QM/MM [24,25] and fragment-based methods in different schemes to combine the accuracy of a QM description with the low computational cost of molecular mechanics.

1.3 EFP Method Development

A promising method is the Effective Fragment Potential (EFP) method [26,27] that serves as a bridge between computational efficiency and a rigorous ab initio-based formulation of interaction energy in fragmented non-covalent systems. The building block or the idea behind EFP, is to fragment the system into subsystems or molecular "fragments" for which parameters can be obtained from an electronic structure calculation in the gas phase. From this calculation, a set of properties such

as point charges and multipoles, static and time-dependent polarizabilities, localized wave functions, etc. are obtained. Thus, EFP method can be thought of as a force field in which both a functional form and parameters originate from first principles. The EFP method describes the energetics of the system as a sum of components related to those parameters:

$$E^{EFP} = E_{Coul} + E_{Pol} + E_{ExRep} + E_{Disp} + E_{CT} \quad (1.1)$$

where E_{Coul} can be thought of as the interaction between electronic densities or distributed multipoles located on each fragment atom and bond-midpoint, E_{Pol} is the energy lowering due to the change in electronic distribution of a fragment in the presence of the electric field resulting from other fragments, E_{ExRep} as the non-classical term arising from the Pauli Exclusion Principle, E_{Disp} as the interaction between instantaneous dipoles within a system, and E_{CT} as the resonance stabilization energy due to the transfer of charge from one fragment to another.

Electrostatic Interactions Simplistic representations of the electrostatic interaction as interactions between atom-centered fractional partial charges are relatively easy to implement. However, these presentations often lack sufficient mathematical flexibility to describe the electrostatic potential of a fragment. This is a concern, as the relative contribution of electrostatic contributions to the total energetic description of the system has the largest magnitude and the longest range of intermolecular interactions components. Williams [28] showed that optimal least-squares fits of atom-centered partial charges resulted in relative rms errors of 3-10% over a set of grid points in a shell outside the surface of a series of small polar molecules. These errors were reduced by 2-3 orders of magnitude when electrostatics are modeled using atomic multipolar representations rather than partial charges. Thus multipolar representations can serve as an alternative method to describe the electrostatic potential. Within the context of the EFP method, multipolar expansion is derived directly from the molecular wave function distributed at atoms and bond midpoints so that the electrostatic

interactions between fragments A and B can be obtained as:

$$\begin{aligned}
 E^{AB} = & q^B [T q^A - T_\alpha \hat{\mu}_\alpha^A + \frac{1}{3} T_{\alpha\beta} \hat{\Theta}_{\alpha\beta}^A - \frac{1}{15} T_{\alpha\beta\gamma} \Omega_{\alpha\beta\gamma}^A + \dots] + \\
 & \hat{\mu}_\alpha^B [T_\alpha q^A - T_{\alpha\beta} \hat{\mu}_\beta^A + \frac{1}{3} T_{\alpha\beta\gamma} \Theta_{\beta\gamma}^A - \dots] + \\
 & \hat{\Theta}_{\alpha\beta}^B [T_{\alpha\beta} q^A - T_{\alpha\beta\gamma} \hat{\mu}_\gamma^A + \frac{1}{3} T_{\alpha\beta\gamma\delta} \hat{\Theta}_\gamma^A - \dots] + \\
 & \hat{\Omega} [-\frac{1}{15} T_{\alpha\beta\gamma} q^A + \dots]
 \end{aligned} \tag{1.2}$$

where q_A is the charge on centroid A and q_B is the charge on centroid B , etc. T are the electrostatic tensors of rank 0,1,2, etc. q , μ , Θ , and Ω are the point charge, dipole, quadrupole, and octupole moments of a fragment. Thus, these discrete Taylor-series type expansions approximate the functional electrostatic interactions. These electrostatic parameters are stored as positions of localized distributed multipoles located at fragment atomic centers and covalent bond midpoints in cartesian space. In a system with multiple fragments, electrostatic interactions between fragments can be obtained using simple classical interactions between the aforementioned distributed multipolar interactions; i.e. point charges, dipoles, quadrupoles and octupoles interaction with those on other fragments to obtain the total electrostatic energy component.

Polarization Interactions In the EFP method, polarization is described as the interaction of induced dipoles on a fragment with the static electric field produced by surrounding fragments:

$$E_{pol} = -\frac{1}{2} \sum_i \mu_{i,A} F(x_i) \tag{1.3}$$

$$F(x_i) = \sum_{B \neq A} \left(\sum_{j \in B} F_j^{mult}(x_i) + \sum_{J \in B} F_J^{nuc}(x_i) \right) \tag{1.4}$$

where $F(x_i)$ refers to the total static electric field acting on polarizability expansion point i on fragment A , and $\mu_{i,A}$ refers to the induced dipole on i computed as:

$$\mu_{i,A} = \alpha_{i,A} F^{total,i} \quad (1.5)$$

where $\alpha_{i,A}$ is the distributed polarizability tensor at i and $F^{total,i}$ is composed from the static field and the field due to other induced dipoles. The number of induced dipoles on a fragment is determined by the number of valence molecular orbitals it has. The location of each induced dipole is placed at a centroid (CT) as the origin of localized molecular orbital (LMO) in the valence shell. The EFP method utilizes these induced dipoles (polarizability tensors) at these LMO centroids.

Dispersion Interactions Dispersion interactions can be represented as a series:

$$E_{Disp} = \frac{C_6}{R^6} + \frac{C_8}{R^8} + \frac{C_{10}}{R^{10}} + \dots \quad (1.6)$$

where each term in the equation represents induced dipole-induced dipole, induced dipole-induced quadrupole etc. interaction. The distributed expression for the first dipole-dipole terms is given as:

$$E_{Disp} = \sum_{k \in A} \sum_{j \in B} \sum_{\alpha\beta\gamma\delta}^{x,y,z} T_{\alpha\beta}^{kj} T_{\gamma\delta}^{kj} \int_0^{\infty} dv \alpha_{\alpha\gamma}^i(iv) \alpha_{\beta\gamma}^j(iv) \quad (1.7)$$

where the dipole-dipole interaction energy is given as a sum of LMO-LMO contributions where k and j are LMO centroids on fragments A and B, $T_{\alpha\beta}^{kj}$ the electric field gradient tensor, and $\alpha_{\alpha\gamma}^k$ and $\alpha_{\beta\delta}^j$ the dynamic polarizability tensor elements for each LMO with respect to the imaginary frequency of the perturbing field iv . Higher order terms for dispersion can be approximated as 1/3 of the C_6 term and the dispersion interaction energy can be rewritten as:

$$E_{disp} = -\frac{3}{\pi} \sum_{k \in A} \sum_{j \in B} \frac{1}{R_{kj}^6} \int_0^{\infty} \alpha^k(i\omega) \alpha^j(i\omega) d\omega \quad (1.8)$$

Exchange Repulsion Interactions Exchange repulsion energetic contributions are the result of interpenetration of electronic charge densities. The total repulsion energy between i and j LMOs on pairs of fragments A and B :

$$E_{ExRep} = \sum_{i \in A} \sum_{j \in B} E_{xr}^{ij} \quad (1.9)$$

where the individual interactions on a system are approximated using static wave functions of fragments A and B :

$$\begin{aligned} E_{ExRep}^{ij} = & 4 \sum_{i \in A} \sum_{j \in B} \sqrt{\frac{-2 \ln S_{ij}}{\pi}} \frac{S_{ij}^2}{R_{ij}} \\ & - \sum_{i \in A} \sum_{j \in B} S_{ij} \left[\sum_{k \in A} F_{ik}^A S_{kj} + \sum_{l \in B} F_{jl}^B S_{li} - 2T_{ij} \right] \quad (1.10) \\ & - \sum_{i \in A} \sum_{i \in B} S_{ij}^2 \left[\sum_{J \in B} -Z_J R_{iJ}^{-1} + 2 \sum_{l \in B} R_{il}^{-1} + \sum_{i \in A} -Z_I R_{Ij}^{-1} + 2 \sum_{k \in A} R_{kj}^{-1} - R_{ij}^{-1} \right] \end{aligned}$$

where S_{ij} and T_{ij} are the overlap and kinetic energy integrals between LMOs i and j . F_{ik}^A is the Fock matrix element between LMOs i and k resulting from the Hamiltonian of fragment A . R_{iJ} is the distance between the centroids of charge of LMOs i and nucleus J (with nuclear charge Z_J).

The summation these EFP energy components is how we quantify and describe the total non-covalent interactions between EFP fragments. However, the original EFP formalisms were developed to model general solute-solvent interfaces and only recently has EFP been developed to study covalently bonded molecular systems [29–31]. In order to validate the EFP method on biologically relevant systems, benchmarking studies were performed on the SSI dataset and basis set recommendations were obtained for the best system description. Those findings are reported in the next chapter.

2. VALIDATION

Noncovalent interactions in biomolecules play an essential role in DNA helical stabilization [32], protein-ligand binding [33], and oligopeptide conformation [34]. Understanding the nature of noncovalent interactions such as hydrogen bonding, π -stacking, hydrophobicity, and metal coordination will enable further insight and eventually control of biochemical or biophysical processes. However, predictive modeling of noncovalent interactions requires using an appropriate level of theory that is both accurate and computationally feasible. While correlated quantum mechanical (QM) methods such as perturbation theory or coupled cluster (CC) methods [35,36] provide sufficient accuracy in describing non-covalent interactions, their computational scaling is N^5 or higher, where N is the size of the system of interest, making them unfeasible for practical simulations of biological complexes. On the other hand, molecular mechanical (MM) force fields such as AMBER [18] and CHARMM [19,37] are extremely efficient but might not be always transferrable from system to system and do not explicitly capture important components of noncovalent interactions such as polarization and charge transfer. Recent developments in density functional theory and fragmentation techniques provide new avenues for predictive modeling of extended systems [38,39]. The Effective Fragment Potential (EFP) method is a computationally efficient alternative technique for obtaining a description of inter-molecular interactions from the first principles

2.1 Prior Studies

The EFP method was originally introduced as a model potential for describing aqueous solvation (so called EFP1 model) [40,41]. Later on, the method was generalized to any solvent [42–44]. Recent work extended EFP to biological polymers [29].

EFP describes non-covalent interactions as a sum of coulomb, polarization, dispersion, exchange-repulsion and charge-transfer terms, all of which are derived from first principles and use information from electronic structure calculations on unique fragments. The EFP method has been previously applied for modeling structures and binding energies in molecular clusters, such as clusters of water, water and alcohols, aromatic molecules, etc. [45–47], as well as properties of complex liquids [48,49]. EFP can be also used as a polarizable force field in hybrid QM/EFP calculations [50]. Previously, we showed that the accuracy of the EFP method in describing noncovalent interactions of the S22 dataset is comparable to that of the second order perturbation theory (MP2) and exceeds that of classical force fields [51]. We also demonstrated that EFP energy components have good correspondence to those of the symmetry adapted perturbation theory (SAPT). However, while S22 dataset is designed as diverse and biologically relevant, it remains to be seen whether EFP preserves high accuracy for native biological complexes. The present work addresses this question by benchmarking EFP against coupled cluster with singles, doubles and perturbative triples [CCSD(T)] method and SAPT on a biological database BFDdb [52].

2.2 EFP on the SSI Dataset

The biomolecular fragment database (BFDdb) is a new large and diverse set of dimers extracted from crystallographic structures of 47 proteins. BFDdb has three components: dimers composed from aminoacid (AA) side chains (side chain - side chain interactions, SSI), dimers where both fragments are peptide groups (backbone-backbone interactions, BBI), and mixed dimers with one monomer being an AA side chain and the other the backbone peptide group (backbone-side chain interactions, BSI). Thus, BFDdb encompasses a great variety of noncovalent interactions present in proteins. It has been shown that the SSI part of BFDdb is more diverse in terms of decomposition of noncolvalent interactions (see Fig.1 from BFDdb [52]) than most of other available databases. In the present work we benchmark EFP against SAPT0 and

CCSD(T) on the SSI database and provide estimates of accuracy of EFP for describing noncovalent interactions in proteins. We also explore how accuracy of EFP depends on the basis set employed for parameterization, and provide recommendations for the optimal strategy of performing EFP simulations on protein complexes.

2.2.1 Basisset Selection Benchmarks

Theoretical and Computational Details The EFP method describes the noncovalent interaction energy as seen Eq 1.1. Here, EFP energy terms are benchmarked against reference values obtained through symmetry adapted perturbation theory (SAPT). The simplest SAPT model, SAPT0, represents the noncovalent interaction energy as:

$$E^{SAPT0} = E_{elst}^{(10)} + E_{exch}^{(10)} + \left[E_{ind,r}^{(20)} + E_{exch-ind,r}^{(20)} + \delta E_{HF}^{(2)} \right]_{ind} + \left[E_{disp}^{(20)} + E_{exch-disp}^{(20)} \right]_{disp} \quad (2.1)$$

where three terms in the first brackets contribute to the induction energy, and two terms in the second brackets describe dispersion energy. $\delta E^{(2)HF}$ is a Hartree-Fock (HF) correction that primarily accounts for polarization correction beyond the second-order $E_{ind}^{(20)}$. Essentially, SAPT0 describes monomers at the HF level and adds explicit dispersion term from the second order perturbation theory to augment electrostatic, induction and exchange-repulsion components derived from the HF treatment of the dimer interaction energy. In this respect, EFP is similar to SAPT0 as EFP parameters are also based on the HF description of fragments, and as such, we should expect the closest agreement between EFP and SAPT0 rather than of EFP with higher-order SAPT models.

Sherrill and coworkers introduced the gold, silver, and bronze standards of SAPT models based on combined performance on several data sets of noncovalent interactions (HSG, S22, NBC10 and HBC6) [53] and computational cost. These are SAPT2+(3) MP2/aug-cc-pVTZ (gold), SAPT2+/aug-cc-pVDZ (silver), and sSAPT0/jun-

cc-pVDZ (bronze), which provide mean absolute errors (MAE) of 0.15 kcal/mol, 0.30 kcal/mol, and 0.49 kcal/mol, respectively [53]. Example computational times of the corresponding methods on a hydrogen-bonding adenine-thymine dimer are 62.9 h, 4.4 h and 0.03 h for gold, silver and bronze standards, respectively, using 16 processors on one node. Using the same metrics, MAE of the SAPT0/jun-cc-pVDZ method is 0.86 kcal/mol, while the computational cost is identical to sSAPT0 in the same basis set.

The highest level of theory applied to the SSI database is the DW-CCSD(T**)-F12 method [54] with the aug-cc-pV(D+d)Z basis set. This level of theory, corresponding to the "silver standard" of the correlated wave function methods, achieves average MAE of 0.06 kcal/mol for the same set of four databases (HSG, S22, NBC10, HBC).

For benchmarking EFP on the SSI dataset, the total interaction EFP energies were compared against those computed with DW-CCSD(T**)-F12/aug-cc-pV(D+d)Z, referred to as CCSD(T)/adz in the following, while the EFP energy components were compared against SAPT0/jun-cc-pVDZ energies, referred to as SAPT0/jdz. The following metrics were used:

$$err_i = E_{EFP,i} - E_{REF,i} \quad (2.2)$$

$$MAD = \frac{1}{N} \sum_{i=1}^N |err_i| \quad (2.3)$$

$$MSD = \frac{1}{N} \sum_{i=1}^N err_i \quad (2.4)$$

$$MRD = \frac{1}{N} \sum_i \frac{|err_i|}{|E_{REF,i}|} * 100\% \quad (2.5)$$

$$STD = \sqrt{\frac{\sum_{i=1}^N (err_i - MSD)^2}{N - 1}} \quad (2.6)$$

where $E_{EFP,i}$ is the EFP total energy or energy component for i th compound of

the dataset, $E_{REF,i}$ is the same for the reference method, i.e., (CCSD(T)/adz or SAPT0/jdz). N is the total number of members of the dataset. MAD, MSD, MRD and STD are mean absolute deviation, mean signed deviation, mean relative deviation and standard deviation, respectively.

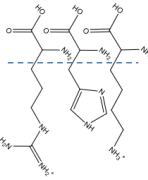
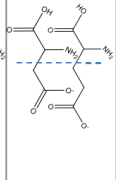
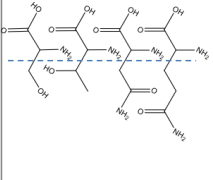
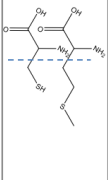
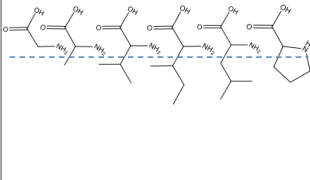
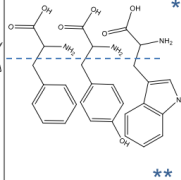
Cationic			Anionic		Polar				Thiol		Aliphatic						Aryl		
R	H	K	D	E	S	T	N	Q	C	M	G	A	V	I	L	P	F	Y	W
Arg	His	Lys	Asp	Glu	Ser	Thr	Asn	Gln	Cys	Met	Gly	Ala	Val	Ile	Leu	Pro	Phe	Tyr	Trp
																			
																			★★

Fig. 2.1.: SSI Amino Acid Monomer Classification

Since the SSI database consists of dimers composed of fragments extracted from amino acid (AA) side chains, the fragments are characterized by their parent amino acid residues (see Fig. 2.1). However, not all fragments match the original amino acid AA residues, because capping of the fragments was made at the closest to the interaction site first occurring sp³ carbon. In some cases, even the charge of an AA residue is different from the charge of a fragment, e.g. a lysine residue might be represented as ethane rather than entire amino acid fragment with 2 sp³ carbons in the alkyl chain rather than 4 sp³ carbons and an amine moiety. However, despite some of the differences, it is still valuable to characterize fragments in the same way as original AAs as cationic (ARG and LYS), anionic (ASP and GLU), polar (SER, THR, ASN and GLN), thiol (CYS and MET), nonpolar or aliphatic (ALA, VAL, ILE, LEU and PRO), and aryl (PHE, TYR, HIE and TRP). Similarly, interactions between fragments can be classified as anionic-polar, polar-polar, polar-non-polar, etc. A more general classification into three major groups of interactions, i.e. ionic, polar and nonpolar, is also used. In this case ionic group includes anionic-anionic, anionic-cationic and cationic-cationic interactions, polar group is composed of polar-

containing monomers, and nonpolar group consists of interactions of non-charged species.

EFP parameters for all fragments were computed in the GAMESS electronic structure package [55]. The basis sets and short-range damping settings used for preparing EFP potentials of interactions fragments are summarized in Table 2.1.

Table 2.1.: EFP Schemes

Name	DMA Basis	NonDMA Basis	Electrostatic Screens	Polarization Screenings	Charge-Transfer
S	S	S	overlap	default	no
M	M	M	overlap	default	no
B	B	B	overlap	default	no
SM	S/M	M	overlap	default	no
SMB	S/M	B	overlap	default	no
SM _{es}	S/M	M	exponential	default	no
SM _{ps}	S/M	M	overlap	pol 0.3 b	no
SMB _{ps}	S/M	B	overlap	pol 0.3 b	no
B _{ct}	B	B	overlap	default	yes
SMB _{ct}	S/M	B	overlap	default	yes
SMB _{ctps}	S/M	B	overlap	pol 0.3 b	yes

All EFP calculations are single point energy calculations on the unoptimized structures from the SSI dataset. EFP calculations with EFP potentials referred to as small (S), medium (M), and small-medium hybrids (SM) were performed in LIBEFP software library [56, 57]. Calculations with EFP potentials using 6-311+G(3df,2p) (B) basis set, i.e. big (B) and SMB hybrids (SMB) were performed in GAMESS.

As summarized in Table 2.1, a number of basis sets and screening options were explored in this work. Small (S) scheme refers to all EFP parameters computed in 6-31G* basis; medium (M) refers to 6-31+G*; big (B) to 6-311++G(3df,2p). Hybrid schemes were generated as follows: in small-medium hybrid (SM), 6-31G* basis is used to compute multipoles in aryl fragments; 6-31+G* basis is used to compute multipoles in all other (non-aryl) fragments and other EFP terms. In small-medium-big hybrids (SMB), 6-311++G(3df,2p) basis is used for calculating polarization, disper-

sion, exchange-repulsion and charge-transfer parameters and a mixture of 6-31G* and 6-31+G* bases is used for computing multipoles (for aryl and non-aryl fragments, respectively). Default screenings are overlap-based screening for Coulomb term, Gaussian screening with default parameter of 0.6 for polarization, and overlap-based screening for dispersion. Exponential (SCREEN2) screening for Coulomb term is notated as "es"; a change in values of polarization damping parameter from 0.6 to 0.3 is indicated by "ps". Charge-transfer term is not included by default; when this term is included, it is noted as "ct" and its contribution is added to the polarization component (similar to how charge-transfer and higher-order polarization terms are combined in SAPT induction term).

In a previous benchmarking study of EFP on S22 and S66 databases we have shown that hybrid SMB scheme provides a balanced description of noncovalent interactions [51]. The present work builds upon this notion and provides a detailed exploration of how different basis sets affect the accuracy of individual EFP terms and the total interaction energies. The overall goal of this work is to provide recommendations on the optimal way to describe a whole plethora of noncovalent interactions occurring in proteins by the EFP method.

Theoretical and Computational Details We start discussion of accuracy of different EFP schemes by analysis of EFP energy components.

Figure 2.2 provides comparison of energy components computed with various EFP schemes and SAPT0. MAD, MSD and STD are reported for ionic, polar and non-polar interactions for the total SSI database. In addition to this general statistics, each dimer interaction in Figure 2.2 is illustrated as a 'thread' whose spatial orientation provides the error with respect to the SAPT energy values and color indicates the dominating force in the non-covalent interaction. The color scheme is adapted from Table II-IV of reference [52]. Namely, red color means that a noncovalent interaction is dominated by the electrostatic term, whereas the blue color corresponds to the dispersion-dominated interactions. Yellow and green colors describe interactions of mixed character. Grey vertical bars determine energy ranges of -10, -5, -2, 0, 2,

		Mean Abs. Error				Mean Sign. Error					
Energy	Basis Set	Non				Non				STD	Error
		Ionic	Polar	polar	Total	Ionic	Polar	polar	Total		
Electrostatics	S	2.82	1.59	0.73	1.13	2.82	1.58	0.73	1.13	1.20	
	M	1.06	1.00	1.52	1.28	1.05	0.94	0.47	0.72	3.18	
	B	1.87	1.22	0.81	0.99	1.87	1.20	0.61	0.88	1.39	
	SM	1.05	1.03	0.67	0.83	1.05	0.99	0.57	0.77	1.18	
	SM_es	0.55	0.73	0.54	0.54	-0.34	0.01	-0.28	-0.17	1.55	
	SMB	0.98	1.01	0.66	0.81	0.97	0.96	0.57	0.75	1.18	
Exchange Repulsion	S	9.75	4.55	2.25	3.42	-9.75	-4.52	-2.25	-3.42	3.26	
	M	2.43	1.24	0.52	0.86	-2.41	-1.15	-0.49	-0.81	1.01	
	B	0.52	0.49	0.31	0.39	-0.34	-0.37	-0.28	-0.32	0.41	
Polarization	S	3.17	1.24	0.27	0.73	3.17	1.24	0.27	0.73	1.27	
	M	0.76	0.77	0.22	0.39	0.49	0.73	0.1	0.27	0.65	
	B	0.86	0.61	0.23	0.36	-0.57	0.57	0.23	0.23	0.58	
	B_ct	2.07	0.34	0.18	0.37	-2.05	0.13	0.17	-0.01	0.82	
	SM	0.76	0.78	0.25	0.41	0.49	0.77	0.25	0.39	0.59	
	SM_es	0.76	0.78	0.25	0.41	0.49	0.77	0.25	0.39	0.59	
	SM_ps	1.91	0.91	0.25	0.52	1.87	0.9	0.25	0.51	0.91	
	SMB	0.96	0.58	0.25	0.37	-0.78	0.53	0.24	0.22	0.59	
	SMB_ps	1.04	0.71	0.25	0.42	0.72	0.67	0.25	0.37	0.7	
	SMB_ct	2.48	0.31	0.21	0.4	-2.46	-0.01	0.21	-0.05	0.93	
	SMB_ct_ps	1.04	0.32	0.21	0.3	-0.96	0.14	0.21	0.1	0.48	
Dispersion	S	0.6	0.45	0.35	0.38	0.58	0.31	0.33	0.34	0.69	
	M	0.33	0.27	0.17	0.21	-0.28	-0.14	0.07	-0.03	0.67	
	B	1.07	0.75	0.39	0.55	-1.07	-0.75	-0.36	-0.54	0.77	
Total	S	3.26	1.54	0.94	1.31	-3.18	-1.32	-0.92	-1.21	1.43	
	M	1.66	0.88	1.41	1.17	-1.15	0.44	0.15	0.16	3.37	
	B	1.13	1.11	0.68	0.79	-0.11	0.72	0.2	0.26	1.62	
	B_ct	1.88	0.82	0.67	0.77	-1.59	0.28	0.15	0.01	1.58	
	SM	1.66	0.93	0.68	0.8	-1.15	0.53	0.4	0.32	1.7	
	SM_es	2.64	1.34	0.9	1.12	-2.54	-0.45	-0.45	-0.61	1.95	
	SM_ps	0.94	1.02	0.68	0.77	0.22	0.66	0.4	0.45	1.67	
	SMB	1.58	0.93	0.55	0.7	-1.22	0.44	0.17	0.12	1.47	
	SMB_ps	1.22	1.05	0.56	0.72	0.27	0.58	0.18	0.27	1.52	
	SMB_ct	3.03	0.8	0.54	0.76	-2.9	-0.1	0.13	-0.16	1.6	
	SMB_ct_ps	1.71	0.82	0.54	0.67	-1.41	0.04	0.14	-0.01	1.38	

Fig. 2.2.: EFP-SAPT MAD| MSD| Errors

5, 10 kcal/mol. Thick black bars indicate average MSD values for each set. Shifting specific values to the right suggests a positive error between EFP and SAPT energy component, i.e., overestimation of repulsive terms (like exchange-repulsion) and un-

derestimation of magnitudes of attractive terms like polarizations, dispersion, and majority of electrostatic interactions.

Electrostatic Interactions We computed electrostatic (Coulomb) interactions using six different ways: in small 6-31G* basis set (S), using 6-31+G* basis (M), using 6-311++G(3df,2p) basis (B), and three hybrid schemes: using 6-31G* basis for aromatic fragments and 6-31+G* basis for other molecules, with overlap-based screening computed in 6-31+G* basis (SM), same combination but with overlap-based screening computed in 6-311++G(3df,2p) basis (SMB), and same combination but with exponential electrostatic screening (SM_es). Small (S) scheme produces the largest errors in describing ionic and polar interactions and the largest overall MSD, suggesting that multipoles computed in 6-31G* basis result in a significant underestimation of absolute values of Coulomb interactions, with electrostatically dominant (red) (and generally stronger) interactions having larger errors. Using 6-31+G* basis for all fragments (M scheme) improves description of ionic and polar interactions but breaks in describing non-polar interactions, with a known culprit of describing aromatic species for which DMA procedure in 6-31+G* basis diverges and produces extremely large values of multipoles. As a result, M scheme with its large standard deviation should not be used in practical calculations. For the same reason, B scheme with 6-311++G(3df,2p) basis might be also non-reliable, even though the problem is less obvious but could be noticed from larger STD and a large spread of errors in yellow-colored (mixed-type) interactions typical in aromatic-containing dimers. Three hybrid schemes that combine 6-31G* basis for computing multipole moments for aromatic molecules and 6-31+G* for others produce a more balanced description of electrostatic interactions. Among them, SM and SMB (two schemes with overlap-based damping) show similar behavior in which all electrostatic interactions are on average underestimated (less attractive) by EFP compared to SAPT0. MAD and MSD errors of both schemes are about 0.8 kcal/mol. Using alternative screening approach (SM_es scheme) leads to quite different results. MAD values of all interaction types decrease (with respect to those in SM and SMB schemes), and

MSEs become smaller and significantly different from MADs, suggesting that now EFP can both underestimate and overestimate electrostatic energies. This scheme produces the smallest overall MAE and MSE, but, interestingly, larger STD than SM and SMB schemes. Overall, SM_{es} scheme produces the most balanced description of electrostatic interactions.

Exchange-repulsion Interactions All EFP schemes underestimate exchange-repulsion with respect to SAPT0, and this tendency is larger in smaller basis sets. As follows from Fig. 3, MAD and MSD in small 6-31G* basis are 3.42 and -3.42 kcal/mol, respectively; they decrease to 0.86 and -0.81 kcal/mol in 6-31+G* and become 0.39 and -0.32 kcal/mol in 6-311++G(3df,2p). The better accuracy of the EFP exchange-repulsion in large diffuse basis sets is due to the assumption of a complete basis set employed in deriving the functional form of the exchange-repulsion term. Using at least one diffuse function (like in medium 6-31+G* basis) is necessary for reliable description of the exchange-repulsion term in EFP.

Dispersion Interactions Quantum-mechanical description of dispersion interactions (and underlying dynamic polarizabilities) is known to be basis-set sensitive, with a general understanding that large basis sets are required for convergence of numerical values of polarizabilities. In the present implementation of EFP, dispersion is described with a scaled C6 term that implicitly accounts for higher-order terms, such that larger bases will result in larger by magnitude but not necessarily more accurate dispersion term. As follows from Fig. 3, increase of the basis set results in an expected increase of the magnitude of dispersion interactions: MSE values change from 0.34 kcal/mol (underestimation) in 6-31G*, to -0.03 kcal/mol in 6-31+G* and -0.54 kcal/mol (overestimation) in 6-311++G(3df,2p), with similar values of STD in all three cases. However, if SAPT0 is taken as a reference, 6-31+G* basis (M scheme)

produces the smallest MSD and MAD values and the most balanced description of dispersion by EFP.

Polarization Interactions Finding the optimal way to describe polarization term is the most challenging task, as polarization energy depends on electrostatic multipoles (i.e. it is coupled to the electrostatic term), on the values of polarizability tensors that are known to slowly converge with the basis set size, as well as on polarization dampings and possible inclusion of the charge-transfer term. We have tried a variety of schemes for analysis of contributions of the different components. S scheme is clearly not sufficient to produce accurate polarization energies, both because of inaccurate multipoles and underestimation of polarizabilities. As a result, polarization energy is underestimated for all types of interactions, with the largest errors coming from strongly bound ionic complexes. Increasing the basis set to 6-31+G* (M and SM schemes) significantly improves description of the polarization energy. However, with these schemes, polar and non-polar interactions are still underestimated, while ionic interactions can be both under- and over-estimated. Further increase of the basis to 6-311++G(3df,2p) (B and SMB schemes) slightly improves description of polar complexes (however, EFP polarization is still underestimated in those) but worsens description of ionic dimers in which EFP polarization becomes overestimated. Overall, in terms of the total MADs, these four schemes (M, SM, B, SMB) perform quite similarly. Increasing strength of the polarization damping generally decreases absolute values of polarization energies, with more pronounced effects observed in strongly polarized complexes, i.e., polar and ionic. This effect is clearly seen by comparing SM - SM_{ps} and SMB - SMB_{ps} pairs. In both cases, increasing strength of polarization damping results in increase of the overall MAD and MSD values. On the other hand, charge-transfer term significantly increases magnitudes of polarization in polar and ionic complexes. In B_{ct} and SMB_{ct} schemes, polar interactions are described more accurately than in B and SMB, with smaller MADs and MSDs, however, effectively all ionic interactions become even more overpolarized. As a result, schemes with

included charge transfer term produce similar overall MADs (as compared to analogous schemes without charge-transfer), smaller MSD values (i.e., they are on average more balanced) but larger STDs, due to larger spread of errors in ionic complexes. SMB_ps_ct scheme combines opposite effects of (repulsive) polarization damping and (attractive) charge-transfer term. When compared to SAPT0 induction energy, this scheme produces the most balanced description of polarization energy of the database as a whole, with the smallest MAD (0.30 kcal/mol), MSD (0.10 kcal/mol) and STD (0.48 kcal/mol). On the other hand, due to inclusion of the charge-transfer term, SMB_ps_ct scheme is computationally more expensive than simpler SM and SMB, with only slight improvements in accuracy.

Total Interaction Energies While it is convenient to analyze EFP energy components with respect to SAPT, it is more meaningful to compare total interaction energies with respect to CCSD(T), as SAPT0 by itself has non-negligible errors. As follows from Fig. 2.3, SAPT0 underestimates all types of noncovalent interactions, with overall MAD=0.57 kcal/mol, MSD=0.51 kcal/mol and STD=0.53 kcal/mol. Specifically, underestimation of ionic interactions (i.e., MSD value) is 0.93 kcal/mol, and underestimation of polar and nonpolar interactions is 0.4 kcal/mol. Thus, while we will refer to SAPT0 total interaction energies, the main effort will be on comparison between EFP and CCSD(T).

As scheme M is not reliable due to large electrostatic errors in aromatic compounds, evidenced by very large STD values, we exclude it from further consideration. EFP schemes from the most bound to the least bound (for the whole dataset) are: S, SM_es, SMB_ct, SMB_ct_ps and B_ct, SMB, B and SMB_ps, SM, SM_ps. SMB_ct_ps scheme has the closest match to SAPT0, while SM_es has the smallest overall MSD with respect to CCSD(T). SMB_ps_ct scheme has the lowest overall MAD and the second lowest STD value, i.e., it is the most balanced scheme. Interestingly, the simplest S scheme shows a good performance compared to the CCSD(T) data, with MAD of 0.87 kcal/mol, which shares second place with schemes based on big or mixed basis













		Mean Abs. Dev.				Mean Sign. Dev.					
Energy	Basis Set	Non				Non				STD	Error
Total	S	2.53	1.19	0.54	0.87	-2.24	-0.89	-0.51	-0.70	1.40	
	M	1.81	1.25	1.58	1.43	-0.21	0.86	0.56	0.67	3.42	
	B	1.48	1.21	0.79	0.93	0.83	1.15	0.61	0.77	1.60	
	B_ct	1.75	0.9	0.76	0.86	-0.65	0.71	0.56	0.53	1.58	
	SM	1.80	1.31	0.9	1.10	-0.21	0.96	0.81	0.83	1.76	
	SM_es	2.05	1.28	0.76	0.95	-1.60	-0.02	-0.04	-0.10	1.96	
	SM_ps	1.54	1.40	0.90	1.10	1.17	1.09	0.81	0.96	1.73	
	SMB	1.64	1.15	0.68	0.88	-0.28	0.87	0.58	0.63	1.51	
	SMB_ps	1.63	1.26	0.68	0.91	1.22	1.01	0.59	0.78	1.54	
	SMB_ct	2.76	0.99	0.65	0.89	-1.95	0.33	0.54	0.36	1.66	
	SMB_ct_ps	1.67	1.01	0.65	0.81	-0.46	0.47	0.55	0.51	1.44	
	SAPT	0.96	0.66	0.43	0.57	0.93	0.42	0.41	0.51	0.53	

Fig. 2.3.: EFP-CCSD MAD| MSD| Errors

sets, and the lowest STD. So, while we do not recommend to use S scheme for analysis of individual energy components, as they are significantly in error with respect to SAPT0, the total interaction energies, even though systematically overestimated, are quite accurate due to favorable error cancelation. On the other hand, two-basis hybrids (SM, SM_ps, SM_es) result in the largest MAD and STD values, and as such, are not recommended if total interaction energies are the main target. However, in applications where EFP serves as a polarizable embedding for a QM region, i.e., when only electrostatic and polarization EFP terms are used, SM_es scheme is the one to be preferred. SMB hybrids show similar performance in terms of MAD (0.81 - 0.91 kcal/mol). As was already mentions, SMB_ct_ps, the scheme with stronger polarization screening and charge transfer term, is the most balanced. For the hybrid scheme without charge-transfer term, SMB produces the most accurate description of the overall dataset. B and B_ct schemes are not better than SMB hybrids, and as they might produce divergent electrostatic energies for aromatic compounds, they are not recommended for a broad use.

Overall, the three most reliable EFP schemes for describing interactions in proteins are (1) SMB_ct_ps hybrid, which is the most balanced but also the most computationally expensive scheme, (2) SMB hybrid, which results in less accuracy in describing

polar interactions while keeping the overall good performance for both total energies and energy components, but without necessity to evaluate expensive charge-transfer term, and (3) S scheme, the simplest and the fastest approach, which manages to keep the overall errors low. However, S scheme should not be used for analysis of individual energy components as its accuracy is based on significant error cancellations. Out of these three schemes, SMB_ct_ps is the best for describing polar interactions, SMB produces the most balanced description of ionic interactions, and S is the best for describing nonpolar interactions. SMB and SMB_ct_ps schemes underestimate most of polar and nonpolar interactions and slightly overestimate strong ionic interactions, while S scheme overstabilizes majority of interactions.

The following discussion considers accuracy of various EFP schemes in describing total interaction energies as compared to the silver standard CCSD(T^{**})-F12/ aug-cc-pV(D+d)Z. These data are summarized in IOWA plots (see Fig. 2.4), and 2.3 showing MAD, MSD and relative errors for specific types of interactions.

IOWA plots 2.4 show signed errors in a scatter plot in which each block represents a pair of AA monomers. Each such block encompasses all occurrences (shown as smaller squares) of the specific interactions in the database. Each entry in the IOWA plot is colored based on a purple-green gradient from -5 to 5 kcal/mol of total interaction energy errors of EFP against CCSD(T)/adz. Purple color corresponds to EFP overbinding and green color means that EFP underbinds. Intensely colored squares reveal interactions with large EFP errors; lighter regions correspond to interactions that are described more accurately.

Overall, two EFP schemes, S and SM_es, overbind compared to the reference CCSD(T)/adz. S overbinds most of the interactions, while SM_es strongly overbinds interactions involving tryptophan residue and most of ionic, polar and mixed ionic-polar interactions.

”Bad Boys” All considered EFP schemes fail to describe interactions in cysteine and methionine dimers, as well as interactions between cysteine and aspartic acid

residue. This is somewhat not surprising as many correlated, DFT and semiempirical methods produce the largest errors in these complexes [52]. As far as accuracy of EFP is under concern, the main problem comes from the HF-based DMA predicting sulfur in MET and CYS more electro-positive compared to neighboring hydrogen atom, such that electrostatic component of the sulfur hydrogen bond is missing. As a consequence, both electrostatic and polarization EFP terms become strongly unbound in these complexes. This issue is somewhat but not completely elevated by exponential electrostatic screening.

This chapter validates the ability of EFP to describe non-covalent interactions within molecular systems. The next chapter on implementation will detail the theoretical and technical modifications to the libefp [56, 57]. These modifications include EFP pairwise interactions between a ligand and individual amino acids and EFP phase-space sampling using Monte-Carlo method.



Fig. 2.4.: EFP-CCSD Total Energy Iowa plots

3. IMPLEMENTATION

Recent advances in computational chemistry have opened exciting applications in drug discovery and have shortened the timeline for drugs to reach the market [58–63]. Understanding the underlying chemistry of drug-target interactions enables faster development of drugs, and computational chemistry has assisted medicinal chemists for decades in achieving this goal [64]. In the recent years, due to major advances in supramolecular chemistry and crystallography, obtaining detailed information on the 3D structures of large macromolecules have become routine [65]. Computational chemistry has played an important role in determining the viability of ligand molecules to be used as drugs for a given target [66]. While molecular mechanics (MM) methods have been widely employed in computer aided drug design (CADD), due to the advent of high performance computing and faster, efficient codes, quantum mechanical (QM) methods are gaining importance in this field [67–69].

In order to accurately predict the binding energy between a ligand-protein in solvent phase, many factors need to be considered: the interaction energy between the two molecules, desolvation penalties due to the removal of solvent molecules occupied by the ligand, other solvent effects, temperature corrections, etc [70, 71]. A simpler means of modeling such a system involving a ligand-protein complex would be to start from a well-studied ligand-protein system, followed by performing substitution to the functional group(s) present in the ligand to measure the effects of substitution. If the only goal is to estimate the relative binding energies, this approach is simpler and faster.

In simulating the ligand-protein interactions, a common assumption made is to account only for local interactions. This local model takes into account only the interaction of the ligand with transferable contacts such as hydrogen bond interactions, pi-pi interactions, ion-pi interactions, halogen bond interactions etc, and is usually

restricted to the drug binding pocket. Here, using a variant of the EFP method, we fragment covalent bonds within a system and obtain EFP parameters to model entire ligand-protein interactions and estimate pairwise binding energies using single point energy calculations (See Sec 3.2). Also, we present the implementation of the Monte-Carlo method in libefp to perform phase-space sampling of ligand-protein with the EFP method with potential applications in protein-ligand docking. In Section 3.3 we present the modifications of libefp to explore the phase space of gas-phase dimers using Monte-Carlo in order to obtain potential energy minima.

3.1 BioEFP

The biomolecular effective fragment potential method (BioEFP), also known as macromolecular effective fragment potential method (mEFP), is an offshoot of the original implementation of the Effective fragment potential method (EFP) and is designed for modeling large covalent molecules such as proteins, lipids, etc [29]. While the original EFP method was constructed with an intent to describe the effect of solvent environment on the properties of the solute, the BioEFP method extends the idea to covalently-connected molecules by systematically fragmenting the macromolecule into smaller fragments.

A protein is a chain of repeating residues connected by covalent bonds. By making use of the idea of repeating residues, we can fragment the chain at designated bonds, cap them using capping fragments (hydrogens, in this case) and obtain the parameters corresponding to the capped fragments. We then remove the parameters that correspond to the capped atoms to obtain the parameters corresponding to the uncapped fragments. Once the parameters corresponding to all the uncapped fragments are obtained, we may perform an EFP simulation that corresponds to the entire protein. A detailed summary of this procedure is available in reference [29].

The electrostatics parameters corresponding to the capped atoms and the bond midpoints connected to the capped atoms are removed. As a result of the removal

of these parameters, the total charge of the fragment is now non-zero (or non-integer in the case of charged fragments). The excess charge resulting due to the removal of capped atoms is then added and redistributed to the nearest neighboring atom using the expand-remove-redistribute procedure [29]. Hence, the net charge of individual fragments and the overall system remains the same before and after this procedure.

As described earlier, the polarization and dispersion parameters are centered at the LMO centroids instead of the atoms themselves. Hence, removal of a capped atom results in the removal of the closest LMO, which is usually the LMO that corresponds to the sigma bond between hydrogen and the neighboring atom. This step is highly necessary to avoid polarization collapse, as the sigma bond LMOs corresponding to the capped hydrogens in neighboring fragments are positioned close to each other and could potentially overpolarize each other. This could result in a larger than expected or diverged polarization energy.

The only computed parameters corresponding to the exchange repulsion term are the non-canonical localized molecular orbital coefficients, basis set coefficients and the Fock matrix. The localized molecular orbitals are then constructed as a linear combination of pre-defined atomic orbital basis functions. When fragmenting the system for the purpose of computing parameters, one can remove the localized molecular orbitals corresponding to the fragmented atoms and bonds, or keep them as-is, since exchange-repulsion is computed in a pairwise fashion. While the former scenario could potentially underestimate the exchange-repulsion energies, the latter could overestimate it. An alternate strategy could be to remove either the orbitals corresponding to C_α carbon or the peptide carbon and maintain this consistency for the entire protein. For the purpose of this study, we have decided not to modify the exchange-repulsion parameters to maintain simplicity.

3.2 Pairwise Energy Component Decompositions using the BioEFP Method

3.2.1 Theoretical and Technical Implementation

Decomposition of total EFP energies into pairwise interactions requires modification of the libefp API and the efpmd program [56, 57]. Data structure `struct cfg` and `void main()` defined and initialized in `main.c` were modified to take user inputs such as ligand index integer and simulation `run_type` `psp`. Subroutine `sim_psp.c` was added specifically for the decomposition of pairwise interactions for a single point energy calculation within the efpmd program. However, the process for decomposition (`run_type psp`) is very similar to the single point energy simulation (`run_type sim_sp` routine). Pairwise Decomposition of total energy related parameters require specifying the index of the EFP fragment of interest as the ligand in the libefp input:

- `run_type [psp]`
- `ligand [default = 0]`

For initialization of the simulation, `void main()` calls on the `run_type sim_pcp()`. `sim_pcp()` uses the coordinates and parameters accessible through header file `common.h`. Using routine `check_fail()`, efpmd is able to communicate with the libefp and call on `efp_compute()` and `efp_get_energy()` in order to create an object `EFP_EXPORT` to return to efpmd program.

Significant modifications to `pol.c` were necessary due to the parallelization of the self-consistent procedure for calculating the polarization contribution for each ligand-fragment interaction. This is because libefp is parallelized in a master-slave scheme [57] and so certain data structure remain 'opaque' - readable but not writable. In order to circumvent this issue, new data structure `p_id_work_data` were created in `pol.c` and accessed using `compute_p_elec_field_range()` and `compute_p_elec_field()`, `get_p_induced_dipole_field()`, `compute_p_id_range()` and `compute_energy_range()`. In order to compute the interaction energies corresponding to the ligand-residue interactions, the following strategy was utilized:

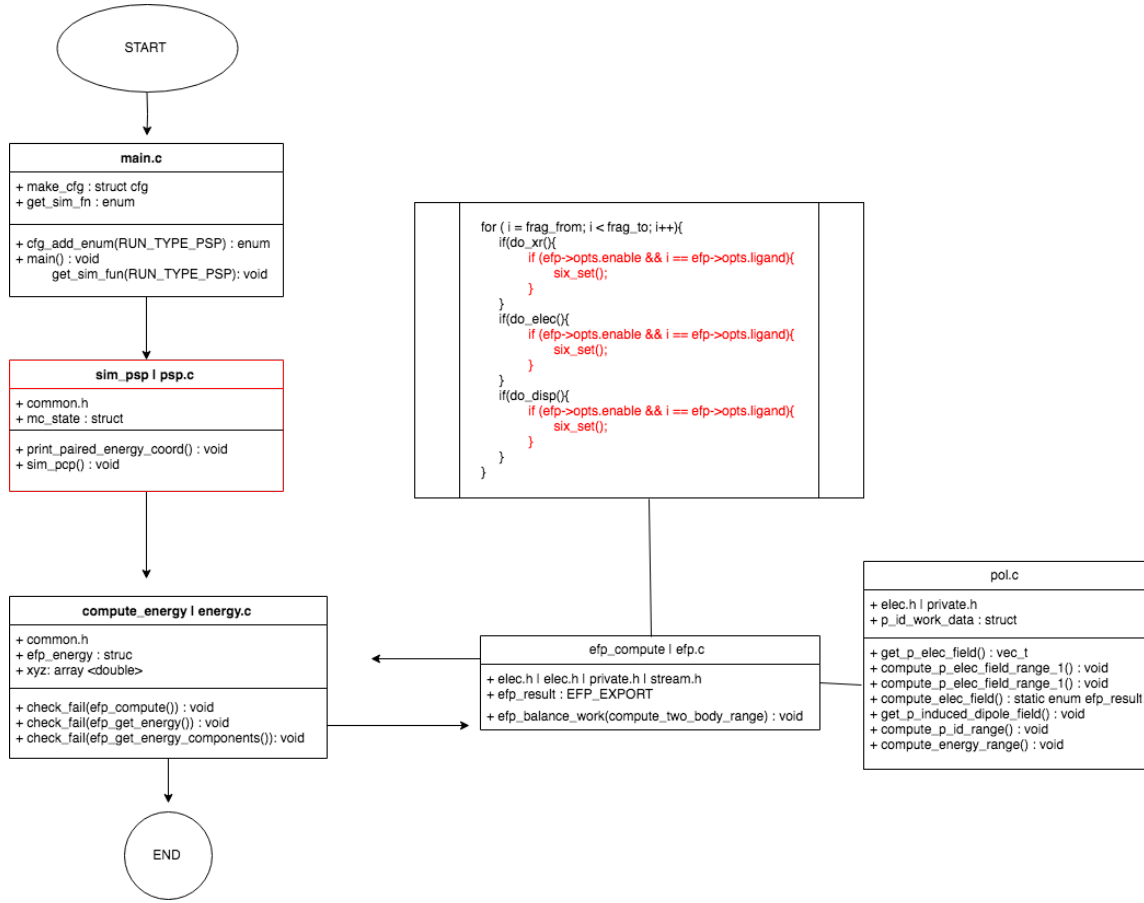


Fig. 3.1.: P-EFP Technical Implementation

1. For electrostatic interactions, the multipole-multipole interaction terms between ligand and the fragment in question were computed separately. From this information, the dimer electrostatic interaction energy was obtained. Hence, the potential due to electrostatic interactions at multipole expansion point $k \in$ fragment of interest LIG due to multipole expansion points $j \in$ fragment FR is given by:

$$\begin{aligned}
E_{Elec}^{LG-FR} = & \sum_{i \in LG} \sum_{j \in FR} q_i \hat{V}(r_{ij}) \\
& + \sum_{i \in LG} \sum_{j \in FR} \sum_{x'}^{x,y,z} \mu_i F_i(r_{ij}) \\
& + \frac{1}{3} \sum_{i \in LG} \sum_{j \in FR} \sum_{x' x''}^{x,y,z} \Theta_i F'_i(r_{ij}) \\
& + \frac{1}{15} \sum_{i \in LG} \sum_{j \in FR} \sum_{x' x'' x'''}^{x,y,z} \Omega_i F''_i(r_{ij})
\end{aligned} \tag{3.1}$$

where N represents the number of multipolar expansion points, $V(r_{ij})$ the electrostatic potential, μ , Θ , and Ω are the dipole, quadrupole, and octupole moments of a fragment respectively on multipolar expansion point i of ligand LG . F , F' , F'' are the electric field, field gradient, and field second derivative operators at point j on fragment FR .

2. For dispersion interactions, the dynamic polarizability tensors ($\alpha(i\omega)$) centered at the localized molecular orbitals corresponding to the ligand and the fragment in question were isolated and the interaction between these LMO points were computed separately. From this information, the dimer dispersion interaction energy between two fragments LG and FR was obtained:

$$E_{disp}^{LG-FR} = -\frac{3}{\pi} \sum_{k \in LG} \sum_{j \in FR} \frac{1}{R_{kj}^6} \int_0^\infty \alpha^k(i\omega) \alpha^j(i\omega) d\omega \tag{3.2}$$

where k and j are localized molecular orbitals in fragments LG and FR respectively.

3. For the exchange-repulsion term, the orbitals corresponding to the ligand and the fragment in question were isolated. Following this, the orbital-orbital exchange-repulsion interactions were computed, and the total exchange-repulsion interaction energy for the dimer is calculated:

$$\begin{aligned}
E_{Exch-Rep}^{LG-FR} = & 4 \sum_{i \in LG} \sum_{j \in FR} \sqrt{\frac{-2 \ln S_{ij}}{\pi}} \frac{(S_{ij})^2}{R_{ij}} \\
& - \sum_{i \in LG} \sum_{j \in FR} S_{ij} \left[\sum_{k \in LG} F_{ik}^{LG} S_{kj} + \sum_{l \in FR} F_{il}^{FR} S_{lj} - 2T_{ij} \right] \\
& - \sum_{i \in LG} \sum_{j \in FR} S_{ij}^2 \left[\sum_{J \in FR} -Z_J R_{iJ}^{-1} + 2 \sum_{l \in FR} R_{il}^{-1} + \sum_{I \in LG} -Z_I R_{ij}^{-1} + 2 \sum_{k \in LG} R_{kj}^{-1} - R_{ij}^{-1} \right]
\end{aligned} \tag{3.3}$$

where S is the overlap integral, R is the distance between orbital centroids i on fragment LG and j on fragment FR , F is the Fock matrix, T is the kinetic energy integral, Z is the nuclear charge corresponding to the two fragments LG and FR .

It must be noted that the electrostatics, dispersion and exchange repulsion terms are pairwise additive, and are directly comparable to the dimer interaction energies. The polarization term, however, is not pairwise additive, and hence must be computed in a fully interacting system.

4. The energy contributions due to polarization are then obtained from the converged induced dipole μ_i and μ_j moments centered at the localized molecular orbitals of fragment of interest LG and other fragment FR , respectively interacting with the electric fields generated by multipoles and nuclear charges of the other fragment $F(x_i)$ and $F(x_j)$.

$$E_{pol}^{LG-FRAG} = \sum_{i \in LG} \mu_i F(x_i) + \sum_{j \in FR} \mu_j F(x_j) \tag{3.4}$$

$$F(x_i) = \sum_{j \in FR} F_j^{mult}(x_{ij}) + \sum_{J \in FR} F_J^{nuc}(x_{iJ}) \tag{3.5}$$

$$F(x_j) = \sum_{i \in LG} F_i^{mult}(x_{ij}) + \sum_{I \in LG} F_I^{nuc}(x_{Ij}) \tag{3.6}$$

3.2.2 Factor-Xa Ligand Binding

Factor Xa is an activated form of thrombokinase, an enzyme that participates in the blood coagulation cascade. Antithrombotic agents corresponding to this enzymes have been developed and studied in detail [72, 73]. The active site of factor Xa consists of four subpockets: S1 - S4. The S1 subpocket plays a major role in selectivity and binding of the factor Xa. Earlier, Sherrill and coworkers [74] probed the effect of functional group modification in the factor Xa ligand using state of the art symmetry adapted perturbation theory (SAPT) methods. The interaction between neutral ligand and the anionic S1 pocket that comprises all the local contacts was computed using Functional SAPT (F-SAPT) [75, 76] and cut-and-cap SAPT (will be referred to as P-SAPT for simplicity) simulations truncated at the zero order (SAPT0) [77, 78]. The aim of the work was to understand the preferential binding of Cl- substituted ligands as compared to Me- substituted ligands. To achieve this aim, the interaction energies between all the sidechain residues and peptide backbone fragments and the ligand were computed (ΔE_{int}^{Cl} and ΔE_{int}^{Me}) in a completely interacting system using F-SAPT and a pairwise interacting system using P-SAPT. Calculating the difference in binding energies ($\Delta\Delta G$) using quantum mechanical methods is a computationally tedious process, hence the difference between the interaction energies ($\Delta\Delta E_{int} = \Delta E_{int}^{Cl} - \Delta E_{int}^{Me}$) can be used as an approximate measure of understanding the former quantity. This approximation can be deemed valid because of two reasons: 1. The vdW radii of the two functional groups are similar if not the same, hence the geometric effects due to substitution can be neglected; and 2. The polarities between the two functional groups are not vastly different, hence the desolvation penalties can be neglected as well. For ligand modifications involving substitution of vastly different functional group, this approximation is expected to break down.

Our intention here is to compare the performance of the biomolecular effective fragment potential method (BioEFP) with SAPT0 and to assess the validity of the local interaction model in predicting ligand-protein interactions.

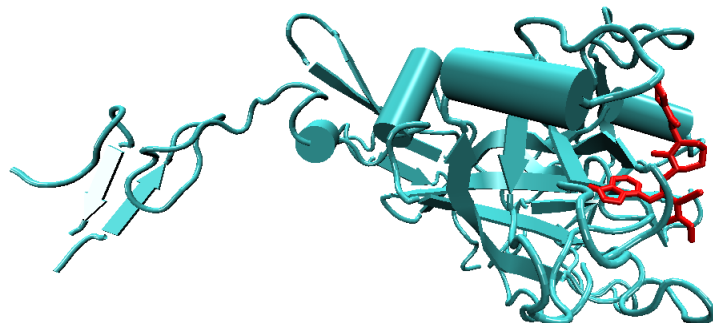


Fig. 3.2.: Structure of the 3ENS-Cl molecule (red) bound to factor Xa.

3.2.3 Methods

The protein conformation was obtained from Ref. [74]. Briefly, the following procedure describes the modification done by Sherrill et. al. to obtain the protonated ligand-protein structure : The geometry for factor Xa (the protein) in complex with methyl(2Z)-3-[(3-chloro-1H-indol-7-yl)amino]-2-cyano-3-[(3S)-2-oxo-1-(2-oxo-2-pyrrolidin-1-ylethyl)azepan-3-yl]aminoacrylate (the drug containing Cl- functional group) was obtained from RCSB database [79]. The Cl- functional group in the ligand was replaced by a carbon atom to obtain the methylated form of the ligand. Structures for both these analogues were prepared using the Protein Preparation Utility in Maestro (Schrodinger), which provides a rational estimate of optimal torsions, protonation states and the orientation of crystal water molecules. Following this, a constrained optimization was performed, to avoid obvious clashes and to minimize steric hindrances.

Earlier studies on the ligand indicate that following the substitution with the methyl ligand, the electron density shifts from the chlorine group to the farther end of the indole ring (Fig. 3.4). This would mean that the change in the nature of stabilizing/destabilizing interactions are not localized to the functional group alone. Further, it was shown that the single substitution enhanced the in-vitro IC_{50} (the concentration of the inhibitor at which the binding is reduced by half) by roughly 50 times. [79]. Major contributions to this enhanced efficacy were attributed to the interactions with Tyr228, Asp189, Gly219 and Cys220 residues. [79–81].

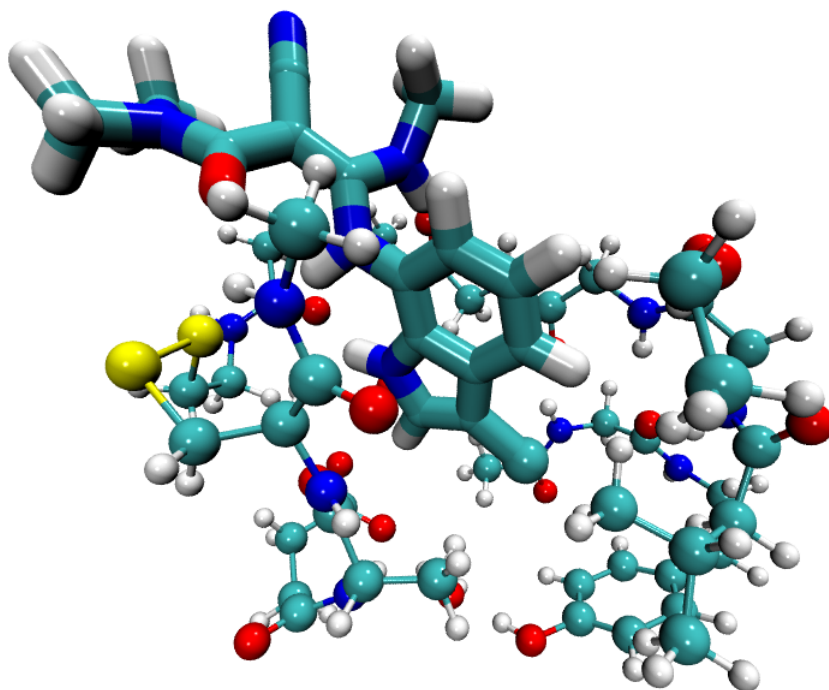


Fig. 3.3.: S1 binding pocket of the 3ENS-Cl ligand-protein system. The pocket is shown in a ball and stick representation, while the ligand is shown using thick sticks.

For the purpose of this study, we considered two model systems: 1. A smaller version of the ligand interacting with the fragments in the S1 pocket (to compare the performance of EFP with SAPT methods); and 2. The unmodified ligand interacting with all the fragments in the protein (to assess the validity of local model and compare

C. Cl vs Me

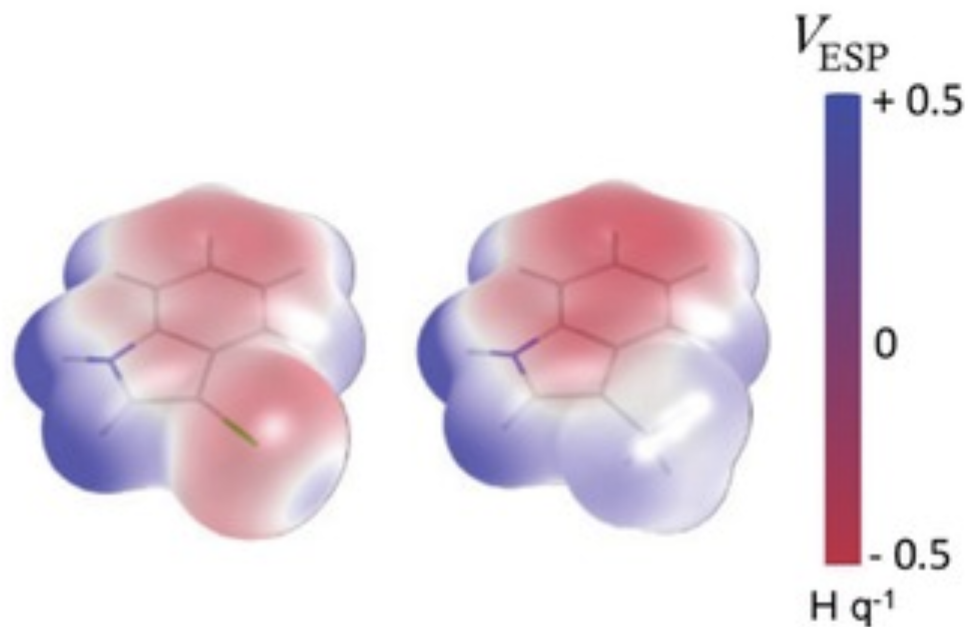


Fig. 3.4.: Electron density map of 3ENS ligands with Cl- and Me- substitutions. Adapted from [74].

it with our extended model). The S1 pocket consists of fragments that directly interact and bind with the ligand.

For the simulation of the S1 pocket with the ligand, only a small portion of the drug molecule that directly interacts with the pocket is included. While it is possible to include the whole drug molecule in the simulation, we decided to fragment the drug molecule and cap it with a hydrogen atom. SAPT0 results available in the literature were performed with this smaller version of the ligand, possibly due to practical restrictions in the simulation time, and hence following a similar strategy would enable a one-to-one comparison of our method with SAPT0. The amino acid residues were fragmented in such a way that the parameters for the sidechain and peptide groups were computed separately. In other words, the residues were fragmented along the $\text{C}\alpha$ -C bond as well as the $\text{C}\alpha$ -N bond (Fig. 3.5). This effectively results in two fragments

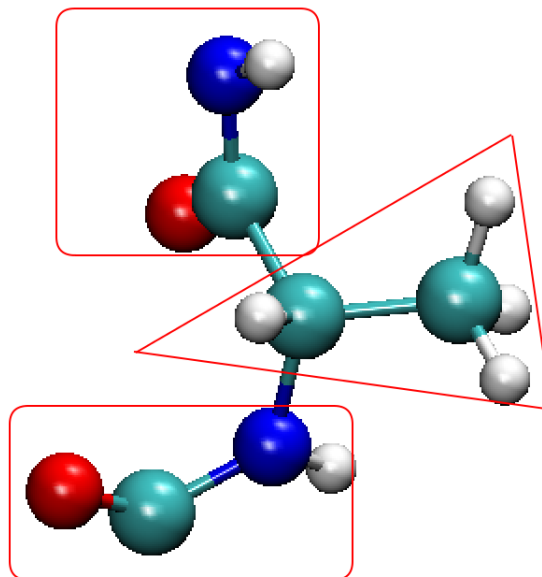


Fig. 3.5.: Cut-and-cap strategy used in BioEFP modeling of ligand-S1 pocket. Each amino acid residue is fragmented across two sites, resulting in a sidechain fragment and a peptide fragment.

per amino acid residue. The amino acid residues included in the S1 pocket are: ASP189, ALA190, CYS191, GLN192, SER195, VAL213, SER214, TRP215, GLY216, GLU217, GLY218, CYS220, GLY226, ILE227, and TYR228. All the fragmented residues were capped with hydrogen atoms. The disulfide bond between the cysteine residues were fragmented as well. As a result, the contribution of the disulfide bond to the total interaction energy is not explicitly computed using a separate disulfide fragment in BioEFP method. However, it must be noted that SAPT0 simulations were computed in such a way that the cysteine residue interactions and disulfide bond interactions were calculated separately. While it is not possible to directly compare the interaction energy components corresponding to these fragments, the sum of interaction energies due to the pair of cysteine EFP fragments can be compared to the sum of interaction energies due to cysteine and disulfide SAPT fragments.

All the EFP parameters were computed using HF/6-31G(d) basis set except for the exchange repulsion term, which was computed using a larger basis set (HF/6-

31+G(d)) for better accuracy. All the EFP-EFP simulations were performed using libefp package using the pairwise energy decomposition feature implemented recently.

3.2.4 Results and Discussion

In order to assess the performance of various interaction terms in the BioEFP method, we compare them directly with P-SAPT or F-SAPT data from ref. [74]. Fig 3.6 shows the contribution of electrostatic interactions to the $\Delta(\Delta E)$ term. Positive $\Delta(\Delta E)$ contributions indicate a preference of the methylated ligand over the chlorinated ligand, while negative $\Delta(\Delta E)$ contributions indicate the opposite. BioEFP correctly predicts the stronger preferential interactions due to peptide bond contacts, especially due to residues 190, 215 and 219. Since the electrostatics term in EFP is computed in a pairwise fashion, it is directly comparable to P-SAPT simulations. It can be seen that EFP predicts the electrostatic interaction energies to within 0.5 kcal/mol in comparison to the P-SAPT simulations. The discrepancies found in $\Delta(\Delta E_{elec})$ in the CYS191, CYS220 and disulfide fragments are due to the different fragmentation schemes employed in EFP and SAPT methods. Electrostatic interactions are the major contribution to the preferential binding energies, in some cases as high as 2 kcal/mol.

Fig 3.7 shows the contribution of polarization interactions to the $\Delta(\Delta E)$ term. For polarization interactions, it is prudent to compare the EFP results with the F-SAPT results, as both the methods obtain interaction energies in a 'fully interacting' system. Again, the qualitative prediction of BioEFP method is reasonable, however, the difference in interaction energies are less than 0.4 kcal/mol in most fragments. In many cases, EFP seems to overestimate the stabilization/destabilization due to binding energy differences, but the errors do not exceed 0.25 kcal/mol. Damping the polarization using screening functions does not affect the results by much, and hence it can be concluded that the polarization interactions do not contribute much to the difference in binding energies, as predicted by SAPT results as well.

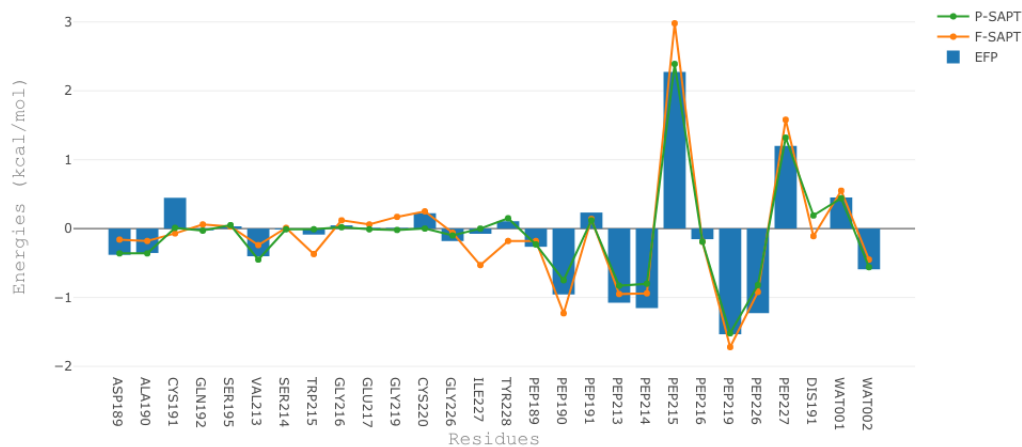


Fig. 3.6.: Contribution of electrostatic interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.

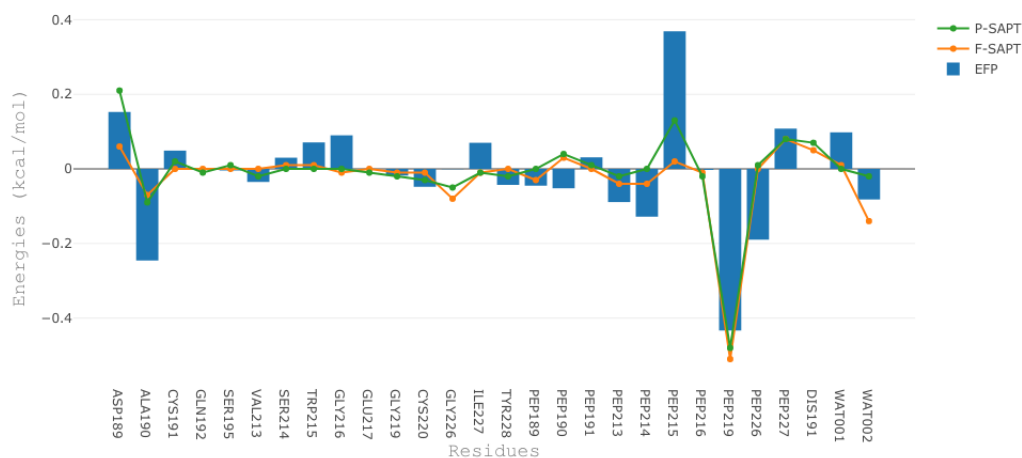


Fig. 3.7.: Contribution of polarization interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.

Fig 3.8 shows the contribution of dispersion interactions to the $\Delta(\Delta E)$ term. For dispersion interactions, we revert back to comparing the EFP results with the P-

SAPT results, as both the methods obtain interaction energies in a pairwise manner. The contribution of dispersion to the $\Delta(\Delta E)$ term is understandably low (less than 0.2 kcal/mol), as the substitution of Cl- with a Me- group does not significantly affect the $\pi - \pi$ interactions or CH- π interactions within the ligand-S1 pocket. This finding is in line with what has been observed in earlier as well [74].

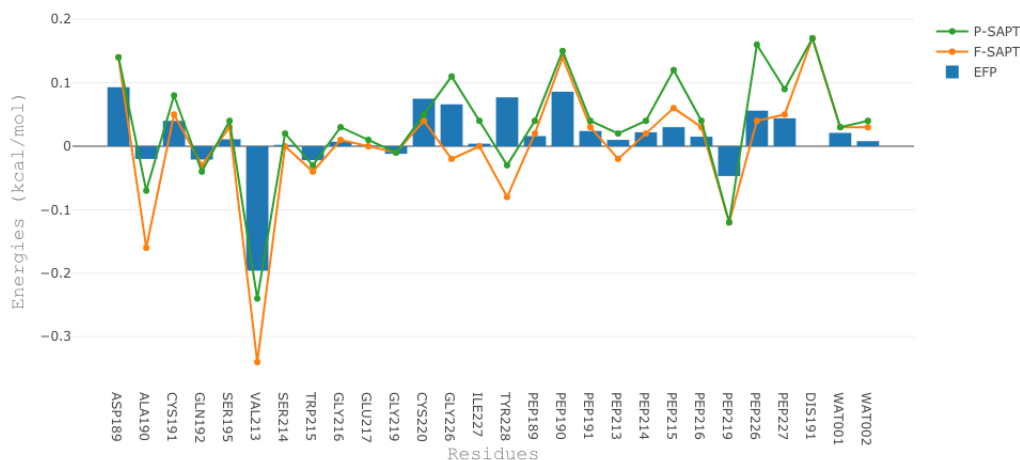


Fig. 3.8.: Contribution of dispersion interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.

Fig 3.9 shows the contribution of exchange repulsion interactions to the $\Delta(\Delta E)$ term. For the exchange-repulsion term, again we compare the EFP results with the P-SAPT results, as both the methods obtain interaction energies in a pairwise manner. Formally, the exchange-repulsion term decays in an exponential manner, and we can assume that only the closest residues that directly interact with the indole ring would contribute to this term. This is evident in the case of ALA190 and VAL213, which is captured by EFP as well as SAPT methods. As explained earlier, the discrepancy in DIS fragment interactions caused by fragmentation are captured in the two CYS fragments.

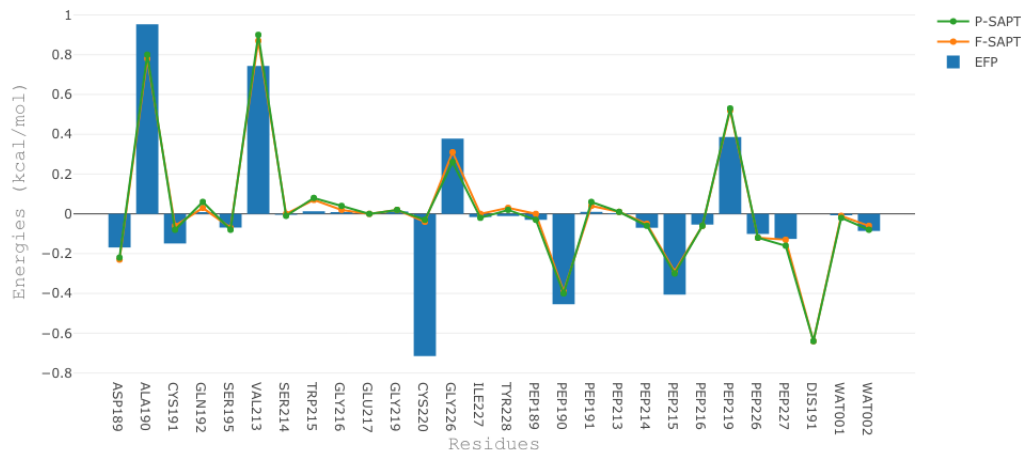


Fig. 3.9.: Contribution of exchange-repulsion interactions to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.

Fig 3.10 shows the sum of all the interaction energy components to the $\Delta(\Delta E)$ term. Fig 3.11 individual contributions of all the four terms that contribute to the $\Delta(\Delta E)$. A point to note here is that while the $\Delta(\Delta E_{total})$ for a single fragment could be negligible, but the contributions of individual interaction energy terms could be non-negligible. Case in point: The contribution of ALA190 and VAL213 residues to the $\Delta(\Delta E)$ is less than 0.5 kcal/mol, while the electrostatic contribution to these residues are closer to 1 kcal/mol, which are then countered by other terms.

Another question we are trying to answer here is the validity of using the S1 pocket as a representative model for simulating the ligand-protein interactions. Fig 3.12 shows the convergence of $\Delta(\Delta E)$ components as a function of distance between centroids of individual residues in the S1 binding pocket. As one can expect, electrostatic interaction term converges very slowly, as the formal decay of charge-dominant interactions as a function of distance is $1/r$. Fragments such as Peptide226 and Peptide227 contribute well over 1 kcal/mol to the $\Delta(\Delta E)$ term, even though they are

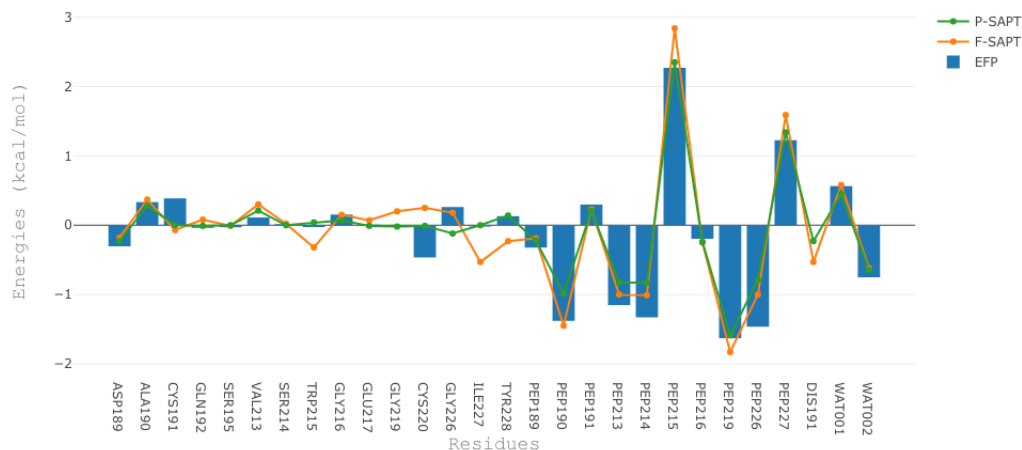


Fig. 3.10.: Contribution of total interaction energies to $\Delta(\Delta E)$ term in the small ligand - S1 pocket model.

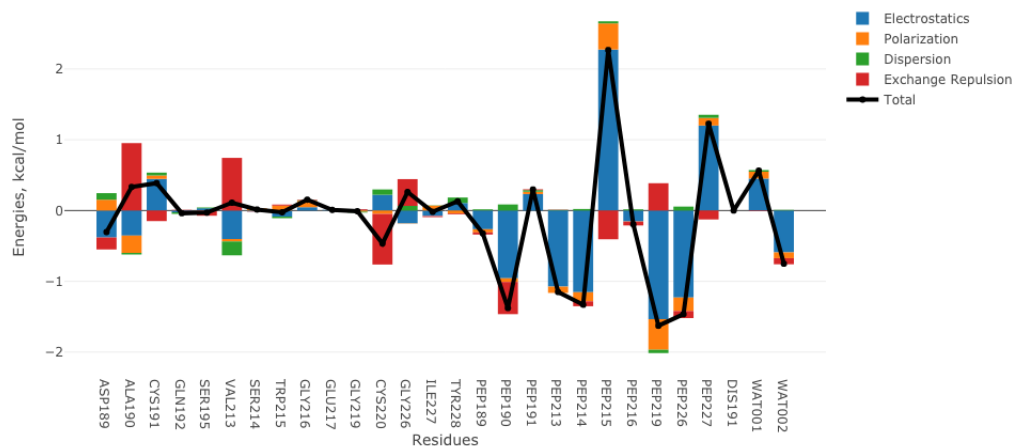


Fig. 3.11.: Total interaction energies and energy component contributions of individual fragments to $\Delta(\Delta E)$ term in the small ligand-S1 pocket model.

located 8-9 Å away from the ligand. This indicates that the ligand-S1 pocket may not be representative of all the significant interactions in the ligand-protein system.

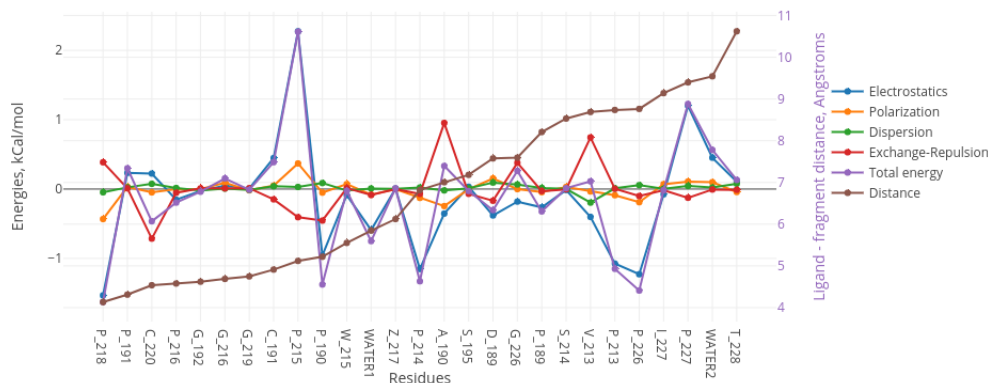


Fig. 3.12.: Pairwise contributions to $\Delta(\Delta E)$ term in small ligand-S1 pocket model as a function of distances between ligand and fragments.

To test this hypothesis, we simulated the entire protein-ligand system using BioEFP. Fig 3.13 shows individual fragment contributions to the $\Delta(\Delta E)$ term as a function of the separation between these fragments and the ligand. We can observe that the electrostatic contributions converge slowly, presenting a few interactions greater than 0.4 kcal/mol at distances beyond 1.5 nm. This is an indication that the substitution in the ligand is stabilized/destabilized by interactions with residues located well beyond the pocket.

Finally, we test the performance of BioEFP method in accurately predicting the preferential binding energies as a result of substitution. Table 3.1 lists the $\Delta\Delta G$ computed using various methods and systems. While the F-SAPT method predicts the total $\Delta\Delta G$ accurately to within 0.2 kcal/mol, P-SAPT underestimates the preferential Cl-ligand binding by 1.1 kcal/mol due to accumulation of errors. BioEFP overestimates the Cl-ligand binding by 1 kcal/mol for the smaller ligand-S1 pocket model, and this can be attributed to the lack of convergence in $\Delta\Delta G$ components for the smaller system. This is more evident in Fig. 3.14, where the presence of very strong stabilizing interactions can be noticed at around 1 nm from the ligand. These

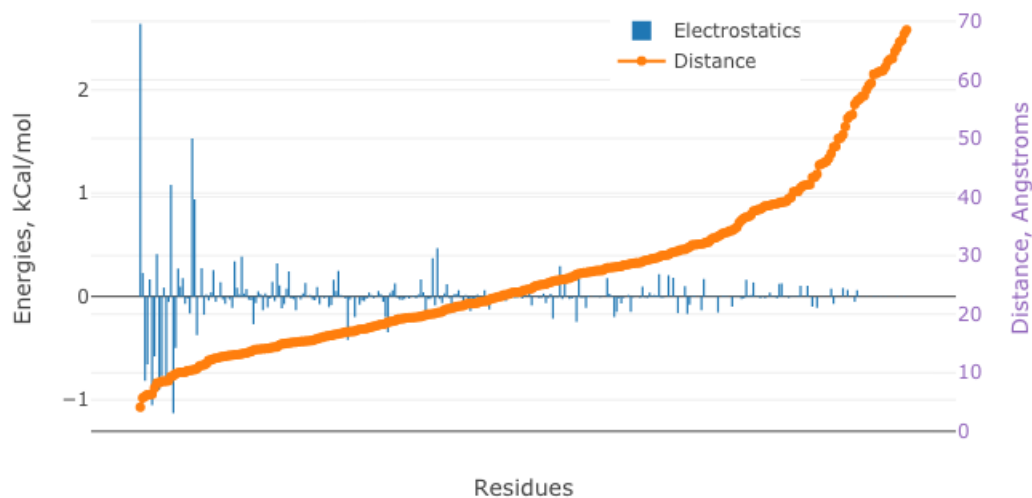


Fig. 3.13.: Convergence of $\Delta(\Delta E_{elec})$ term in the ligand - protein complex as a function of distance.

interactions are then countered by several destabilizing interactions beyond the 1 nm sphere, and the $\Delta\Delta G$ converges at around 4 nm from the ligand.

Table 3.1.: Difference in binding energies between Cl- and Me- ligands, computed using different methods.

Method	$\Delta\Delta E$ (kcal/mol)
F-SAPT (small)	-2.464
P-SAPT (small)	-1.208
EFP (small)	-3.314
EFP (large)	-2.436
Experiment	-2.3

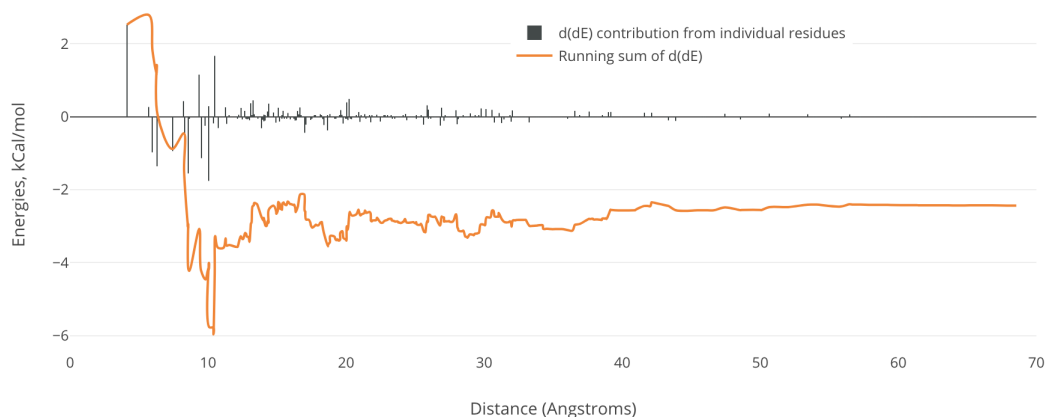


Fig. 3.14.: Convergence of $\Delta(\Delta E_{total})$ term in the ligand - protein complex as a function of distance. A running sum of the $\Delta(\Delta E_{total})$ term is plotted (orange line).

3.2.5 Conclusions

The performance of BioEFP as a tool for computing the intermolecular interaction energies and binding energy differences is probed. The results presented here are in good agreement with accurate first principles methods. Simulations of the ligand in the whole protein indicate that the electrostatic energy differences due to varying substituents do converge at distances far exceeding the size of the pocket previously used for modeling $\Delta(\Delta G)$ in this system. Performing simulation of the ligand with protein ensures that the interactions unaccounted for in the smaller system are now properly accounted for. BioEFP provides a viable option for performing such simulations at a much faster timeframe. Additional effects that need to be considered for obtaining rigorous binding energy differences are solvent effects and configurational sampling, which will be explored in future work.

3.3 EFPMD-MC

It is important to know the effect of cofactors on cocrystal formulations for active pharmaceutical ingredients. However, performing an exhaustive experimental screen-

ing is experimentally expensive. Thus, there is a need for a tool that is relatively accurate yet computationally cost effective for the virtual screening of drugs based on changes in functional group interaction energies in order to approximate free energy changes. Here, we describe the implementation of the Monte-Carlo method and perform some simple benchmarking studies.

3.3.1 Monte-Carlo Sampling

Here we describe the technical implementation of a Monte-Carlo integration over configuration space using the effective fragment potential (EFP) method as a description for a chemical system. Using the effective fragment potential method, is possible to calculate energetic properties of any substance of a system of N interacting EFP fragments:

$$E^{EFP} = E_{coul} + E_{ind} + E_{exch} + E_{disp} + (E_{CT}) \quad (3.7)$$

In order to explore the configurational space, random sampling of points are obtained through moving each EFP fragments from an initial configuration in succession:

$$x = x + \alpha \xi_x \quad (3.8)$$

$$y = y + \alpha \xi_y \quad (3.9)$$

$$z = z + \alpha \xi_z \quad (3.10)$$

$$a = a + \alpha \xi_a \quad (3.11)$$

$$b = b + \alpha \xi_b \quad (3.12)$$

$$c = c + \alpha \xi_c \quad (3.13)$$

where α is the maximum allowed displacement, and $\xi_x, \xi_y, \xi_z, \xi_a, \xi_b, \xi_c$ are random numbers between (-1) and 1. x, y , and z refer to the Cartesian coordinates for the center of mass of a fragment. a, b , and c refer the fragment's Euler angles. The change in energy of the system ΔE following the move is calculated. If $\Delta E > 0$, then the

move is evaluated against probability of $\exp(\Delta E/kT)$ where a random number ξ_4 between 0 and 1 is evaluated against $\exp(\Delta E/kT)$. If $\xi_4 < \exp(\Delta E/kT)$, then the system configuration is accepted despite $\Delta E > 0$. However if $\xi_4 > \exp(\Delta E/kT)$, then the move is rejected and the system is returned to its prior configuration.

3.3.2 Technical Implementation

The general EFP method has been implemented as a library API called libefp [56]. The libefp library and efpmd program are written in fully portable standard C language and parallelized using OpenMP. Here we describe some brief additional functions to efpmd program that serves as the integrator for a Monte-Carlo simulation with the EFP method.

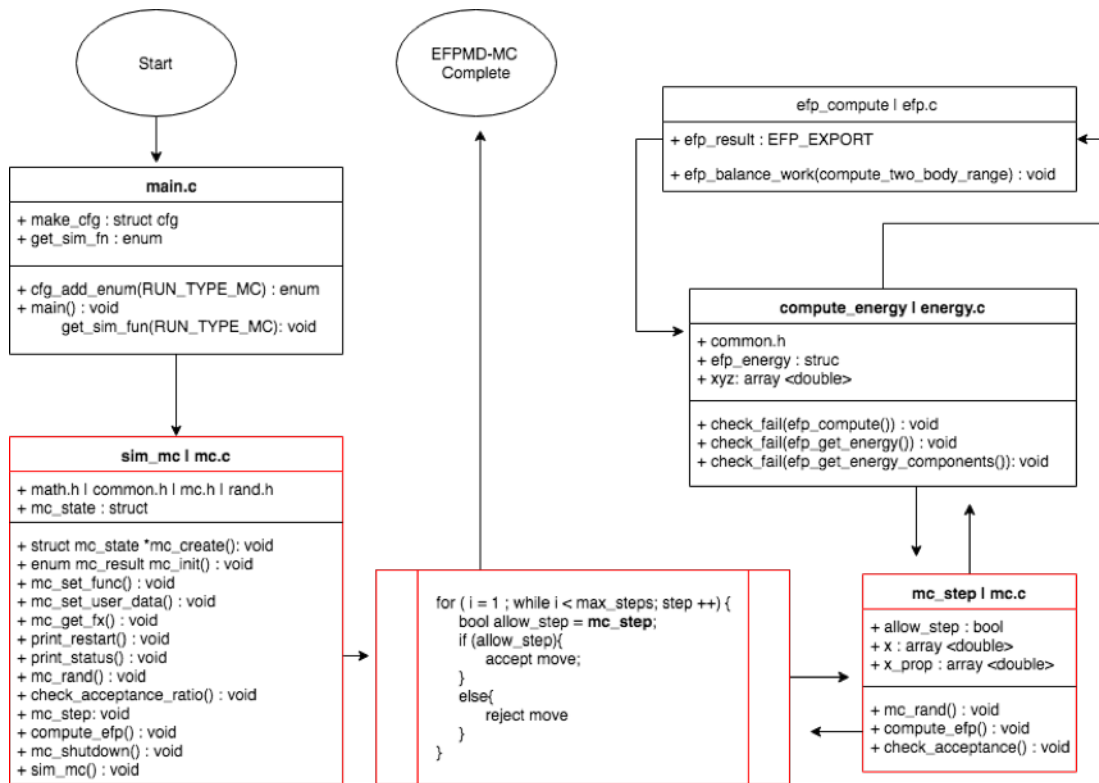


Fig. 3.15.: Monte-Carlo implementation scheme within *efpmd*

In order to perform Monte-Carlo using the EFP method, a new subroutine `sim_mc()` was introduced and linked to the `efpmd` program. The simulation configurations for performing a EFP-MC calculations, involve the `run_type` flag introduced as `'run_type mc'` and `max_step` and `temp` general parameters for the simulation. Similarly to other `run_types` in `efpmd`, `sim_mc()` is initialized in its own header file (`mc.h`) and declared in `mc.c`, both located in the `/efpmd/src/` directory. `sim_mc` includes/calls on other functions in the `efpmd` program through header files `math.h`, `common.h`, and `rand.h`. Running a EFP-MC related parameters that control the step size are:

- Max Displacement Step Size: `dismag_threshold` [default = 0.05]
- Update Displacement Step Size: `dismag_modifier` [default = 0.95]
- Frequency of Updating Displacement: `dismag_modify_steps` [default = 100]

System information is parsed and stored in struct `mc_state` in the void `mc_create`. It is within `mc_state` that the initial configuration for the original configuration is accessible. `Mc_state` also contains a dynamic array for proposed configuration `x_prop` that is updated at each step of the simulation. Enum `mc_init` is the function that allocates the memory for `mc_state`, along with initializing the Monte-Carlo step counter `'step'`, the initial accepted step `n_accept` and rejected steps `n_reject`.

Accessory functions such `mc_set_func()` and `mc_set_user_data()` are provided in order to transfer data structures and simulation configuration information populated within the initial `main()` function in `main.c` to data structures within `sim_mc()`. `Sim_mc` which is able to communicate with the library function `libefp` through subroutine `check_fail()` that passes the atomic coordinate and point charge coordinate information for each fragment.

When running a Monte-Carlo simulation, `sim_mc` calls on `mc_step()` multiple times in a while loop for the number of `max_steps` the user specifies. At each step, `mc_rand()` is called to randomize the center of mass (COM) of one fragment in the simulation. The energy of the state is obtained through `compute_efp()` that calls on `libefp` function `efp_compute()` through `compute_energy()`. With each Monte-Carlo step, evaluation of

the move is done through `check_acceptance()`. If the move is accepted, the proposed coordinates stored in struct `x_prop` will be copied to struct `x`. Else, another step is taken and the proposed configuration is evaluated.

3.3.3 S22 Dataset

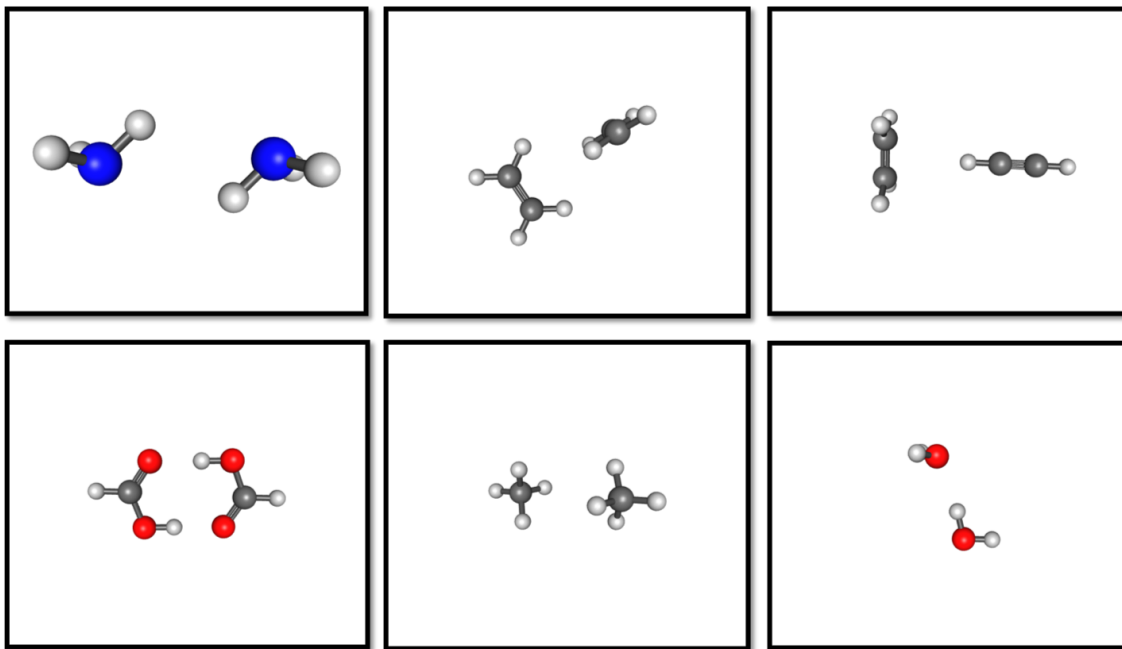


Fig. 3.16.: Selected S22 Dimer Complexes

In order to examine the robust of the Monte-Carlo code it was necessary to examine the ability of the program to sample different phase space of different systems. Previous efp parameters [51] generated for the S22 dataset [82] are readily available through the libefp package. The Coulomb part of these parameters was obtained with analytic Stone DMA, using HF/6-31+G(d)(42-44) and HF/6-31G(d) for nonaromatic and aromatic molecules, respectively. The rest of the potential, that is, static and dynamic polarizability tensors, wave function, Fock matrix, etc., were obtained with the 6-311++G(3df,2p) basis set.(44-46) To account for the short-range charge-penetration

effects, overlap-based electrostatic and dispersion screenings as well as Gaussian-like polarization screening were employed [51].

Finding the Potential Minima From an Optimized Geometry S22 is a data set of dimer complexes are divided into three subgroups: (i) hydrogen bonded complexes; (ii) complexes with predominant dispersion stabilization; (iii) mixed complexes in which electrostatic and dispersion contributions that are similar in magnitude. Of the 22 complexes, 6 of the dimers were optimized at the CCSD(T) level in cc-pVTZ and cc-pVQZ basis sets, and so were selected as initial configurations (see Fig. 3.16).

Each configuration served as the initial step, for an EFP-MC simulation with 10,000. Each simulation was run with a displacement maximum threshold of 0.05 and displacement modifier of 0.95 utilized ever 500 steps. After the simulation, the configuration with the lowest energy was obtained for each dimer and geometry optimized using efpmd. The geometry optimized structure following Monte-Carlo simulation (EFP MC_Opt) was then compared against a EFP geometry optimized structure (EFP Opt) and the initial CCSD(T) structure itself (CCSD Opt).

Table 3.2.: EFP Energies (kcal/mol) Obtained through Geometry Optimization or Monte-Carlo on S22 Equilibrium Geometries

Complex	EFP MC	EFP MC_Opt	EFP Opt	CCSD Opt
ammonia	-4.72	-5.37	-5.37	-3.17
ethene	-2.70	-3.16	-3.16	-1.51
ethene-ethyne	-1.92	-2.32	-2.32	-1.53
formic acid	-980.38	-16.65	-16.65	-18.61
methane	-0.92	-1.04	-1.04	-0.53
water	-5.90	-5.99	-5.99	-5.02

Using the Monte-Carlo method with EFP would likely find a non-equilibrium configuration close to one of the local EFP minima. Optimization of that non-equilibrium configuration, would result in finding this local EFP minimum. EFP-MC values for

the lowest energy minima are reported in 3.2 using a geometry optimized EFP-MC (EFP MC_Opt), EFP-MC (EFP MC), EFP optimized structure (EFP Opt), and CCSD(T) initial reference structure (CCSD Opt).

Finding Potential Minima From Non-equilibrium Geometries Rather than focus on finding a single minima/stationary point, it was important to see if the Monte-Carlo algorithm in conjunction with EFP would be able to explore and find local minima beyond the minima obtained through geometry optimization. Thus, the EFP parameters for 6 dimer complex from the S22 dataset were once more utilized. However, the initial configurations were not the obtained CCSD(T) geometry optimized structures, but were randomly placed twice the original intermolecular distance apart. These geometries were obtained from the S22 nonequilibrium geometries [83]. For each dimer complex, a Monte-Carlo simulation was run for 10,000 steps, displacement maximum threshold of 0.05 and displacement modifier of 0.95 utilized ever 500 steps. After the simulation, the configuration with the lowest energy was obtained for each dimer and geometry optimized using efpmd. The geometry optimized structure following Monte-Carlo simulation was then compared against a reference efpmd geometry optimized starting from the 10 angstrom apart dimer structure.

Table 3.3.: EFP Energies (kcal/mol) Obtained through Geometry Optimization or Monte-Carlo on S22 Nonequilibrium Geometries

Complex	EFP MC_Opt	EFP Opt	CCSD Opt
ammonia	-	-1.67	-0.36
ethene	-1.86	-3.15	-0.03
ethene-ethyne	-1.55	-1.54	-0.15
formic acid	-16.61	-16.61	-3.63
methane	-1.03	-1.03	-0.01
water	-7.07	-5.71	-0.96

In Table 3.3 we see that EFP using Monte-Carlo was able to obtain different minima (EFP MC_Opt vs. EFP Opt) rather than just falling into the original geometry optimized minima presented in 3.2 - implying that the system was able to overcome energy barriers and explore different potential energy minima. With Ammonia, we were not able to obtain a optimizable local minima using the same simulation with Monte-Carlo and so is not provided in this table. The ability to EFP-MC explore potential minima was encouraging and thus we attempted to find local stationary points on well established potential energy surfaces of a water dimer.

Water Dimer Local Minima Here, the water dimer configuration obtained from S22 nonequilibrium dimer configuration served as the initial step. The simulation was ran with 10,000 steps using efpmd with a displacement maximum threshold of 0.05 and displacement modifier of 0.95 utilized ever 500 steps with periodic conditions of 10 Angstrom. After the simulation, each Monte-Carlo step was then geometry optimized using efpmd. The initial Monte-Carlo obtained configurations are depicted in Fig 3.17 as energetic states distinguishable by intermolecular distance (without periodic boundary conditions applied). Although, the obtained energies are obtained with periodic boundary conditions, it was easier to distinguish individual states with respect to their coordinates without PBC in order to 'smear' the density of points and see energy groupings.

When looking at the obtained energy states (See Fig. 3.18 from the initial configurations (see Fig 3.17), we see a reduction in the number configurations and it is easy to ascertain that there are approximately 3 'lines' or minima of a water dimer in Fig. 3.18. From these configurations, the 3 final configurations found (see Fig. 3.19 correspond to the linear (I), cyclic (II), and bifurcated (III) CI potential minima by Matsoka et. al [84]. EFP energies and CI literature energies [84] are reported in Table 3.4. This is a promising result that EFP is able to minimally obtain previously cited local minimum obtained using CI methods.

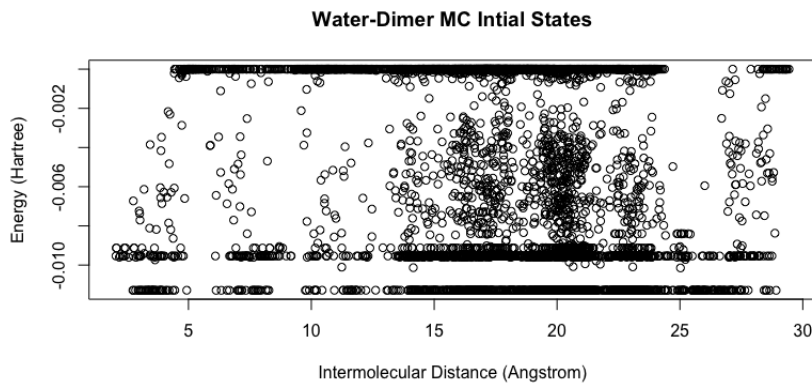


Fig. 3.17.: Initial Monte-Carlo Water Dimer Configurations

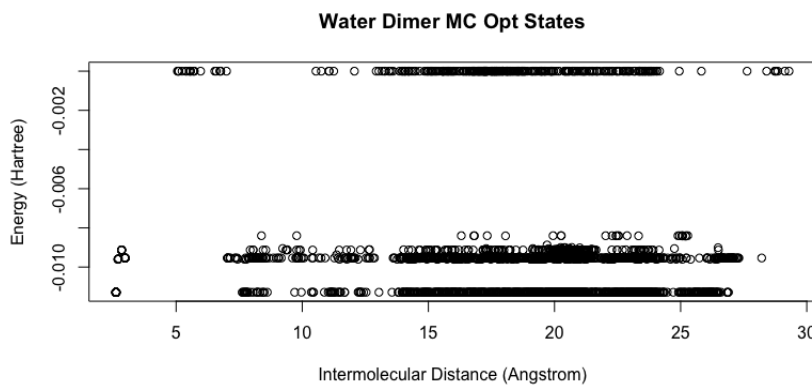


Fig. 3.18.: Geometry Optimized Monte-Carlo Water Dimer Configurations

Table 3.4.: EFP vs CI Reference Energies (kcal/mol) on Water Dimer Local Minima

Configuration	EFP	CI
Linear	-5.98	-5.6
Cyclic	-7.08	-4.9
Bifurcated	-5.72	-4.2

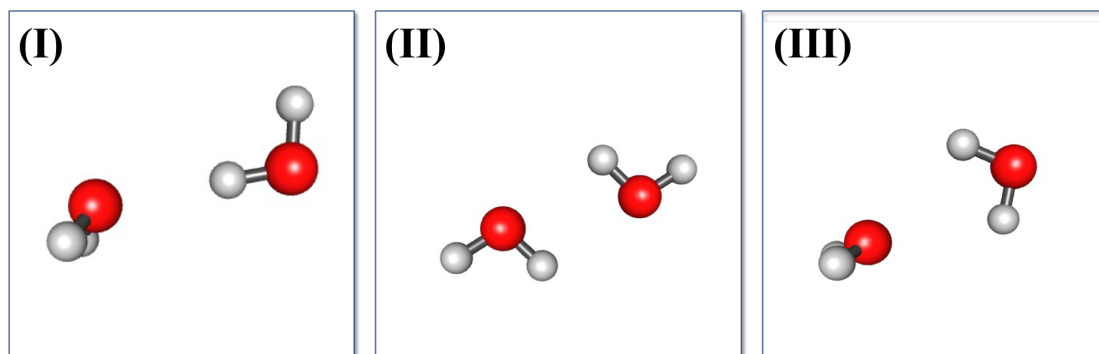


Fig. 3.19.: Local Minima Obtained From Monte-Carlo Optimized Water Dimer Configurations

3.3.4 Conclusions

The theoretical and technical implementation of the Monte-Carlo method in the libefp package is reported. Benchmark studies on the ability of EFP-MC to appropriately phase-space sample relative to both CCSD(T) geometry optimized equilibrium and non-equilibrium geometries on the S22 dataset are reported. The configurations obtained and presented here are in good agreement with those obtained using accurate first principles methods as demonstrated by finding the linear, cyclic, and bifurcated structures with comparable energetics. These results demonstrate EFP-MC as method for obtaining local and global minima on a potential energy surface of a molecular model. Thus, this work serves as a promising means to perform co-crystal screening with EFP-MC as a sophisticated alternative to docking methods.

4. SIMILARITY MEASURE

However, determining the robustness of a method against other methods is usually closer to the ‘downstream’ or final stages of a workbench simulation process investigating dynamical properties of a molecular system. The precursor to any sort of benchmarking or validation of a method on a particular model involves analysis of the system with respect to its varying parameters. With the case of our benchmarking and phase-space sampling studies, molecular configuration and orientation should be closely analyzed. Thus, for our large sets of data, this involves utilizing automated and robust structure comparison methods in order to assess the robustness of a set of configurations by assigning some sort of quantitative value through the use of a similarity metric. Ideally, a similarity measure should be able to:

- provide an index value that ranks structures/configurations of a molecular system based on their similarity with a minimal overlap - providing a high degree of resolution between clusters of structures.
- provide an intuitive visual interpretation.
- be robust, relevant and generalizable

The appropriate choice of method that measures similarity and dissimilarity (distance) measures then becomes of great importance in pattern analysis problems such as classification, clustering, and recognition. Over the last century there have been great deal of effort toward finding meaningful similarity and distance based measures in various fields. Subsequently, they’ve been applied in biology [85, 86], fingerprint analysis [87], image retrieval [88], etc. This is no different then in analysis of chemical structure classification [89].

When looking at biologically relevant similarity metrics utilized for proteins, methods generally fall into two classes: positional distance-based and contact-based. With regards to the first class (and the more popular), positional euclidean distance-based measures require super-positioning reference atoms (selection of appropriate super positioning is also not an easy task) in Cartesian space in order to minimize the distance between shared reference points/atoms. Typically, similarity in the super-imposed configurations is measured using Root Mean Square Deviation (RMSD):

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n d_i^2} \quad (4.1)$$

where d_i is the distance between two atoms in the i -pair of all atoms N for comparison. However, RMSD provides an average of the distances between pairs of atoms, and as such can become dominated by the most deviated fragments. Another cause for concern with RMSD is the internal symmetry of the system. With systems of high degrees of symmetry, it becomes difficult to determine unambiguous atomic-pairings between configurations as some atoms within the structure are topologically equivalent to each other. This issue continues to plague scientists studying protein similarity through alignment and RMSD similarity metrics [90] .

The second class, contact-based measures serve as an alternative to avoiding super-positioning atoms. Contact-based measures are determined by overall differences between the distribution of pairwise distances from one configuration to the next, rather than distinguishing between structures by averaging the pairwise distances between the configurations. When utilizing contact-measure methods, the general protocol, as outlined by Abagyan and Totrov [91] is to assign a contact area difference (CAD) number as a similarity ranking measure to evaluate protein structures. In method, they determine the "contact strength" of two amino acid residues i and j within a protein as the overlap of van der Waals surface area of residue atoms A_{ij} . This is done for all pairs of residues in the protein and the stored as elements in matrix

$\{A\}$. When comparing contact matrices for reference structure R to trial structure T , the elements of the difference matrix between R and T will be:

$$\Delta A_{ij}^{RT} = (A_{ij}^R - A_{ij}^T) \quad (4.2)$$

Thus, non-zero elements in ΔA^{RT} will provide information about differences between fragment R and T in regards to specific residue pairs i - j . This representation of contact differences between fragment R and T can then be represented as a single CAD number of the total unnormalized contact errors as :

$$\Delta A = \sum_{i,j} |(A_{ij}^R - A_{ij}^T)| \quad (4.3)$$

However, a variant to contact-based differences is Cosine-Similarity that is popular for document similarity in text analysis. Like CAD similarity measure, Cosine-Similarity measure factors in non-zero matches between the trials A^R and A^T measures the similarity between the inner product space of two vectors by determining the angle between the two vectors:

$$\text{Cosine-Similarity} = \frac{A^R \cdot A^T}{||A^R|| ||A^T||} \quad (4.4)$$

where A_R and A_T correspond to a reference vector and trial vector, $||A||$ the euclidean norm of vector $A = (a_1^R, a_2^R, \dots, a_i^R)$, $||A_T||$ the euclidean norm of vector $A_T = (a_1^T, a_2^T, \dots, a_j^T)$. Thus, as cosine-similarity computes the angle between vectors A_R and A_T indicating whether the vectors are alike (cosine-distance = 1) or dissimilar (cosine-distance = 0). Thus, the closer the cosine value to 1, the smaller the cosine angle between two vectors, and the greater the match between vectors. Normalization of the Cosine-Similarity values:

$$\text{Cosine-Distance} = 1 - 2 \cos^{-1} \left(\frac{A^R \cdot A^T}{||A^R|| ||A^T||} \right) \quad (4.5)$$

provides the angular similarity or Cosine-Distance functional between vectors as a distance metric between vectors that provides a more intuitive ordering of similarity from structure to structure. Thus, in the search for an appropriate similarity measure

to analyze and categorize similar and dissimilar structures in an automated fashion, we have chosen to examine RMSD and Cosine-similarity distance metrics on two representative methods of chemical configurations: Cartesian- based distance matrix and Pairwise Radial Distribution Functional (PRDF)- based distance matrix.

4.1 Fingerprinting

However, the ability of similarity measure to capture the differences from structure to structure is also affected by the degrees of freedom of the chemical representation. For larger subsystems, these degrees of freedom are reduced into euclidean distance measures between arbitrarily designated center of masses. As such, we attempt to represent molecular structures as unique identifiers or 'fingerprints' using an intramolecular distance matrix representation and also contact-based matrix representation utilizing pairwise radial distribution functions. Using both representation, we test the ability of a general RMSD similarity measure and cosine-distance measure to rank 'fingerprints' in a visual representation that is intuitively interpretable. We also present the technical python implementation for said conversion of chemically relevant Cartesian-space data structures to redundant internal coordinates as a distance-matrix representation and a pairwise radial distribution functional representation.

4.1.1 Distance-Matrix

First to remove dependency on superpositions, the chemical representation for each atom within a system is converted from an array of points in cartesian space, to an internal coordinate presentation known as a 'distance-matrix' that provides a description of each atom in terms of its atomic type, bond length, angle, and dihedral angle. Rather than performing analysis of two structures using a typical 'atom x y z' data structure, molecules are represented in a linked list made up of a linear series of nodes and stored in an abstract data type known as a 'tree' that simulates a hierarchical tree structure with a root value and 'branches' or subtrees of children

with a main or parent node. Projecting each atom as nodes in Cartesian space, intramolecular bond distances and angles are computed in an iterative fashion to obtain a set of redundant internal coordinates and stored as a distance matrix. Using a set of rule-based conditionals, atoms are considered covalently bonded depending on intramolecular bond distances between pairs of atoms forming the basis for the molecules fingerprint-like ‘bonding tree’. Thus, using a brute force method, it is possible to create a representative ‘bonding graph’ matrix through computing all possible permutations of different variations of ‘bonding trees’ using different atoms as the parent node in a brute force method.

Structural Isomerism Using this bonding graph it is then possible to compare potential bonding trees of molecule A to molecule B and classify if molecule A is a structural isomer of molecule B if at least one of the bonding tree of molecule B is within a similarity threshold using RMSD with a user specified tolerance to the bonding tree of molecule A. This is done by converting the binary tree to a matrix of redundant bond distances between atoms, ranking the similarity between those matrices corresponding to bonding tree B and bonding tree A, and determining whether there is structural isomerism between fragments B within a certain threshold to the matrix for fragment A.

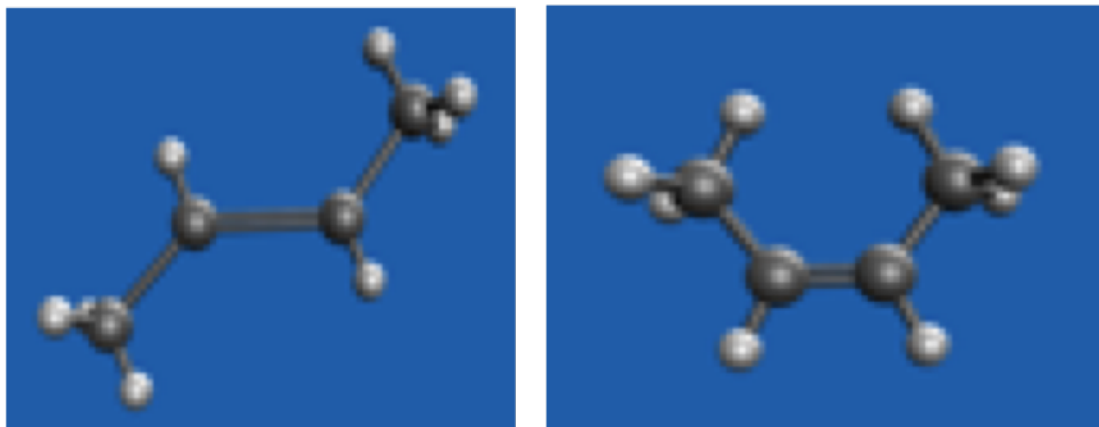


Fig. 4.1.: Example of Isomers

Cis-Trans Isomerism Following identification of an acceptable bonding tree to molecule A to molecule B, it is possible to determine cis-trans isomerizations using those bonding trees to compare angles to neighboring nodes for each pair of adjacent nodes. This is done by aligning the backbone of the nodes and building a plane perpendicular to the backbone. The adjacent nodes to the backbone of the modules are then rotated to minimize the distance between all nodes and the plane. If distances between nodes of fragment A to fragment B are within a given threshold, molecule A and B are considered optical isomers.

Technical Implementation Firstly, we convert the pdb file to xyz. Then, we use xyz coordinates to query for EFPdB. We then proceed with getting the fragment coordinates for the formula:

```
frag_id: 6
O  0.0 0.1191094785 0.0
H -1.422305967 -0.9451766865 0.0
H  1.422305967 -0.9451766865 0.0
```

We then proceed with checking if each fragment is a configurational isomer. Before the scanning of the atomic coordinates is started, comparisons between the number of different atom types in the reference coordinate file and the trial coordinate file are performed to make sure they are the same.

The most critical part of our program is that we built a script to convert xyz file to z matrix, which will tell us very important information about the relationships between the atoms, in another words, the bonds, to build the bonding trees. The implementation that borrows parts of the code from [92] converts xyz to distance matrix is provided here:

https://github.com/jialincheoh/iSpiEFP_Database_Search_Engine/blob/master/XYZ-to-ZMAT.py

To build the bonding trees on the input file and the database molecules, we compute the distances and unit vectors between all pairs of atoms. We also compute the angles between all triplets of atoms and then proceed with determining which atoms are covalently bonded based on the bonding criteria to come up with the bonding pattern. We then build bonding trees, which are the graphs, for the input and database molecules. The following scripts to generate bonds and angles are provided at:

```
https://github.com/jialincheoh/iSpiEFP\_Database\_Search\_Engine/blob/  
master/bonds.py
```

and

```
https://github.com/jialincheoh/iSpiEFP\_Database\_Search\_Engine/blob/  
master/angles.py
```

Afterwards, we perform permutations of the bonding graph of the input molecule to match the graph of the database molecule. Exact match of the two graphs in terms of having the same number of same atom types separates out the structural isomers. The main script that we use to handle structural isomers is found here:

```
https://github.com/jialincheoh/iSpiEFP\_Database\_Search\_Engine/blob/  
master/structural\_isomers.py
```

If the monomers pass the criteria for structural isomerism the representation is then passed to the script that we use to handle stereoisomers:

```
https://github.com/jialincheoh/iSpiEFP\_Database\_Search\_Engine/blob/  
master/isomers.py
```

This script allows us to distinguish between optical isomers, by determining the closest distance between two identical nodes on Reference Bonding Trees and Trial Reference Trees. Briefly, a hidden plane perpendicular to a line connecting the two

nodes is determined. We then proceed with rotating both the Reference and Trial bonding tree to minimize the distances between all nodes and the hidden plane. Afterwards, distances of the atoms from the Reference and Trial Tree to the hidden plane are computed and averaged. If the differences in the given distances are within a given threshold, the input and database molecules are optical isomers.

Using these scripts we can then distinguish similarity between bonding trees and then determine structural by comparing pairwise intramolecular distances and stereoisomerism by comparing angles to neighbouring nodes for each pair of adjacent nodes.

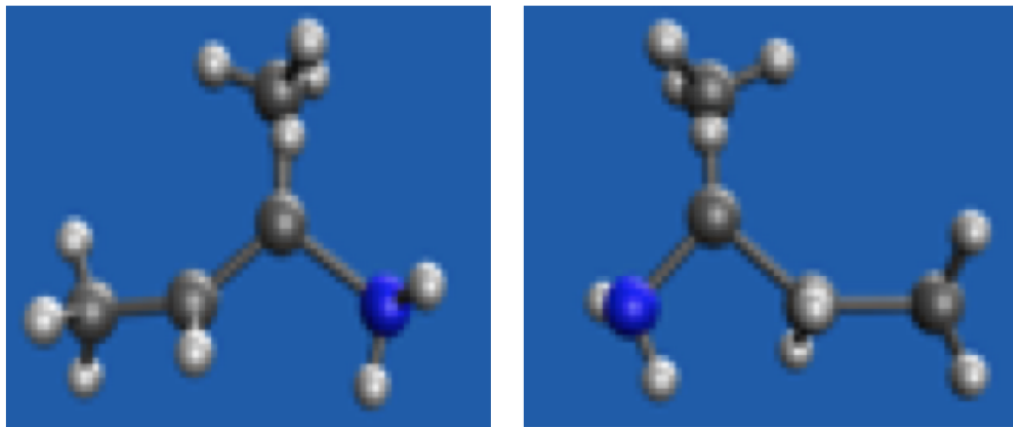


Fig. 4.2.: Optical Isomers

4.1.2 PRDF

One method for a contact-based matrix representation is to utilize a pairwise radial distribution function (PRDF). In this case the PRDF fingerprint for a molecular system is obtained by calculating pairwise radial atomic distribution distances that serve as structural signatures. Previous methods and implementations have already been utilized for material cartography to represent crystal structure subunits [93]. Using this method for chemical fingerprint, it has been demonstrated to be able to (i) query large databases of materials using similarity measures, (ii) map the connectivity

of materials space (i.e., as a materials cartograms) for identifying regions with unique trends/properties. In this implementation, the molecular fingerprint of a system is represented as a concatenated set of histograms detailing the distribution of atoms within a given space. The goal of this implementation is to provide an accurate method for chemical search queries for a EFP parameter database to be described in Chapter 5. Here, we detail a python implementation for the derivation of a PRDF structural fingerprint for chemical system as defined in Cartesian space.

Technical Implementation From data science perspective, these molecules are just data structures represented as 4-dimensional arrays with inputs stored within a text file. Rows refer to instances of atoms with the floating types of the Cartesian coordinates in the x, y, and z direction. Using this type of representation, a distance matrix can be computed in a pairwise fashion between atoms within the system. Thus, for a H₂O and NH₃ system shown in Fig. 4.3, a data structure would need to be initialized as a $N \times N$ array, where N represents the number of atoms. Each element within that array then will contain a list of pairwise distances specific to that particular atom-atom type.

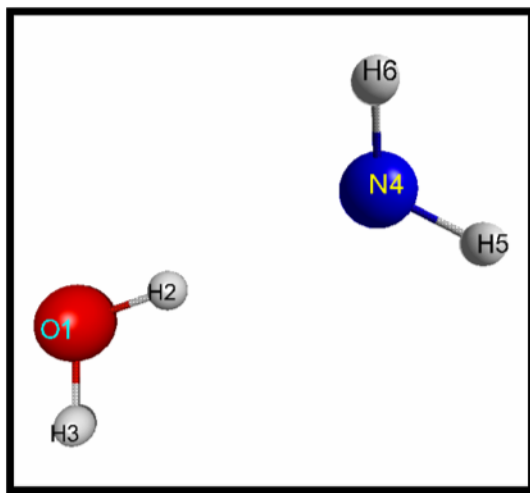


Fig. 4.3.: Water and Ammonia Dimer with Atom Labels

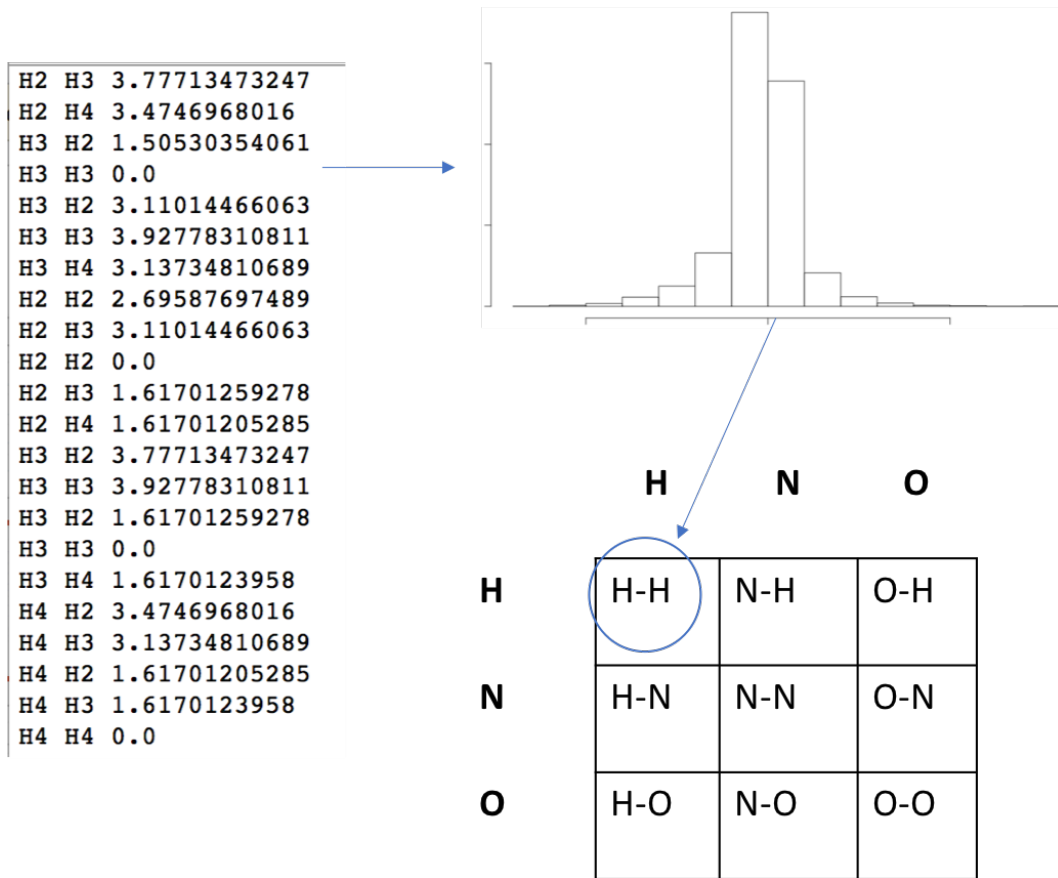


Fig. 4.4.: Sample H-H Intermolecular Conversation to Histogram In Data Structure Array

In our current implementation, a distance matrix is computed and used to obtain a diagonal - the greatest distance between two atoms. This diagonal is used as a threshold for normalizing pairwise distances and computing the particular density for each atom type. Then, a histogram of pairwise distances for two specific types of atoms are obtained iteratively for all elements in the array. In a general sense, each element in the array is a distribution described by:

$$F_{AB}(R) = \sum_{A_i} \sum_{B_j} \frac{R_{ij}}{4\pi R_{ij}^2 (N_A N_B / V_d)} \quad (4.6)$$

where i iterates over all atoms N_A of type A within the molecular system and j runs over all atoms N_B of type B . R_{ij} refers to the interatomic distance between atoms i and j and V_d volume of the molecular space. F_{AB} becomes a list of pairwise distances of type A - B . This list of pairwise distances then is accumulated into a histogram of bin size 0.05 \AA .

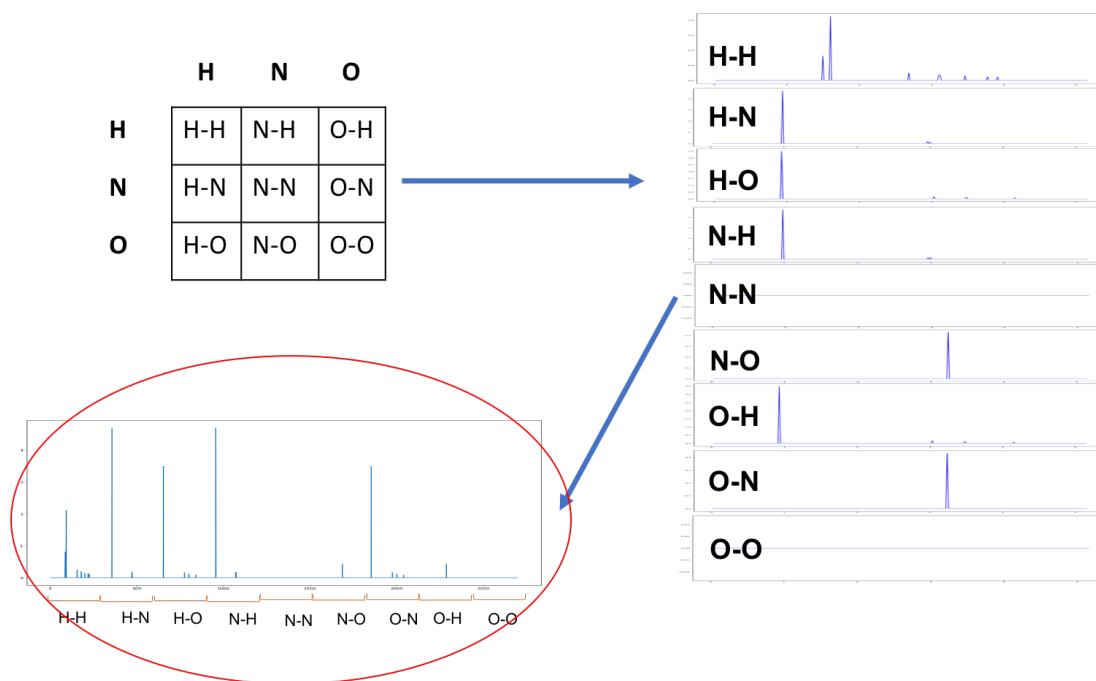


Fig. 4.5.: Conversion of Histograms to 'Molecular Fingerprint'

Once all of the histograms are obtained for each element in the array they are concatenated linearly to form a 2D dimensional array representing interatomic distances between pairs of atomtype A and B and the distribution of those distances (see Fig 4.5 and Fig 4.6). It should be noted that it is not possible to interconvert between atomic cartesian, distance matrix, and PRDF representations. This is due to the loss of information as one transforms the data from one type to the next.

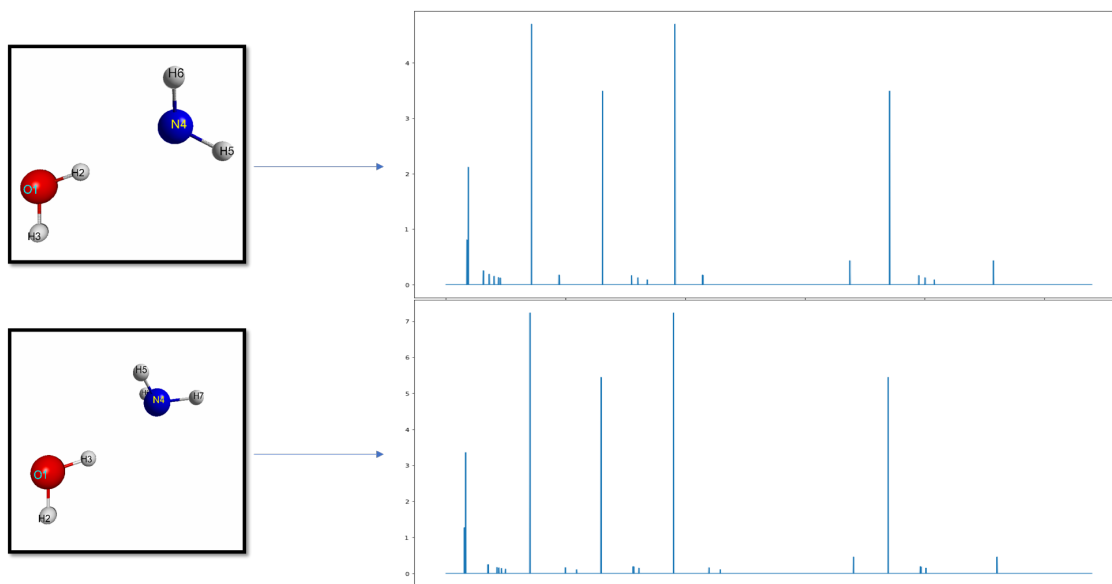


Fig. 4.6.: Comparison of the Molecular Fingerprints of Configurations of Ammonia and Water

4.2 Lysine Monomer

Here, we attempt to examine the ability of RMSD and Consine-Distance measures to distinguish similar and dissimilar configurations of a molecule within a biologically relevant context - amino-acid residue lysine. This amino acid fragment was chosen for its high degree of freedom resulting from its long alkyl side chain. Our dataset for the lysine monome contains 145 configurations extracted from ten different crystal structures (1NWA, 1OT9, 2QI7, 3PYP, 2WUR, 3E5T, 4GF6, 4Q9W, 3ENI, 3EOJ) for the Photoactive Yellow Protein (PYP) bacterial photoreceptor, Green Fluoresent Protein (GFP), and the Fenna-Matthews-Olsen (FMO) protein.

Each lysine monomer in the cartesian format/distance matrix representation then sequentially served as a reference by which RMSD and Cosine-Distance values were computed for against all other monomer-configurations in an iterative fashion (See Fig. 4.7(A-B)). This process was performed again for the configurations in the PRDF representation (See Fig. 4.7(C-D)). One can interpret Fig. 4.7 as a scatterplot of

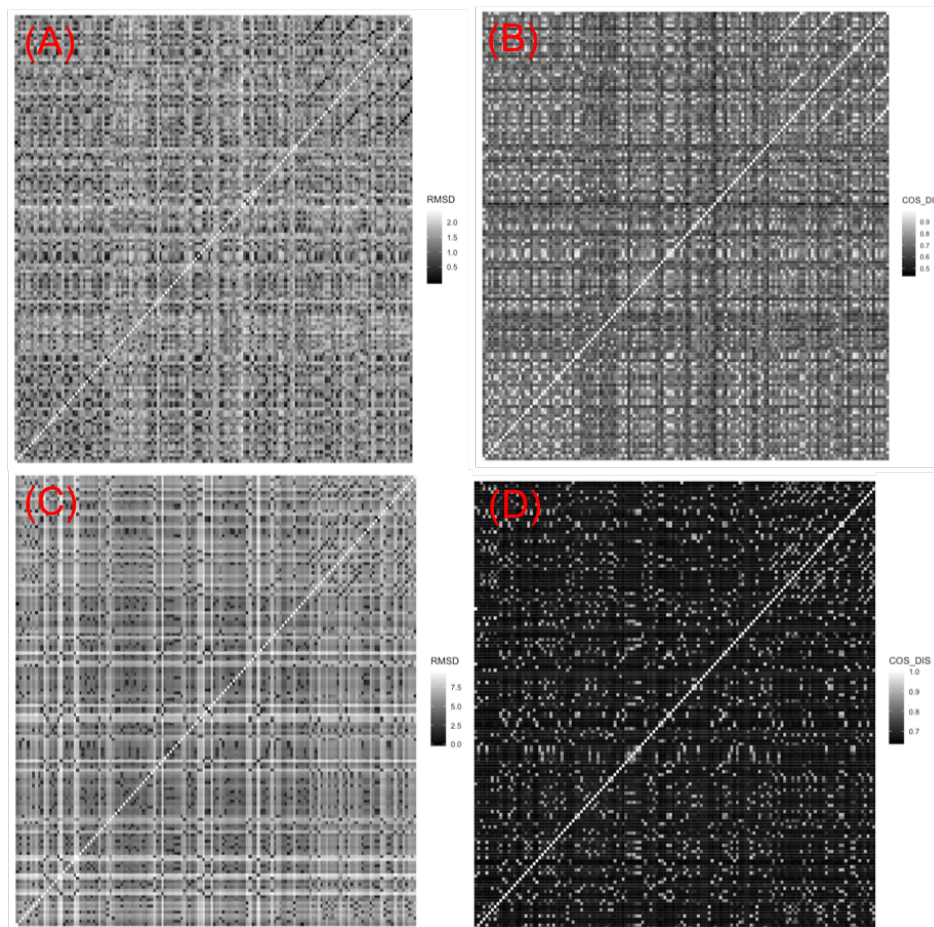


Fig. 4.7.: Similarity Comparison of Molecular Representations of Lysine Residues; A) RMSD similarity comparison using Cartesian Representation; B) RMSD similarity comparison using PRDF Representation; C) Cosine-Distance similarity comparison using Cartesian Representation; D) Cosine-Distance similarity comparison using PRDF Representation;

RMSD or Cosine-Similarity values between a lysine monomer against another lysine monomer, whose color along the black-white gradient scale indicates the degree of similarity between the two monomers. For plots demonstrating RMSD between lysine monomers (Fig. 4.7 A and C), an RMSD closer to ‘black’ is indicative of a similar structure as that value corresponds to zero. However, for the plots demonstrating Cosine-Distance values, points closer to ‘white’ are indicative of a similar structure as

Table 4.1.: LYS Similarity Average Measures

Representation	Measure	Min	Max	Median	Mean	SD
Cartesian	RMSD	0.00	2.40	1.41	1.36	0.41
Cartesian	Cos-Dis	0.45	1.00	0.70	0.71	0.09
PRDF	RMSD	0.00	9.57	5.57	5.55	1.64
PRDF	Cos-Dis	0.035	1.00	0.08	0.14	0.17

that value corresponds to 1. This is because Cos-Similarity measures the similarity between the structures as the inner product space of two vectors (in N Dimensional Space). Typically, similarity values range from -1 indicating ‘exact opposite’, or 1 meaning ‘exactly the same’. However, PRDF has values only in positive space so the range of values is from 0-1 (Dissimilar to Exact). In Fig. 4.7, we can see that the resolution between the configuration landscape of the lysine monomers provided by the PRDF fingerprint (Fig. 4.7 (C-D)) is higher than that of the Cartesian fingerprint (Fig 4.7 (A-B)) using both RMSD or Cosine-Distance metrics.

The average higher range (Average Max-Min) demonstrated by using both Cos-Distance and RMSD metrics with the PRDF representations (See Table 4.1) indicates PRDF might be able to distinguish on a broader range. This is confirmed looking at the qualitative color gradient of PRDF values versus RMSD as individual squares are easily distinguishable as similar as one scans across a row or column on either PRDF plot. Based off Fig. 4.7, Cosine Distance gives similar statistics on the distribution to RMSD, only normalized with respect to the particle density.

4.3 Random Amino Acid RMSD Cosine-Distance Measures

One limitation of RMSD, is that it is only able to distinguish against configurational isomers. In theory, because the PRDF is a reduction of 3D atomic coordinate to a 1D space that represents the pairwise distribution of indistinguishable atomic

centroids, it is possible to compare the molecular fingerprint as a whole, rather than element between two vectors. Thus, we wanted to examine the ability RMSD and Cosine Distance to detect similarity between molecules of different chemical compositions and configurations using the PRDF representation. We attempted to obtain RMSD similarity metrics using PRDF representations of 100 randomly selected amino acid residues extracted from ten different crystal structures (1NWA, 1OT9, 2QI7, 3PYP, 2WUR, 3E5T, 4GF6, 4Q9W, 3ENI, 3EOJ) for the Photoactive Yellow Protein (PYP) bacterial photoreceptor, Green Fluorescent Protein (GFP), and the Fenna-Matthews-Olsen (FMO) protein.

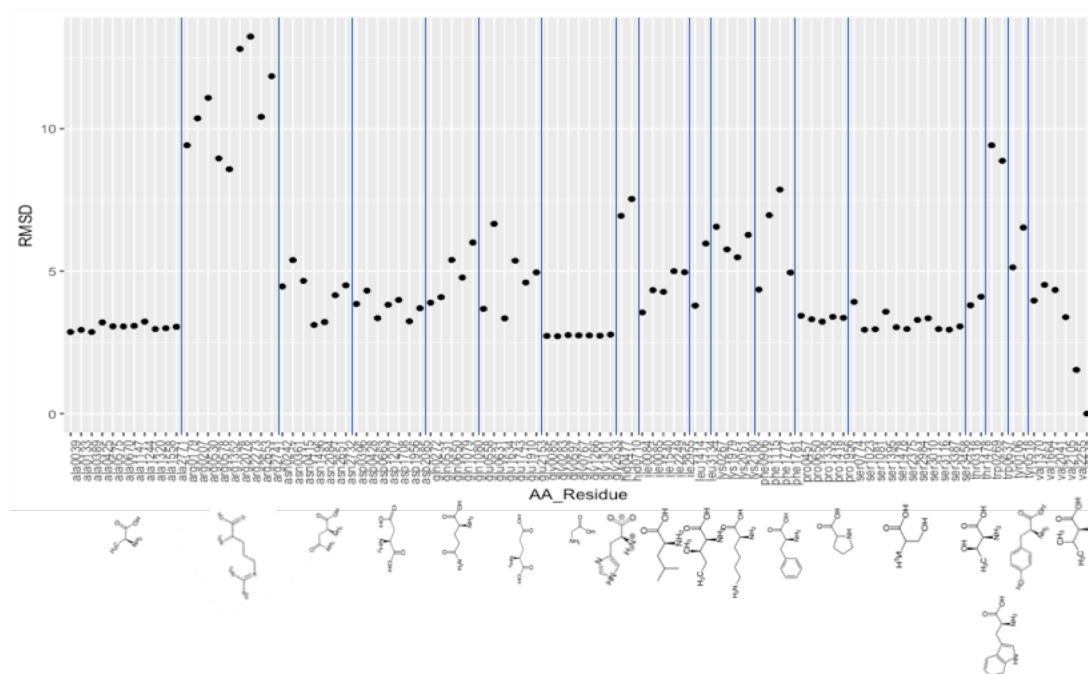


Fig. 4.8.: RMSD of Random Amino Acid Residues Versus Valine3947

Ninety-nine of those random structures were compared using RMSD to amino acid Valine3047 and reported in Fig. 4.9. As expected, we can see the amino acids with larger sidechains (Arg, Lys, and Trp) provide the highest RMSD values greater than 5 Angstrom of when compared to the reference Valine3947 molecule. Amino acids of the same type/size/shape provide similar RMSD values when compared to

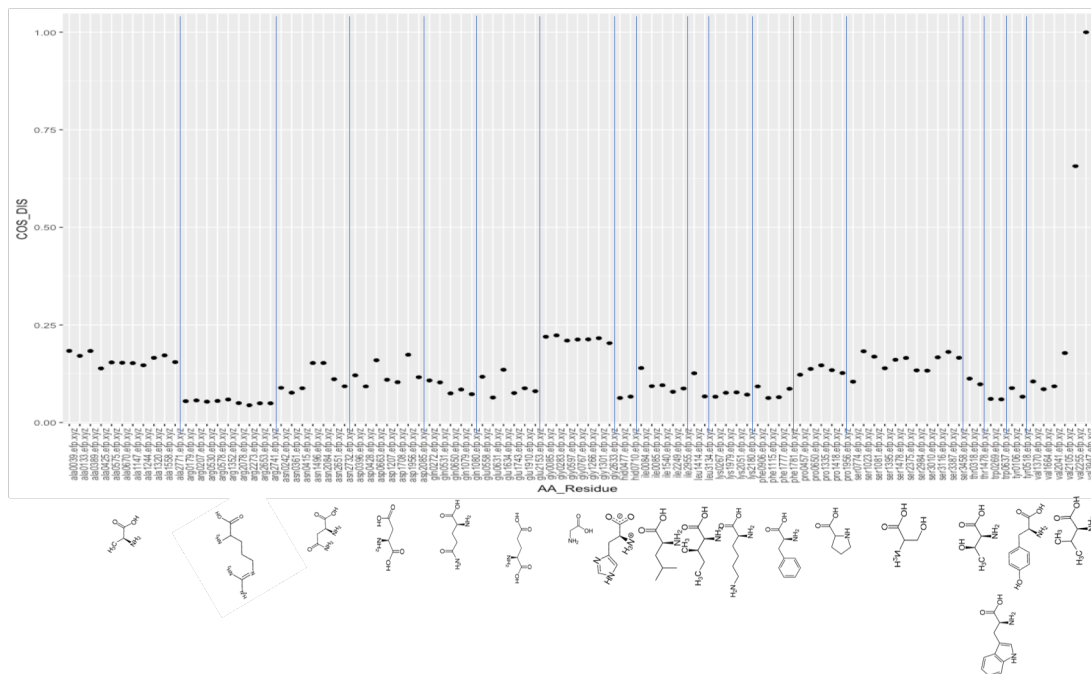


Fig. 4.9.: Cosine Distance of Random Amino Acid Residues Versus Valine3947

Valine 3947. Specifically, alanine, glycine, and proline behave uniformly across with RMSD values of 3 Angstroms. When we examine cosine-distance values on the same subset of configurations represented in the PRDF-based representation, we see that the measure is able to provide only one structure val2255 with a Cosine Distance value of greater than 0.5.

4.4 Conclusions

Here we present two methods for configurational chemical representations: Cartesian-based Distance Matrix and Pairwise Radial Distribution Functional-Based. On a dataset of 145 Lysine Configurations, we were able to see that PRDF provided higher resolution and distinguishability between configurations. Also, on a dataset of 100 randomly selected Amino-Acid configurations, the PRDF representation was able to distinguish between molecules of different chemical formulas and configurations. This implies that PRDF coupled with either RMSD or Cosine-Similarity

metric will be able to provide a relevant quantitative index assignment to chemical configurations that is robust and generalizable through a quantitative and qualitative visual interpretation.

We see a similar trend in Fig. 4.9. However, it seems as if the threshold of Cosine-Distance between structures of varying chemical composition and size is greater than 0.50 - something not possible utilizing similarity metrics such as RMSD on a cartesian-based representation.

5. TOOLS

Here, we present the development of EFPdB and iSpiEFP as tools enabling access to the efp method for the computational community. iSpiEFP provides a more streamlined route for data parsing simulation configurations and molecular geometry representation offered in a 'point-in-click' window. EFPdB, offers automation through the use of iSpiEFP, at obtaining and selecting similar/representative EFP fragment types.

5.1 EFPdB

The database serves as a repository of standard EFP fragment used for solvating systems of interests and standard amino-acid fragments using different fragmentation schemes [29]. Each instance within the database will be a EFP fragment with features necessarily to calculate EFP energy components. These features correspond to each EFP energy component.

Electrostatic Interactions Parameters related to EFP electrostatic interactions are stored as positions of distributed multipoles located at fragment atomic centers and covalent bond midpoints in cartesian space. Thus, each instance in the database contains data regarding a unique identifier for each distributed multipole on the molecular fragment, it's spatial orientation in XYZ format, atomic mass and estimated atomic charge in the following format:

```
COORDINATES (BOHR)
A01N1      -2.2226017038  -1.9159829515   0.0182086197  14.0030700  7.0
A02C2      -0.0280802012  -3.2801178867  -0.0192540569  12.0000000  6.0
A03H3      -0.2553060522  -5.3063424470  -0.0939411006   1.0078250  1.0
```

A04C4	2.2590197868	-2.1542088239	0.0242778590	12.00000000	6.0
A05H5	3.9781932028	-3.2403873432	-0.0070582589	1.0078250	1.0
A06C6	2.4278424431	0.5829159519	0.0284915022	12.00000000	6.0
A0707	4.3722507626	1.8211330344	-0.0997943006	15.9949100	8.0
A08N8	0.0720292042	1.7826039205	0.1583106263	14.0030700	7.0
A09H9	0.0849581956	3.6946891232	0.0627495002	1.0078250	1.0
A10C10	-2.2979131233	0.6950513523	0.0299271627	12.00000000	6.0
A11011	-4.2546425969	1.8941419493	-0.0847426590	15.9949100	8.0
A12H12	-3.9040343067	-2.7697787998	-0.2410336476	1.0078250	1.0
B021	-1.1253409525	-2.5980504191	-0.0005227186	0.00000000	0.0
B032	-0.1416931267	-4.2932301669	-0.0565975787	0.00000000	0.0
B042	1.1154697928	-2.7171633553	0.0025119011	0.00000000	0.0
B054	3.1186064948	-2.6972980835	0.0086098000	0.00000000	0.0
B064	2.3434311150	-0.7856464360	0.0263846806	0.00000000	0.0
B076	3.4000466029	1.2020244931	-0.0356513992	0.00000000	0.0
B086	1.2499358237	1.1827599362	0.0934010643	0.00000000	0.0
B098	0.0784936999	2.7386465218	0.1105300633	0.00000000	0.0
B0101	-2.2602574135	-0.6104657996	0.0240678912	0.00000000	0.0
B0108	-1.1129419595	1.2388276364	0.0941188945	0.00000000	0.0
B01110	-3.2762778601	1.2945966508	-0.0274077482	0.00000000	0.0
B0121	-3.0633180052	-2.3428808757	-0.1114125140	0.00000000	0.0

Related electrostatic parameters for monopoles, dipoles, quadrupoles, and octupoles are stored in a separate section as described in the Table 5.1:

In a system with multiple fragments, electrostatic interactions between fragments can be obtained using simple classical interactions between the aforementioned distributed multipolar interactions. i.e. point charges, dipoles, quadrupoles and octupoles interaction with those on other fragments to obtain the total electrostatic energy component.

Table 5.1.: Sample EFP Electrostatic Parameters

Electrostatic DM Parameter	Description	Ex:
Monopole	AtomID Electron_Charge Nuclear_Charge BondID Electron_Charge Nuclear_Charge	A01C -5.5529040510 6.00000 BO21 -0.3679899573 0.00000
Dipole	AtomID X Y Z BondID X Y Z	A01C 0.0026390711 -0.1017025592 0.0000793040 BO21 -0.0550374750 -0.0063061599 -0.0001622745
Quadrupole	AtomID xx yy zz xy xz yz BondID xx yy zz xy xz yz	A01C -3.3866990384 -3.6375574876 - 3.7632182910 -0.0001867188 > -0.0000098072 -0.0000451636 BO21 0.1200321819 0.1391251378 - 0.4631397796. 0.1078396202 > -0.0000534634 -0.0000016527
Octupole	AtomID xxx yyy zzz xxy xxz xyy xyz xzz yyy yzz BondID xxx xxx yyy zzz xxy xxz xyy xyz xzz yyz yzz	A01C 0.027476932 -0.667528614 0.000329639 -0.045657940 > 0.000088282 0.011664784 0.000089971 0.003128359 > -0.090760843 0.000004116 BO21 -0.266164335 -0.095983087 - 0.000765879 -0.057532632 > -0.000341883 -0.182715179 -0.000130804 -0.184308730 >

Polarization Interactions In the EFP method, polarization is described as the interaction of induced dipoles on a fragment with the static electric field produced by surrounding fragments. Individual polarizable points for an EFP fragment are stored in the database in the manner as described by Table 5.2:

The number of induced dipoles on a fragment is determined by the number of valence molecular orbitals it is. The location of each polarizability is placed at a centroid (CT) of localized molecular orbital in the valence shell.

Dispersion Interactions Dispersion interactions are derived from instantaneous electronic densities represented as dynamic polarizability tensors distributed on LMO centroids at atomic centers and lone pairs.

Table 5.2.: Sample Polarization EFP Parameters

Polarization Parameter	Description	Ex:
Polarizability Tensors	UniqCT X Y Z (9 Component Polarizability Tensor)	CT1 -3.1518596318 -3.3980274071 0.0018185767 0.8103624630 3.5034026252 0.2848509668 1.2801236336 > -0.0026967104 -0.0067993635 0.0624753347 0.0003226178 >- 0.0056448562

Table 5.3.: Sample Dispersion EFP Parameters

Dispersion Parameter	Description	Ex:
Dynamic polarizability Tensors	UniqCT (9 Component of dynamic polarizability tensor)	CT1 -3.1518596318 -3.3980274071 0.0018185767 0.8103624630 3.5034026252 0.2848509668 1.2801236336 > -0.0026967104 -0.0067993635 0.0624753347 0.0003226178 > -0.0056448562

Exchange Repulsion Interactions

Exchange Repulsion interactions are related to basis set, localized wavefunctions, fock matrix elements, and LMO positions. These parameters are described in Tab 5.4:

Table 5.4.: Sample Exchange-Repulsion EFP Parameters

Exchange Repulsion Parameter	Description	Ex
Basis Set	UniqCT X Y Z Nuclear_Charge Orbital_type. Total_No. Orb_No. parameter parameter . . . Orbital_type. Total_No. Orb_No. parameter parameter . . .	A01C 1.3509909583 2.6366038344 0.0001873532 4.0 S 6 1 3047.5248800000 0.53634519 2 457.3695180000 0.98945214 3 103.9486850000 1.59728255 4 29.2101553000 2.07918728 5 9.2866629600 1.77417427 6 3.1639269600 0.61257974 L 3 7 7.8682723500 -0.39955639 1.29608216 8 1.8812885400 -0.18415517 0.99375360 9 0.5442492580 0.51639033 0.49595269 L 1 10 0.1687144782 0.18761794 0.15412764 D 1 11 0.8000000000 1.11382493
Projection WaveFunction	Localized WaveFunction No_LMOs No.basis.func LMO_i lmo_ln_ref LMO coefficients of WF LMO_i lmo_ln_ref LMO coefficients of WF	PROJECTION WAVEFUNCTION 21 132 1 1 1.53278293E-03-4.79184552E-03 5.29943618E-03 2.91001237E-03-3.73745666E-06 1 2 6.32052854E-03-3.86473758E-05 2.83054133E-03 1.77609963E-06-7.89940230E-04 1 3 7.60592470E-04-2.31275003E-04 4.40984665E-04 4.08127880E-06 9.53873714E-07 1 4 5.71279932E-04-8.97114501E-04 1.15961842E-03 1.28996791E-04 3.76387139E-05 . . . 21 19-1.58756302E-03 4.57174406E-03 8.26105821E-03-7.43054183E-03-1.57523599E-02 21 20 1.02485302E-02 1.23449667E-02-9.35981670E-03-2.69385493E-02 2.91305432E-04 21 21-3.30371382E-04 3.98784854E-04 3.93956299E-04 1.67626668E-03-7.91171895E-04 21 22 7.48461994E-04-2.81970678E-03 7.63720877E-04 5.62099609E-04 9.45424748E-03 21 23 1.96343674E-02 2.05940299E-03 8.11906612E-03 3.79357585E-03-1.17720800E-02 21 24-2.65885372E-02-1.70924679E-03-1.81843091E-02-1.22538478E-02-3.12060537E-02 21 25 2.77171453E-03-4.88684145E-02 3.22585019E-03-3.88837701E-03-9.68301262E-04 21 26-8.76138201E-04 5.91140185E-03 3.18816324E-03-1.74293768E-02-4.28790371E-02 21 27-8.94331017E-03-1.11010965E-02

Table 5.4.: Sample Exchange-Repulsion EFP Parameters

Exchange Repulsion Parameter	Description	Ex
FOCK MATRIX ELEMENTS	N*(N-1)/ symmetry-unique elements of a Fock matrix (upper triangular) of the fragment	-0.9409453277 0.0116286090 -0.9164766932 -0.0080637991 > -0.0005717440 -0.7980661453 -0.0017895432 -0.0088798180 > 0.0267839152 -0.7978589349 0.0405575563 -0.0393923537 > 0.0190377046 -0.0235183752 -0.8363369067 0.0237564231 > -0.0034958749 -0.0805527622 0.0050317792 0.0834429837 > -0.6307848820 -0.0056021641 0.1850454834 0.0033275226 > -0.0256707849 0.1017850398 0.0149663042 -0.9502804183 > -0.1796628455 -0.0057640245 -0.0289225831 0.0036479624 > -0.0995291659 0.0677969945 0.0148548427 -0.9584640636 > -0.0079655065 0.0390328825 0.0052299671 -0.1142339239 > -0.1193959544 0.0165417437 0.0927266686 0.0240921650 >
LMO CENTROIDS	CT.No. X Y Z CT.No. X Y Z	CT1 -3.1518596318 -3.3980274071 0.0018185767 CT2 -3.8562547963 1.9486651754 -0.0002204416 CT3 2.4757925339 -2.0044963935 -0.0000491399 CT4 2.4847109642 1.9723120765 0.0001597885 CT5 -0.9739691261 -0.0062907392 0.0000995724 CT6 0.3070950616 -1.7972409353 -0.5520500971 CT7 -2.3242991568 2.0812634269 0.0000236813 CT8 -2.3579085977 -2.0676004504 -0.0000022473 CT9 0.1896907532 1.9929475593 0.0001496406 CT10 1.3135403389 -4.0581981413 0.0002587672

Database Composition And Structure Currently, EFP parameters for standard solvents and the S22 dataset are available on github in the raw GAMESS text-based document format. Those parameters are also included in the database. Standard amino acid parameters were obtained from the 3ENI pdb crystal structure using basis set 6-31G*. Amino acid fragments were fragmented along the C $_{\alpha}$ -C bond as demonstrated in Fig. 3.5. We provide the “similarity landscape” between 100 randomly selected amino acid residues for each amino acid monomer type using both cartesian- and PRDF- based representations along with average similarity metrics using RMSD and Cosine-Similarity. The average RMSD for the randomly selected monomer units is 1.11 Angstrom with a range of 0.11 to 1.92. As expected, we see that amino acids with longer alkyl tails such as ARG, ASN, ASP, etc. have higher average RMSD compared to smaller more compact amino acids. Overall, using the RMSD metric on the Cartesian- based representation indicates that the fragments are very similar. However, when examining similarity using RMSD and Cosine-Distance, we see

that the average total RMSD value for the subset is more than twice at about 2.34 Angstrom and average total cosine-distance values at 0.35. This implies that there is a lot more dissimilarity - or distinguishability than originally anticipated. This allows us to believe that the EFPdB database, contains a high degree of unique and thus representative instances of all 20 amino acids. Implications of this, is that the database should be able to provide adequate parameters for download - and standard amino acid parameters might not need to be recomputed.

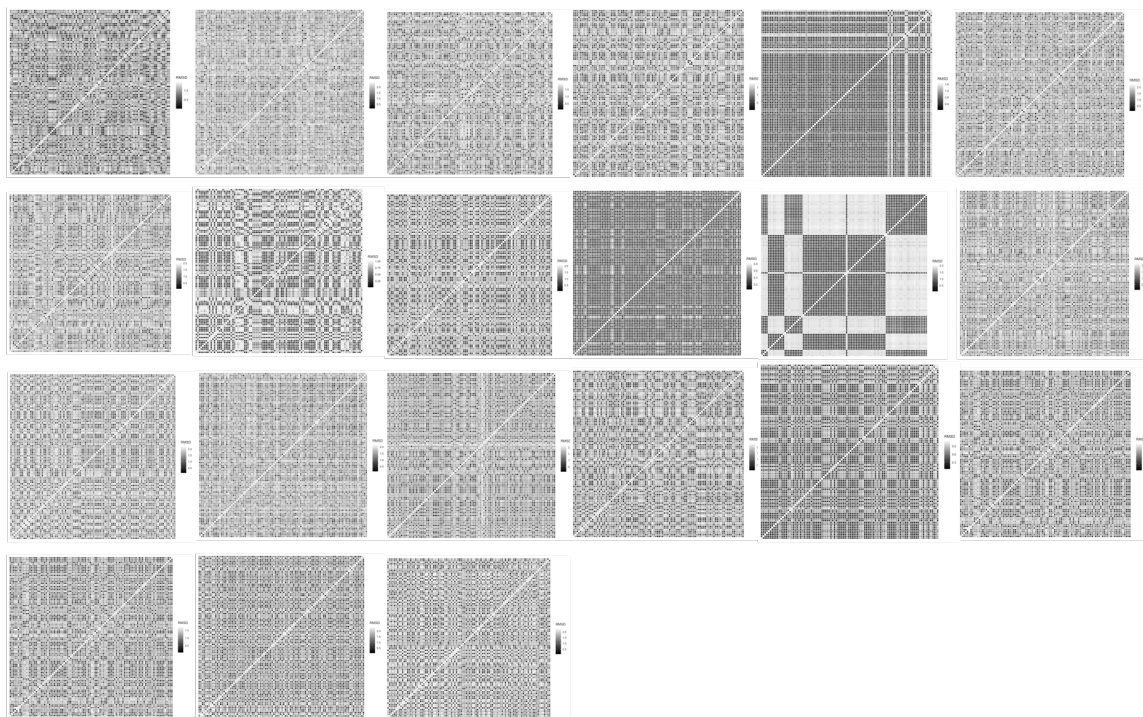


Fig. 5.1.: EFPdB Standard Amino Acid Cartesian Representation RMSD values.

Parameters will also be made available on Anaconda [94] as a package of datasets grouped into individual amino acid residues types and general solvents. Python querying scripts will also be provided to aid in parameter selection. Access to the database stored on an Amazon Web Server (AWS) is streamlined using a molecular visualizer called iSpiEFP that enables a streamed lined protocol for EFP parameters selection for multi-fragment system using visualization. Documentation for database access

Table 5.5.: EFPdB Cartesian Average Amino Acid RMSD Values

Amino Acid	Min	Max	Median	Mean	SD
ala	0.06	1.48	0.93	0.86	0.33
arg	0.19	2.46	1.70	1.65	0.35
asn	0.13	1.95	1.31	1.22	0.37
asp	0.08	1.80	1.12	1.06	0.38
cys	0.07	1.69	0.48	0.57	0.37
gln	0.15	2.25	1.45	1.40	0.38
glu	0.12	2.00	1.34	1.28	0.36
gly	0.04	1.03	0.64	0.58	0.32
hid	0.11	2.09	1.37	1.17	0.57
hie	0.07	0.83	0.26	0.29	0.12
hip	0.10	2.05	0.75	1.09	0.75
ile	0.14	2.22	1.45	1.41	0.35
leu	0.14	2.23	1.48	1.42	0.41
lys	0.18	2.18	1.42	1.39	0.34
met	0.17	2.19	1.41	1.32	0.38
phe	0.12	2.32	1.44	1.31	0.53
pro	0.08	1.11	0.50	0.48	0.27
ser	0.08	1.75	1.06	0.98	0.37
thr	0.10	1.94	1.05	1.06	0.40
trp	0.11	2.37	1.33	1.27	0.52
tyr	0.14	2.32	1.51	1.38	0.56
val	0.13	2.00	1.38	1.31	0.35
Total	0.11	1.92	1.15	1.11	0.40

Table 5.6.: EFPdB Cartesian Average Amino Acid Cosine Distance Values

Amino Acid	Min	Max	Median	Mean	SD
ala	0.63	1.00	0.87	0.87	0.08
arg	0.63	1.00	0.85	0.85	0.05
asn	0.60	1.00	0.83	0.84	0.08
asp	0.64	1.00	0.87	0.86	0.09
cys	0.62	1.00	0.97	0.93	0.08
gln	0.59	1.00	0.83	0.84	0.08
glu	0.61	1.00	0.84	0.84	0.08
gly	0.75	1.00	0.90	0.90	0.08
hid	0.67	1.00	0.87	0.88	0.09
hie	0.95	1.00	1.00	0.99	0.01
hip	0.65	1.00	0.95	0.85	0.14
ile	0.58	1.00	0.82	0.82	0.08
leu	0.50	1.00	0.81	0.81	0.10
lys	0.67	1.00	0.87	0.87	0.06
met	0.66	1.00	0.85	0.86	0.07
phe	0.61	1.00	0.87	0.87	0.09
pro	0.85	1.00	0.96	0.96	0.04
ser	0.58	1.00	0.85	0.85	0.09
thr	0.55	1.00	0.89	0.87	0.08
trp	0.70	1.00	0.93	0.92	0.05
tyr	0.65	1.00	0.87	0.87	0.08
val	0.55	1.00	0.80	0.81	0.09
TOTAL	0.65	1.00	0.88	0.87	0.08

Table 5.7.: EFPdB PRDF Average Amino Acid RMSD Values

Amino Acid	Min	Max	Median	Mean	SD
ala	0.21	1.07	0.80	0.77	0.16
arg	1.25	10.12	6.46	6.22	1.62
asn	0.45	3.16	2.25	2.18	0.49
asp	0.47	2.02	1.41	1.37	0.27
cys	0.25	1.32	0.66	0.65	0.14
gln	0.56	4.71	2.86	2.82	0.71
glu	0.53	3.52	2.13	2.11	0.53
gly	0.10	0.51	0.39	0.39	0.07
hid	0.76	4.24	2.71	2.67	0.61
hie	0.94	3.51	2.52	2.39	0.58
hip	0.59	2.61	1.80	1.74	0.40
ile	0.41	3.64	2.23	2.20	0.58
leu	0.47	3.54	2.62	2.46	0.62
lys	0.81	6.70	4.62	4.48	1.07
met	0.59	3.96	2.86	2.71	0.67
phe	0.73	5.03	3.40	3.25	0.82
pro	0.36	1.53	1.30	1.09	0.36
ser	0.23	1.52	0.97	0.94	0.26
thr	0.35	2.00	1.56	1.48	0.30
trp	0.71	7.65	4.75	4.68	1.44
tyr	0.68	4.93	3.54	3.39	0.84
val	0.41	2.13	1.59	1.51	0.35
TOTAL	0.54	1.07	2.43	2.34	0.59

Table 5.8.: EFPdB PRDF Average Amino Acid Cosine Distance Values

Amino Acid	Min	Max	Median	Mean	SD
ala	0.44	0.88	0.56	0.58	0.08
arg	0.04	0.81	0.08	0.15	0.18
asn	0.15	0.79	0.25	0.29	0.12
asp	0.26	0.77	0.37	0.40	0.10
cys	0.46	0.86	0.62	0.63	0.06
gln	0.08	0.76	0.19	0.23	0.13
glu	0.14	0.78	0.24	0.28	0.12
gly	0.68	0.92	0.77	0.77	0.04
hid	0.10	0.80	0.21	0.28	0.15
hie	0.14	0.77	0.31	0.35	0.16
hip	0.19	0.79	0.31	0.36	0.13
ile	0.12	0.84	0.22	0.27	0.15
leu	0.13	0.82	0.20	0.26	0.16
lys	0.05	0.82	0.11	0.16	0.16
met	0.10	0.80	0.16	0.21	0.15
phe	0.09	0.81	0.17	0.23	0.15
pro	0.34	0.83	0.40	0.53	0.17
ser	0.32	0.85	0.50	0.52	0.09
thr	0.26	0.79	0.34	0.39	0.12
trp	0.06	0.79	0.13	0.24	0.21
tyr	0.08	0.80	0.15	0.22	0.16
val	0.24	0.84	0.32	0.37	0.13
TOTAL	0.20	0.81	0.30	0.35	0.13

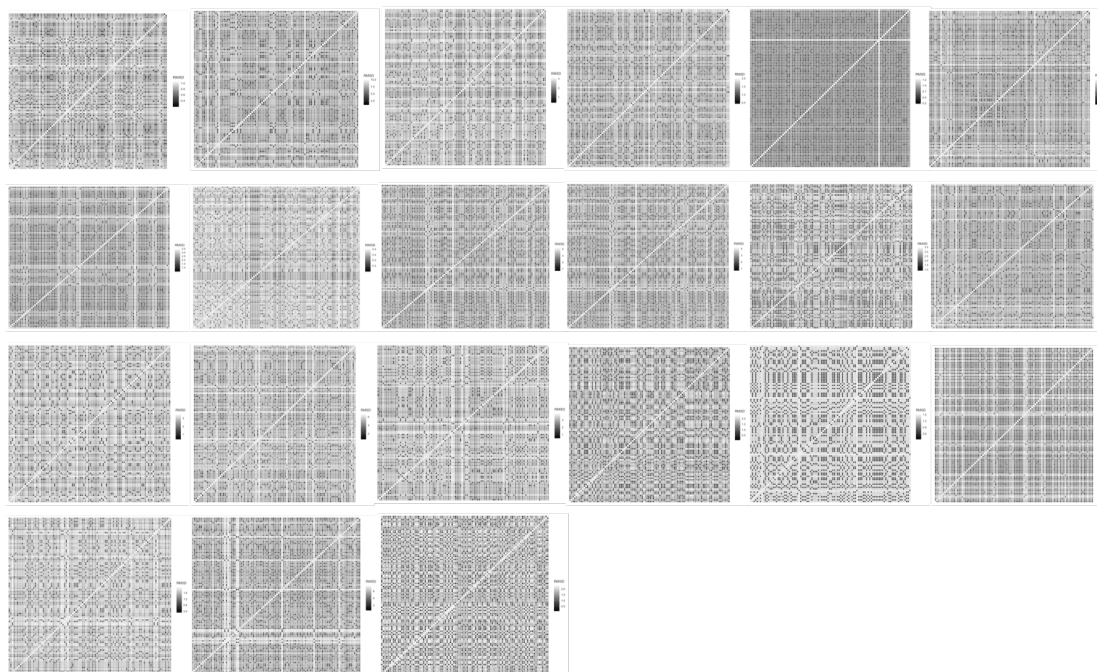


Fig. 5.2.: EFPdB Standard Amino Acid PRDF Representation RMSD values.

and tools and system fragmentation will be provided in the final chapter of this thesis.

5.2 iSpiEFP

Computational molecular modeling has made great strides in providing support for experimental studies in chemistry, physics and biology following continual advancements in hardware, networking, and data management. However, utilizing computational methods itself is daunting for the novice user as molecular simulations require not only thorough theoretical background, but also technical experience in data parsing between various programs, data visualization of large data sets, and terminal-based programming. Moreover, working with increasingly larger datasets and various data formats becomes a serious time sink and error source even for the most experienced users. To address questions of data compatibility, analysis and visualization, we introduce iSpiEFP - a local graphical user interface (GUI) that streamlines multi-

scale calculations with Effective Fragment Potential (EFP) - a sophisticated ab initio based method for modeling non-covalent interactions.

iSpiEFP serves as a workflow manager for system visualization, access to a cloud-based amazon web server (AWS) database of EFP parameters, high-performance simulations, and data analysis. It will serve as a tool that makes molecular modeling more accessible. This is necessary because molecular modeling is essential in

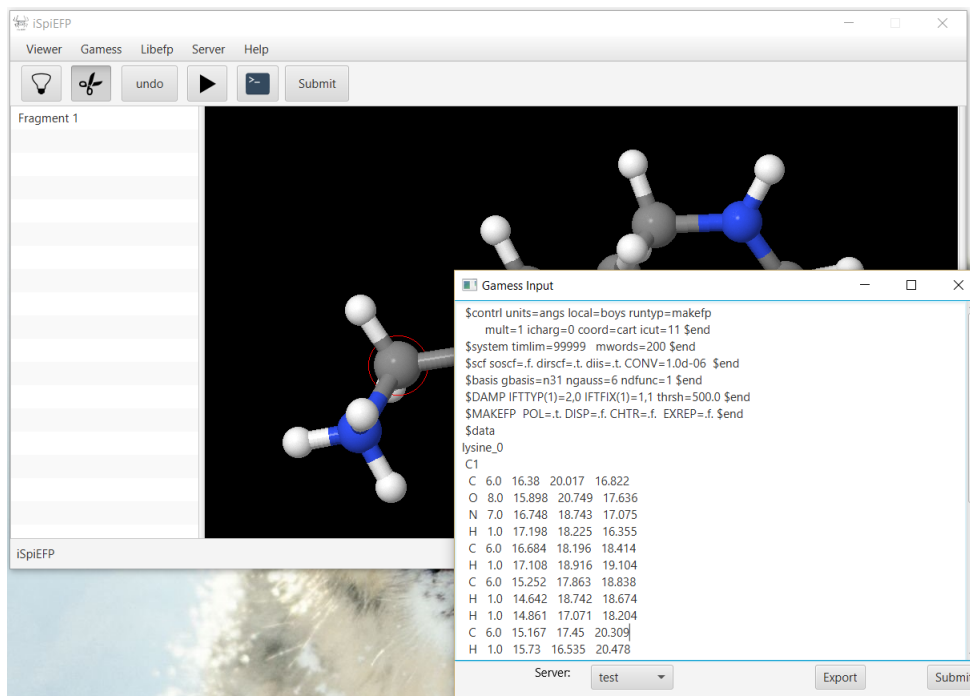


Fig. 5.3.: iSpiEFP Desktop Application

many applications of chemical, biological, medicinal and materials research, such as catalysis, drug design and design of new materials with unique chemical, optical and conducting properties. Extensive machinery of molecular modeling, including algorithms and software, has been actively developed over the past several decades. For example, tremendous progress in molecular modeling recognized by the 2013 Nobel Prize in Chemistry has advanced in silico drug design, where potential drug molecules are selected based on their interactions in a key-lock fashion with a target disease-inducing protein. iSpiEFP was designed to address two major challenges within the

field of molecular modeling at large. The first being, obtaining the appropriate compromise between accuracy of the model and its computational cost, such that more accurate models and more efficient algorithms are always in need. The other, modeling of complicated phenomena and systems imposes severe burden on user due to necessity to manage and coordinate many entangled jobs and job sequences.

Recently, EFP formalism was extended to modeling non-covalent interactions in polymers and macromolecules, which opened up new avenues for application of the method [29]. However, modeling larger or more complex systems is associated not only with higher computational cost (which often can be mitigated by scalable algorithms and more powerful computational resources), but also with a cumbersome setup of computational jobs and a non-trivial analysis of the obtained results. Indeed, molecular simulations today rarely involve only one or two calculations, but typically employ a combination of tools for modeling both static and dynamic properties of systems of interest. While a typical software deals with single jobs in isolation, the user sees all the calculations as somehow related and views all of the data in a research project as a whole. Performing molecular simulations on a scale with high throughput adds another dimension to the challenges. The process of setting up, scheduling, running computational jobs, checking for possible errors and reading the output must be automated to the largest extent possible. Furthermore, the data obtained from these massive simulations need to be consolidated so that appropriate data analysis and machine learning tools could be applied. With this tool, we aim to revisit these issues and provide a new tool iSpiEFP.

General Overview of iSpiEFP (Version 0-Alpha)

iSpiEFP (Version 0-Alpha) is a graphical user interface written in java (1.8.0_101) and designed with JavaFX - a software platform for creating and delivering desktop applications, as well as Rich Internet applications (RIAs) that can run across a wide variety of devices. JavaFX is intended to replace Swing as the standard GUI library for Java SE, but both will be included for the foreseeable future. The program's

main functionalities enable molecular system fragmentation, access to EFP parameter representation, and interfacing with remote HPCC for simulation. The visual for the general workflow for an LIBEFP calculation performed using iSpiEFP is provided in Fig. 5.4. Fragmentation, Representation, and Simulation routines are highlighted in blue, purple, and green respectively.

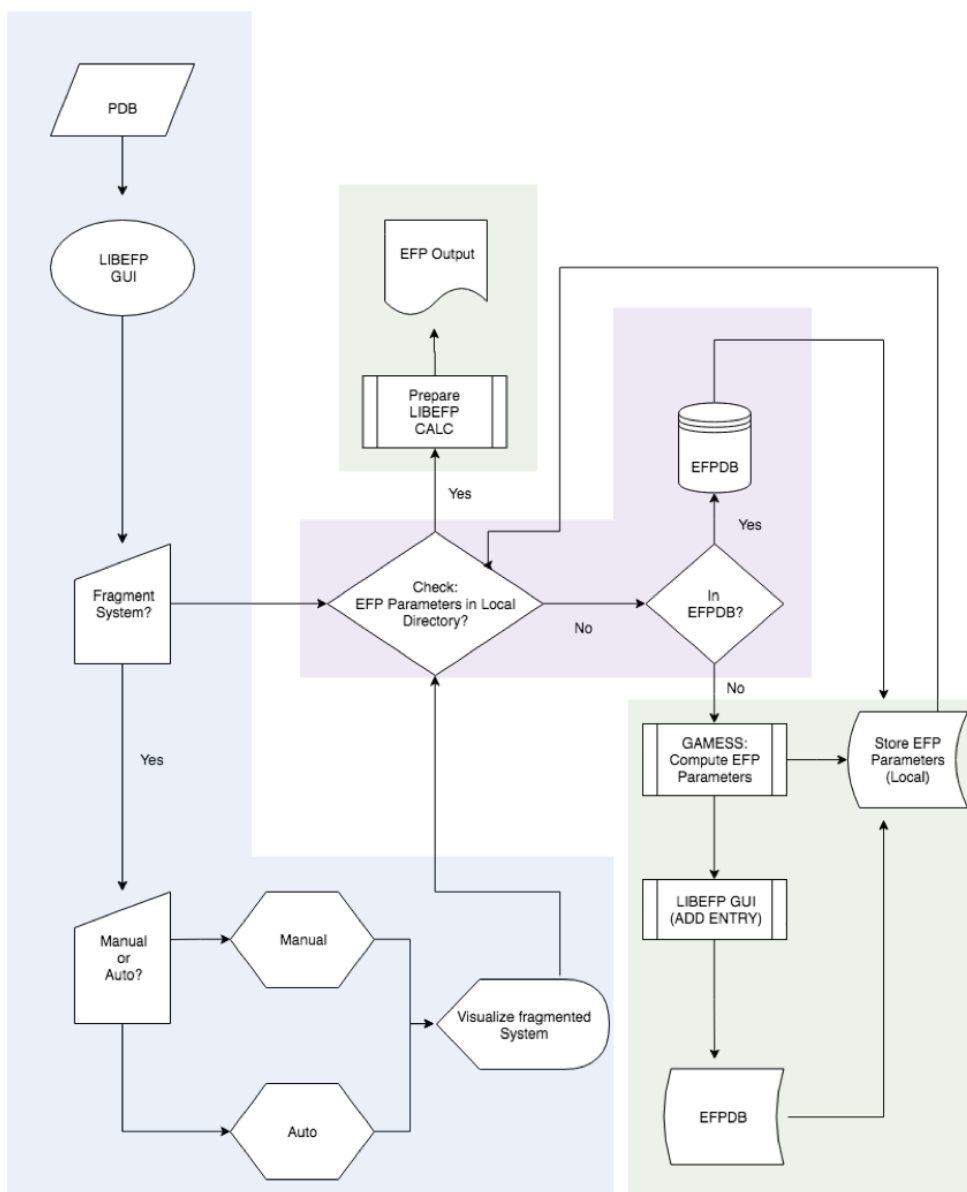


Fig. 5.4.: iSpiEFP General Workflow

Fragmentation Using the GUI, user will be able to visualize their molecular system with a text file in the PDB format. Because of the nature of EFP, it is desirable to fragment the initial molecular system into smaller subsystems called EFP ‘fragments’ in order to speed of LIBEFP calculations - however this step is optional. If fragmentation is desired, it is possible to fragment the system by calling on a ‘manual’ or ‘automated’ fragmentation routine.

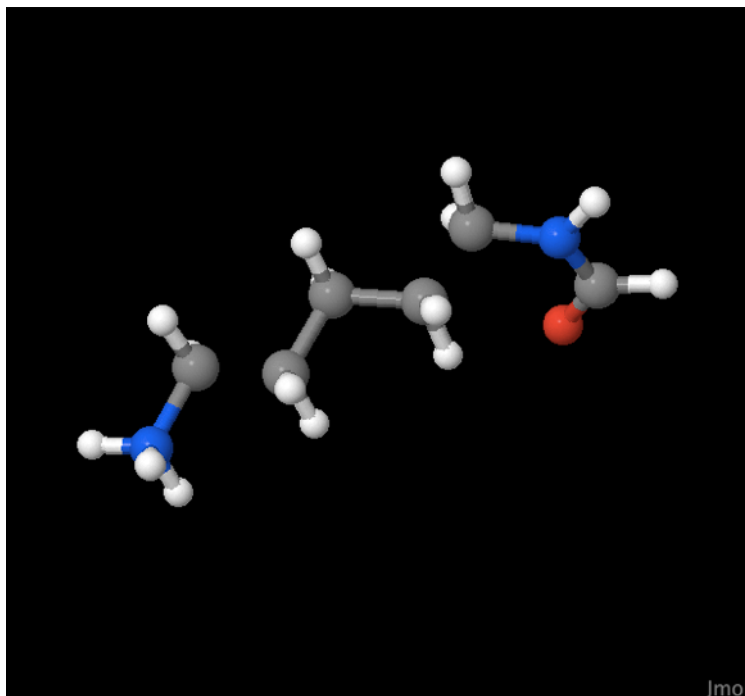


Fig. 5.5.: Lysine Molecule Fragmented Into EFP Fragments along covalent bonds.

- Manual. Manual fragmentation is more ideal for larger gas phase molecules that can be broken down into repeating subunits. This is provided with the user indicating covalent bonds to ‘cut’ using a ‘point-n-click’ routine (See Fig. 5.5). Fragmentation along covalent bonds involves the BioEFP method to ‘cut and cap’ [29].

- Automated. Automated fragmentation requires the PDB text file be in the Brookhaven format. The entire molecular system will be fragmented according to specified Residue Number and Residence Sequence Number.

In this way, biopolymers such as proteins, DNA, RNA, etc. can be fragmented automatically into amino acids and nucleoside monomer fragments. Small fragments with high degrees of free can be fragmented manually.

Representation Following fragmentation, it necessary to obtain EFP parameters for an appropriate description of each ‘unique’ fragment based on atomic composition and internal geometry. Because EFP is a rigid body-based chemical description, it is possible for two EFP fragments with the same molecular formula to require two different sets parameters. EFP parameters are obtained following fragmentation by using the geometry of the fragment and performing an electronic structure calculation of those fragments in the gas phase for each unique fragment. From those electronic structure calculations, EFP parameters can be obtained as a set of fragment potentials such as a point charges and multipoles, static, and localized wave functions. These parameters can then be thought of a sophisticated and transferrable ‘force field’ obtained at the quantum level. Determining the appropriate set of EFP parameters then is dependent upon the internal coordinates of the structure and the level of theory for fragment chemical description. iSpiEFP is connected to an online repository of already standard amino acid fragments capable of providing relevantly similar EFP parameters - in a more sophisticated search query utilizing Cosine-Distance metrics (See Chapter - Implementation).

This is helpful, because when running a libefp calculation, the method requires a set of EFP parameters for each unique fragment in the molecular system in the local directory following fragmentation. If the parameters are not within the directory, it possible to perform a MAKEFP calculation using the GAMESS [55] program to obtain the necessary EFP parameters or query the online EFPdB repository for standard solvent and amino acid EFP parameters. Parameter selection will be determined

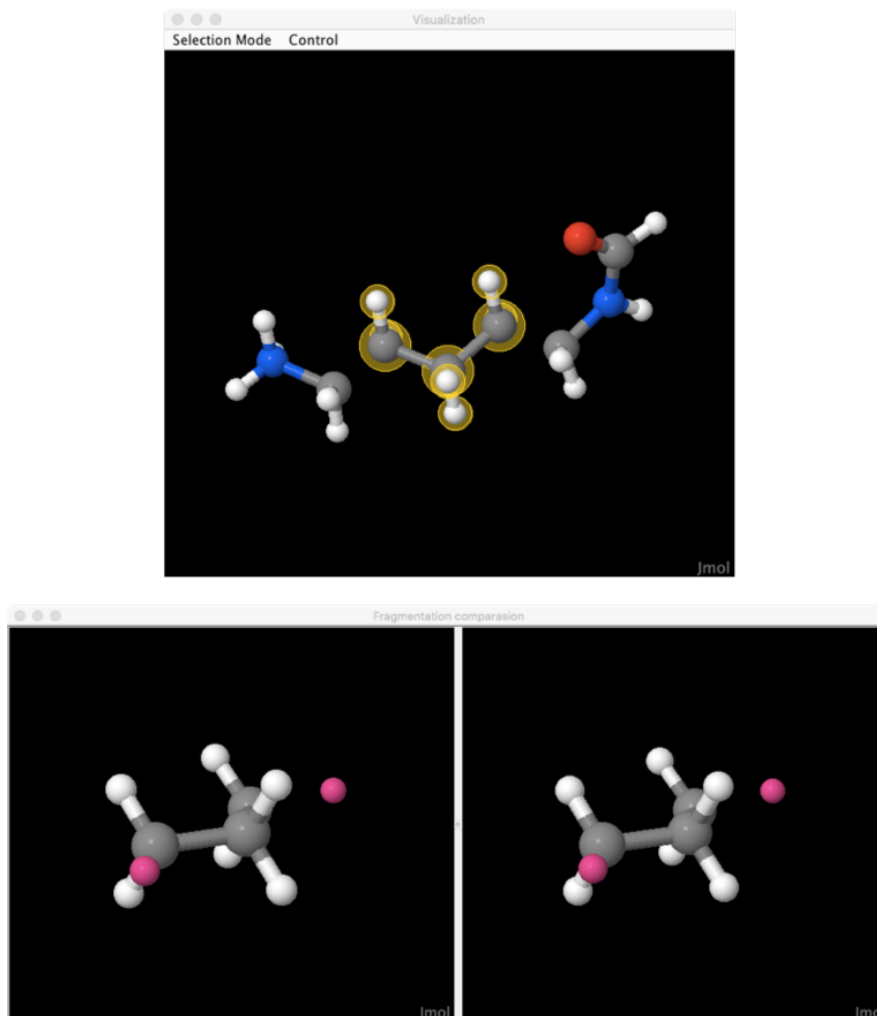


Fig. 5.6.: iSpiEFP enables fragment parameter similarity selection using RMSD and molecular fingerprint

by similarity in geometric configuration using RMSD and Cosine-Distance similarity metrics. If a suitable EFP parameter is not found on the database, a MAKEFP calculation will be suggested. All MAKEFP calculations run using iSpiEFP will be screened after submission for potential addition to the official EFPdb database of fragments. It is also possible for to current and potential collaborators to register with our group for special permission for unpublished project specific parameters.

Simulation After fragmentation and obtaining the appropriate EFP parameters for a libefp calculations, iSpiEFP provide an interactive dialog box to generate the configuration file for a libefp calculation. Prior to simulation, it is necessary to change the defaults of the SSH configurations and enter the necessary credentials to access a high performance cluster (HPC) cluster for a libefp job submission. Following simulation, the resulting log file will be provided in the iSpiEFP working directory.

Analysis iSpiEFP will also provide a feature for charting simulation log files. This enables basic graphing functionalities in order to generate simple graphs and diagrams such as scatterplot, line plots, histograms, etc. Simulations specific variables such as a time, energy, temperature, etc. can be exported as a dataframe in a comma separated (CSV) or space-delimited text file.

Technical Overview of Software Architecture

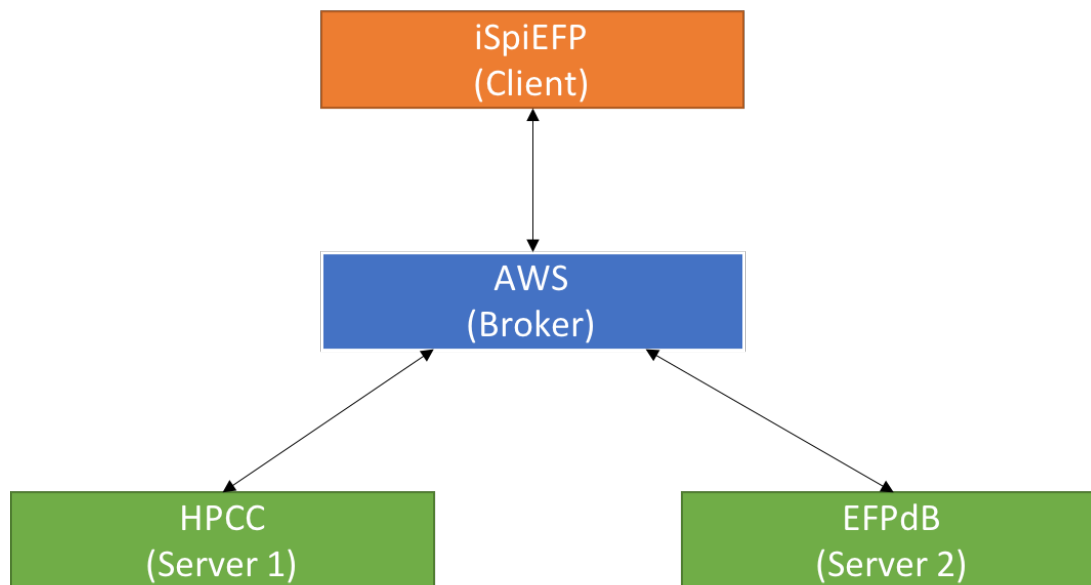


Fig. 5.7.: iSpiEFP workflow components

The iSpiEFP workflow operates on broker-type pattern in which distributed systems interact using remote service invocations (see Fig. 5.7). In this case, the 'broker' is an amazon web server (AWS) that serves in facilitating the coordination of communicating between iSpiEFP on the local client side and high performance computing cluster (HPCC) via SSH. In general sense, this means iSpiEFP (desktop client) will request a specific calculation or information set from AWS (broker), and iSpiEFP will then redirect the request to the client's remote HPCC (server 1) or to the EFPdb (server 2). For future iSpiEFP development, it is worth reporting the java class software architecture depicted in the Fig ():

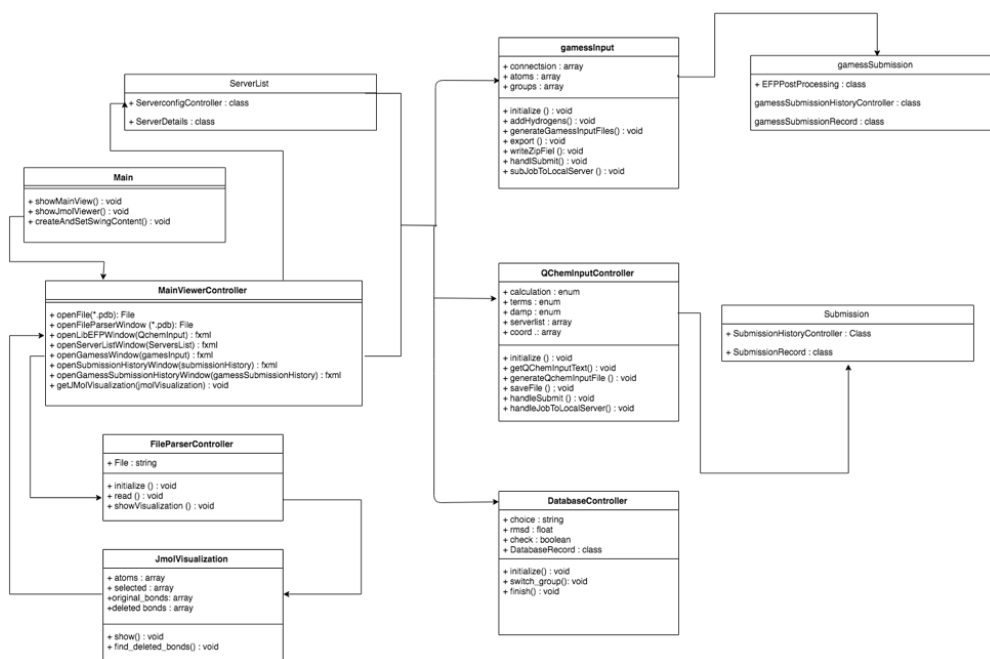


Fig. 5.8.: iSpiEFP Software Overview

- Main.java - This is the main file which launches the application by loading Main-View.fxml file. Every fxml file has an associated controller file. MainView.fxml is linked with MainViewController.java.
- MainViewController.java - This class is responsible for handling calls to open/visualize files. The method openFile has the appropriate logic for either visualizing the

xyz or pdb files/ or parsing the file (in case of libefp input/output file). Xyz and pdb files are directly passed to JmolVisualization.java for visualization. If not, the file is opened in a new window (for parsing) by loading FileParser.fxml. The other method is openLibEFPWindow (or openQChemWindow - The name is changed to libefp to indicate that appropriate changes need to be done to implement libefp instead of QChem, QChem acts as a base for its implementation). This method's name is responsible for generating libefp/Qchem Input files necessary for submitting jobs to libefp server. The other method is openServersListWindow which is responsible for loading the ServersList.fxml used for editing or configuring servers.

- JmolVisualization.java - This class contains code for integration with JMol for visualizing the xyz or pdb files. JMol uses swing's framework for displaying the molecules. However, as we're using Javafx framework it was tricky to open this JMol window through javafx. Also, there were some bugs which couldn't be fixed easily when we tried to integrate Jmol with javaFX. Hence we've adopted to open the files separately in a new JFrame window for visualization. The current file has two main methods - show and showMultipleFiles. Show method is called for displaying a single xyz file at a time. showMultipleFiles is used to visualize multiple files simultaneously within a single window. Both the methods logics are similar however, showMultipleFiles method has an additional logic of displaying prev and next buttons used for navigating across the xyz files.
- QchemInputController.java - This class is responsible for generating QChem Input file. All the parameters in the file are written in such a way that a Qchem input file can be generated. However, we need to modify this so that LibEFP input can be generated appropriately. All the fields are initialized in the initialize method of the controller, hence this method should be modified appropriately. Slight changes might need to be done to the generateQChemInputFile method. Submitting the current input file to a server is handled by the handleSubmit

method. This method is currently implemented for a local server scenario. In case of SSH scenario, first a connection need to be established before executing any command. Jsch jar can be used for this purpose.

- `ServerConfigController.java` - This class handles the adding/editing the servers list in the application by loading `ServersList.fxml`. As mentioned earlier, jobs can be submitted to any of the configured server. Users can add and edit servers using this module. Each server is implemented as a serializable class named `ServerDetails`. For testing purposes, a dummy server is created and added by default and is shown to the user when the user first launches the application (This can be deleted or can be replaced with appropriate default libefp server). However, the user can add/remove/edit servers. These calls are handled by methods - `handleAddServer`, `handleEditServer` and `handleDeleteServer`. Once any of these methods is successfully executed (i.e., either user clicks 'ok' in the next module or deletes a server), the `updateServerDetailsListInPreferences` method is called which updates the preferences with appropriate servers list. Java Utils Preferences class is used for storing this user level information. When the user clicks add/edit server, `ServerEditView.fxml` is loaded which enables the user to edit the server accordingly. A default set of queue options (used for PBS) are loaded when server hits the add server. This method can be accordingly for a libefp server.
- `ServerEditViewController.java`- This class handles adding/editing a server by loading `ServerEditView.fxml`. `setServerDetails` method needs to be called before loading this module. There are three main methods in this class - `handleOk`, `handleCancel` and `handleConfigure`. Once the input is validated, `handleOK` sets the appropriate values in the `serverDetails` and returns. The `handleCancel` method closes the window directly. The `handleConfigure` method loads the `ServerEditConfigView.fxml` file for editing the queue options (necessary for PBS and in other cases).

- `ServerEditConfigViewController.java` - This class sets the values of the queue options for that server by loading `ServerEditConfigView.fxml` file. It does a basic validation of the user inputs and set the queue options appropriately. Queue options is implemented as a new serializable `QueueOptions` class inside the `ServerDetails` class. This class might need to be updated for libefp. Once input validation is done, queueoptions are set appropriately which are then used in the `ServerEditViewController` class.

5.3 Conclusions

Here, we have presented two tools developed for the computational community: EFPdB and iSpiEFP. EFPdB is a online-repository of diverse and potentially transferable standard amino acid fragments capable of a sophisticated search query. iSpiEFP is a graphically user interface to the libefp package and EFPdB database. Both tools enable a more stream-lined experience when setting and running a libefp calculation enabling access to new and experienced users in regards to steps that were once considered time-sinks (data parsing and analysis).

6. SUMMARY

From EFP formalisms, to more recent methods of implementations of the EFP method in the libefp package, we have covered a brief technical overview of recent libefp implementations utilizing pairwise energy decomposition of the total interaction energy on a biologically relevant protein system Factor Xa and a monte-carlo based sampling method capable of finding potential local minima. We also provided some biologically relevant EFP benchmarks on the SSI dataset from which we can make basis set recommendations when generating EFP parameters. And lastly, we introduce EFPdB and iSpiEFP - two computational tools that serve to stream and automate the workflow of generating efp parameter, data parsing simulation inputs/parameters and data analysis.

REFERENCES

REFERENCES

- [1] J. K. Gilbert, C. Boulter, and M. Rutherford, "Models in explanations, part 1: Horses for courses?" *International Journal of Science Education*, vol. 20, no. 1, pp. 83–97, 1998. [Online]. Available: <https://doi.org/10.1080/0950069980200106>
- [2] "Virtual and physical molecular modeling: Fostering model perception and spatial understanding," *Educational technology society : journal of International Forum of Educational Technology Society and IEEE Learning Technology Task Force.*, vol. 4, no. 1, 2001.
- [3] G. N. Lewis, "The atom and the molecule." *Journal of the American Chemical Society*, vol. 38, no. 4, pp. 762–785, 1916. [Online]. Available: <https://doi.org/10.1021/ja02261a002>
- [4] —, "Valence and the structure of atoms and molecules," *Chemical Catalog: New York, 1923*, 1923. [Online]. Available: <https://doi.org/10.1021/ja02261a002>
- [5] M. Turner, "Ball and stick models for organic chemistry," *Journal of Chemical Education*, vol. 48, p. 407, Jun. 1971.
- [6] W. Koltun, "Space filling atomic units and connectors for molecular models," *U.S. Patent 3170246*, February 23, 1965.
- [7] R. B. Corey and L. Pauling, "Molecular models of amino acids, peptides, and proteins," *Review of Scientific Instruments*, vol. 24, no. 8, pp. 621–627, 1953.
- [8] S. E. Wheeler and J. W. G. Bloom, "Toward a more complete understanding of noncovalent interactions involving aromatic rings," *Journal of Physical Chemistry A*, vol. 118, no. 32, pp. 6133–6147, 2014.
- [9] D. a. Dougherty, "Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp." *Science (New York, N.Y.)*, vol. 271, no. 5246, pp. 163–168, 1996.
- [10] —, "The cation- π interaction," *Accounts of Chemical Research*, vol. 46, no. 4, pp. 885–893, 2013.
- [11] a. S. Mahadevi and G. N. Sastry, "Cation π Interaction: Its Role and Relevance in Chemistry, Biology, and Material Science," *Chem. Rev.*, vol. 113, no. 3, pp. 2100–2138, 2012. [Online]. Available: <http://dx.doi.org/10.1021/cr300222d>
- [12] C. Møller and M. S. Plesset, "Note on an approximation treatment for many-electron systems," *Phys. Rev.*, vol. 46, pp. 618–622, Oct 1934. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRev.46.618>

- [13] R. J. Bartlett and M. Musiał, "Coupled-cluster theory in quantum chemistry," *Rev. Mod. Phys.*, vol. 79, pp. 291–352, Feb 2007. [Online]. Available: <https://link.aps.org/doi/10.1103/RevModPhys.79.291>
- [14] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, "A fifth-order perturbation comparison of electron correlation theories," *Chemical Physics Letters*, vol. 157, no. 6, pp. 479 – 483, 1989. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0009261489873956>
- [15] R. O. Jones and O. Gunnarsson, "The density functional formalism, its applications and prospects," *Rev. Mod. Phys.*, vol. 61, pp. 689–746, Jul 1989. [Online]. Available: <https://link.aps.org/doi/10.1103/RevModPhys.61.689>
- [16] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical review.*, vol. 136, no. 3B, pp. B864–B871, 1964.
- [17] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Physical review.*, vol. 140, no. 4A, pp. A1133–A1138, 1965.
- [18] J. W. Ponder and D. A. Case, *Force Fields for Protein Simulations*, ser. Protein Simulations. Amsterdam :: Academic Press, 2003, vol. 66.
- [19] N. Foloppe and A. D. MacKerell, Jr, "All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data," *Journal of computational chemistry*, vol. 21, no. 2, pp. 86–104, 2000.
- [20] H. F. Schaefer III, "Methods of electronic structure theory," 1977.
- [21] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6," *Journal of computational chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.
- [22] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the gromos force-field versions 54a7 and 54b7," *European biophysics journal.*, vol. 40, no. 7, pp. 843–856, 2011.
- [23] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids," *Journal of the American Chemical Society.*, vol. 118, no. 45, pp. 11 225–11 236, 1996.
- [24] H. Lin and D. G. Truhlar, "Qm/mm: what have we learned, where are we, and where do we go from here?" *Theoretical chemistry accounts*, vol. 117, no. 2, pp. 185–199, 2007.
- [25] H. M. Senn and W. Thiel, "Qm/mm methods for biomolecular systems," *Angewandte Chemie International Edition*, vol. 48, no. 7, pp. 1198–1229. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200802019>
- [26] M. S. Gordon, M. A. Freitag, P. Bandyopadhyay, J. H. Jensen, V. Kairys, and W. J. Stevens, "The effective fragment potential method: A qm-based mm approach to modeling environmental effects in chemistry," *The Journal of Physical Chemistry A*, vol. 105, no. 2, pp. 293–307, 2001.

- [27] P. N. Day, J. H. Jensen, M. S. Gordon, S. P. Webb, W. J. Stevens, M. Krauss, D. Garmer, H. Basch, and D. Cohen, "An effective fragment method for modeling solvent effects in quantum mechanical calculations," *The Journal of chemical physics*, vol. 105, no. 5, pp. 1968–1986, 1996.
- [28] D. E. Williams, "Representation of the molecular electrostatic potential by atomic multipole and bond dipole models," *Journal of Computational Chemistry*, vol. 9, no. 7, pp. 745–763. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.540090705>
- [29] P. K. Gurunathan, A. Acharya, D. Ghosh, D. Kosenkov, I. Kaliman, Y. Shao, A. I. Krylov, and L. V. Slipchenko, "Extension of the effective fragment potential method to macromolecules," *The Journal of Physical Chemistry B*, vol. 120, no. 27, pp. 6562–6574, 2016.
- [30] A. V. Nemukhin, B. L. Grigorenko, A. V. Rogov, I. A. Topol, and S. K. Burt, "Modeling of serine protease prototype reactions with the flexible effective fragment potential quantum mechanical/molecular mechanical method," *Theoretical chemistry accounts*, vol. 111, no. 1, pp. 36–48, 2004.
- [31] J. H. Jensen, H. Li, A. D. Robertson, and P. A. Molina, "Prediction and rationalization of protein p k a values using qm and qm/mm methods," *The journal of physical chemistry.*, vol. 109, no. 30, pp. 6634–6643, 2005.
- [32] D. A. Bondarev, W. J. Skawinski, and C. A. Venanzi, "Nature of Intercalator Amiloride - Nucleobase Stacking . An Empirical Potential and ab Initio Electron Correlation Study," *J. Phys. Chem. B*, vol. 104, pp. 815–822, 2000.
- [33] R. Malham, S. Johnstone, R. J. Bingham, E. Barratt, S. E. V. Phillips, C. A. Laughton, S. W. Homans, V. Uni, and N. Ng, "Strong Solute - Solute Dispersive Interactions in a Protein - Ligand Complex," *J. AM. CHEM. SOC.*, vol. 127, no. 7, pp. 17 061–17 067, 2005.
- [34] P. Hobza and C. Jir, "Non-covalent interactions in biomacromolecules," *Phys. Chem. Chem. Phys.*, vol. 9, pp. 5291–5303, 2007.
- [35] R. J. Bartlett and M. Musiał, "Coupled-cluster theory in quantum chemistry," *Reviews of Modern Physics*, vol. 79, no. 1, pp. 291–352, 2007.
- [36] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, "Reprint of: A fifth-order perturbation comparison of electron correlation theories," *Chemical Physics Letters*, vol. 589, pp. 37–40, 2013. [Online]. Available: <http://dx.doi.org/10.1016/j.cplett.2013.08.064>
- [37] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *Journal of Computational Chemistry*, vol. 31, no. 4, p. NA, Mar. 2009. [Online]. Available: <http://dx.doi.org/10.1002/jcc.21367>
- [38] A. J. Cohen, P. Mori-Snchez, and W. Yang, "Challenges for density functional theory," *Chemical reviews.*, vol. 112, no. 1, pp. 289–320, 2012.

- [39] M. Casida and M. Huix-Rotllant, "Progress in time-dependent density-functional theory," *Annual review of physical chemistry*., vol. 63, no. 1, pp. 287–323, 2012.
- [40] S. P. Webb and M. S. Gordon, "Solvation of the menshutkin reaction: a rigorous test of the effective fragment method," *The journal of physical chemistry*., vol. 103, no. 9, pp. 1265–1273, 1999.
- [41] I. Adamovic and M. S. Gordon, "Solvent effects on the s_n2 reaction: application of the density functional theory-based effective fragment potential method," *The journal of physical chemistry*., vol. 109, no. 8, pp. 1629–1636, 2005.
- [42] K. Ohta, Y. Yoshioka, K. Morokuma, and K. Kitaura, "The effective fragment potential method. An approximate ab initio mo method for large molecules," *Chemical Physics Letters*, vol. 101, no. 1, pp. 12–17, 1983.
- [43] M. S. Gordon, Q. A. Smith, P. Xu, and L. V. Slipchenko, "Accurate first principles model potentials for intermolecular interactions," *Annu. Rev. Phys. Chem.*, vol. 64, p. 553, 2013.
- [44] M. S. Gordon, L. Slipchenko, H. Li, and J. H. Jensen, "The effective fragment potential: a general method for predicting intermolecular interactions," *Annual reports in computational chemistry*, vol. 3, pp. 177–193, 2007.
- [45] M. D. Hands and L. V. Slipchenko, "Intermolecular interactions in complex liquids: Effective fragment potential investigation of water tert -butanol mixtures," *The journal of physical chemistry*., vol. 116, no. 9, pp. 2775–2786, 2012.
- [46] Q. A. Smith, M. S. Gordon, and L. V. Slipchenko, "Benzenepyridine interactions predicted by the effective fragment potential method," *The journal of physical chemistry*., vol. 115, no. 18, pp. 4598–4609, 2011.
- [47] L. V. Slipchenko and M. S. Gordon, "Water-benzene interactions: An effective fragment potential and correlated quantum chemistry study," *Journal of Physical Chemistry A*, vol. 113, no. 10, pp. 2092–2102, 2009.
- [48] K. Gierszal and J. Davis, "Pi-Hydrogen Bonding in Liquid Water," *The Journal of Physical Chemistry Letters*, vol. 2, no. 22, pp. 2930–2933, 2011. [Online]. Available: <http://dx.doi.org/10.1021/jz201373e>
<http://pubs.acs.org/doi/abs/10.1021/jz201373e>
<http://pubs.acs.org/doi/full/10.1021/jz201373e>
<http://pubs.acs.org/doi/pdf/10.1021/jz201373e>
- [49] B. M. Rankin, M. D. Hands, D. S. Wilcox, K. R. Fega, L. V. Slipchenko, and D. Ben-Amotz, "Interactions between halide anions and a molecular hydrophobic interface," *Faraday discussions*, vol. 160, pp. 255–270, 2013.
- [50] D. Kosenkov and L. V. Slipchenko, "Solvent effects on the electronic transitions of p-nitroaniline: A qm/efp study," *The Journal of Physical Chemistry A*, vol. 115, no. 4, pp. 392–401, 2010.
- [51] J. C. Flick, D. Kosenkov, E. G. Hohenstein, C. D. Sherrill, and L. V. Slipchenko, "Accurate prediction of noncovalent interaction energies with the effective fragment potential method: Comparison of energy components to symmetry-adapted perturbation theory for the S22 test set," *Journal of Chemical Theory and Computation*, vol. 8, no. 8, pp. 2835–2843, 2012.

- [52] L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerell, K. M. Merz, and C. D. Sherrill, "The biofragment database (bfdb): An open-data platform for computational chemistry analysis of noncovalent interactions," *Journal of chemical physics.*, vol. 147, no. 16, 2017.
- [53] T. M. Parker, L. A. Burns, R. M. Parrish, A. G. Ryno, and C. D. Sherrill, "Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies," *Journal of Chemical Physics*, vol. 140, no. 9, 2014.
- [54] M. S. Marshall and C. D. Sherrill, "Dispersion-weighted explicitly correlated coupled-cluster theory [dw-ccsd(t**)-f12]," *Journal of chemical theory and computation : JCTC.*, vol. 7, no. 12, pp. 3978–3982, 2011.
- [55] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su *et al.*, "General atomic and molecular electronic structure system," *Journal of computational chemistry*, vol. 14, no. 11, pp. 1347–1363, 1993.
- [56] I. A. Kaliman and L. V. Slipchenko, "Libefp: A new parallel implementation of the effective fragment potential method as a portable software library," *Journal of computational chemistry*, vol. 34, no. 26, pp. 2284–2292, 2013.
- [57] —, "Hybrid mpi/openmp parallelization of the effective fragment potential method in the libefp software library," *Journal of computational chemistry*, vol. 36, no. 2, pp. 129–135, 2015.
- [58] N. Homeyer and H. Gohlke, "Free energy calculations by the molecular mechanics poisson- boltzmann surface area method," *Molecular Informatics*, vol. 31, no. 2, pp. 114–122, 2012.
- [59] N. Huang, C. Kalyanaraman, K. Bernacki, and M. P. Jacobson, "Molecular mechanics methods for predicting protein–ligand binding," *Physical Chemistry Chemical Physics*, vol. 8, no. 44, pp. 5166–5177, 2006.
- [60] C. McInnes, "Virtual screening strategies in drug discovery," *Current opinion in chemical biology*, vol. 11, no. 5, pp. 494–502, 2007.
- [61] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, "Docking and scoring in virtual screening for drug discovery: methods and applications," *Nature reviews Drug discovery*, vol. 3, no. 11, p. 935, 2004.
- [62] Y. Chen and D. Zhi, "Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule," *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 2, pp. 217–226, 2001.
- [63] W. L. Jorgensen, "The many roles of computation in drug discovery," *Science*, vol. 303, no. 5665, pp. 1813–1818, 2004.
- [64] G. R. Marshall, "Computer-aided drug design," *Annual review of pharmacology and toxicology*, vol. 27, no. 1, pp. 193–213, 1987.
- [65] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures," *The FEBS journal*, vol. 275, no. 1, pp. 1–21, 2008.

- [66] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.
- [67] K. Raha, M. B. Peters, B. Wang, N. Yu, A. M. Wollacott, L. M. Westerhoff, and K. M. Merz Jr, "The role of quantum mechanics in structure-based drug design," *Drug discovery today*, vol. 12, no. 17-18, pp. 725–731, 2007.
- [68] M. De Vivo, "Bridging quantum mechanics and structure-based drug design," *optimization*, vol. 7, p. 8, 2011.
- [69] M. W. van der Kamp and A. J. Mulholland, "Combined quantum mechanics/molecular mechanics (qm/mm) methods in computational enzymology," *Biochemistry*, vol. 52, no. 16, pp. 2708–2728, 2013.
- [70] M. K. Gilson and H.-X. Zhou, "Calculation of protein-ligand binding affinities," *Annual review of biophysics and biomolecular structure*, vol. 36, 2007.
- [71] K. M. Merz Jr, "Limits of free energy computation for protein- ligand interactions," *Journal of chemical theory and computation*, vol. 6, no. 5, pp. 1769–1776, 2010.
- [72] N. R. Patel, D. V. Patel, P. R. Murumkar, and M. R. Yadav, "Contemporary developments in the discovery of selective factor xa inhibitors: a review," *European journal of medicinal chemistry*, vol. 121, pp. 671–698, 2016.
- [73] D. J. Pinto, J. M. Smallheer, D. L. Cheney, R. M. Knabb, and R. R. Wexler, "Factor xa inhibitors: next-generation antithrombotic agents," *Journal of medicinal chemistry*, vol. 53, no. 17, pp. 6243–6274, 2010.
- [74] R. M. Parrish, D. F. Sitkoff, D. L. Cheney, and C. D. Sherrill, "The surprising importance of peptide bond contacts in drug–protein interactions," *Chemistry–A European Journal*, vol. 23, no. 33, pp. 7887–7890, 2017.
- [75] R. M. Parrish and C. D. Sherrill, "Spatial assignment of symmetry adapted perturbation theory interaction energy components: The atomic sapt partition," *The Journal of chemical physics*, vol. 141, no. 4, p. 044115, 2014.
- [76] R. M. Parrish, T. M. Parker, and C. D. Sherrill, "Chemical assignment of symmetry-adapted perturbation theory interaction energy components: the functional-group sapt partition," *Journal of chemical theory and computation*, vol. 10, no. 10, pp. 4417–4431, 2014.
- [77] B. Jeziorski, R. Moszynski, and K. Szalewicz, "Perturbation theory approach to intermolecular potential energy surfaces of van der waals complexes," *Chemical Reviews*, vol. 94, no. 7, pp. 1887–1930, 1994.
- [78] K. Szalewicz, "Symmetry-adapted perturbation theory of intermolecular forces," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 2, pp. 254–272, 2012.
- [79] Y. Shi, D. Sitkoff, J. Zhang, H. E. Klei, K. Kish, E. C.-K. Liu, K. S. Hartl, S. M. Seiler, M. Chang, C. Huang *et al.*, "Design, structure- activity relationships, x-ray crystal structure, and energetic contributions of a critical p1 pharmacophore: 3-chloroindole-7-yl-based factor xa inhibitors," *Journal of medicinal chemistry*, vol. 51, no. 23, pp. 7541–7551, 2008.

- [80] H. Matter, M. Nazaré, S. Güssregen, D. W. Will, H. Schreuder, A. Bauer, M. Urmann, K. Ritter, M. Wagner, and V. Wehner, "Evidence for c cl/c br π interactions as an important contribution to protein–ligand binding affinity," *Angewandte Chemie International Edition*, vol. 48, no. 16, pp. 2911–2916, 2009.
- [81] H. G. Wallnoefer, T. Fox, K. R. Liedl, and C. S. Tautermann, "Dispersion dominated halogen– π interactions: energies and locations of minima," *Physical Chemistry Chemical Physics*, vol. 12, no. 45, pp. 14 941–14 949, 2010.
- [82] P. Jureka, J. poner, J. ern, and P. Hobza, "Benchmark database of accurate (mp2 and ccscd(t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs," *Physical chemistry chemical physics*, vol. 8, no. 17, pp. 1985–1993, 2006.
- [83] L. Grfov, M. Pitok, J. ez, and P. Hobza, "Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended s22 data set," *Journal of chemical theory and computation : JCTC.*, vol. 6, no. 8, pp. 2365–2376, 2010.
- [84] O. Matsuoka, E. Clementi, and M. Yoshimine, "Ci study of the water dimer potential surface," *Journal of chemical physics*, vol. 64, no. 4, pp. 1351–1361, 1976.
- [85] Z. HUBLEK, "Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation," *Biological reviews of the Cambridge Philosophical Society.*, vol. 57, no. 4, pp. 669–689, 1982.
- [86] "Marine ecology and the coefficient of association: a plea in behalf of quantitative biology," *Journal of ecology.*, vol. 8, no. 1, 1920.
- [87] B. S. G. Britain), "Similarity-based approaches to virtual screening." *Biochemical Society Transactions.*, vol. 31, no. Pt 3, pp. 603–606.
- [88] I. S. P. S. S. C. Author, "Automated binary texture feature sets for image retrieval," in *1996 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-96*, ser. 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 4. [Place of publication not identified]: IEEE, 1996, pp. 2239–2242.
- [89] P. Willett, J. M. Barnard, and G. M. Downs, "Chemical similarity searching," *Journal of chemical information and computer sciences.*, vol. 38, no. 6, pp. 983–996, 1998.
- [90] S. C. Li, "The difficulty of protein structure alignment under the rmsd," *Algorithms for molecular biology.*, vol. 8, no. 1, 2013.
- [91] R. A. Abagyan and M. M. Totrov, "Contact area difference (cad): a robust measure to evaluate accuracy of protein models," *Journal of molecular biology.*, vol. 268, no. 3, pp. 678–685, 1997.
- [92] "Calculate root-mean-square deviation (rmsd) of two molecules using rotation."
- [93] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, "Materials cartography: Representing and mining materials space using structural and electronic fingerprints," *Chemistry of materials.*, vol. 27, no. 3, pp. 735–743, 2015.

- [94] “Anaconda software distribution.” *Computer software. Vers. 2-2.4.0. Anaconda, Nov. 2016. Web. [jhttps://anaconda.com](https://anaconda.com)*.