Summer 8-5-2019

Formulation of Hybrid Knowledge-Based/Molecular Mechanics Potentials for Protein Structure Refinement and a Novel Graph Theoretical Protein Structure Comparison and Analysis Technique

Aaron Maus
University of New Orleans, amaus@uno.edu

Formulation of Hybrid Knowledge-Based/Molecular Mechanics Potentials for Protein Structure Refinement and A Novel Graph Theoretical Protein Structure Comparison and Analysis Technique

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering & Applied Sciences
Computer Science

Aaron P. Maus

B.S. University of New Orleans, 2011
B.S. University of New Orleans, 2011

August, 2019

I would like to dedicate this dissertation to my parents David and Tracy whose endless support and encouragement has allowed me to take advantage of the opportunities afforded me. I would also like to dedicate this to my brother and sister, Brandon and Kristen, and to my grandparents Bernard and Sylvia Warren, and Carl and Georgiana Maus. All of your support has made this possible.

Acknowledgements

First and foremost, I would like to thank my advisor Christopher Summa for years of guidance, assistance, and friendship. He has guided me towards interesting projects while still giving me the freedom to follow my interests. His open-door policy and unending willingness to entertain ideas and offer feedback have been instrumental in this work coming to fruition. I have learned from him how to be a scientist and I will eternally be grateful for his introduction to the enormous world of the small, of molecular biology.

I would also like to thank my other committee members: Md Tamjidul Hoque, Steven Rick, David Worthylake, and Dimitrios Charalampidis. Their support, feedback, and patience throughout this process has been greatly appreciated. In particular Md Tamjidul Hoque's feedback, criticism, and comments have been instrumental in helping me ask the right questions and decide on avenues to explore.

I would also like to acknowledge current and past members of the Summa Lab and of the Bioinformatics and Machine Learning group. These include Jonathan Redmann, Devin Villegas, Manuel Zubieta, Avdesh Mishra, Sumaiya Iqbal, and Denson Smith. Their friendship, support, feedback, and great conversations throughout my years in the program have made them rewarding ones.

I would like to thank the University of New Orleans and the Department of Computer Science for providing an excellent research environment and for the financial support. I especially thank Mahdi Abdelguerfi, who has sponsored and supported me throughout my undergraduate

and graduate career at the University of New Orleans. I only hope that one day I can pay it forward.

Lastly, but definitely not the least, I would like to express my gratitude for the love, support, and patience of my family throughout my graduate studies. I wish to thank my partner, Shaina Monet, for her understanding, patience, and support as well, bearing with the years of this process and the months of my single-minded focus as I finished up this work. None of this would have been possible without my family, for the motivation to follow my dreams from my parents and for the support and encouragement of my brother and sister. Thank you all.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
CAD	Contact Area Difference
CASP	Critical Assessment of Protein Structure Prediction
CATH	Class Architecture Topology Homologous Superfamily (Protein Structure Classification Database)
ENCAD	Energy Calculation and Dynamics
ERoS	Expanded Regions of Similarity
GDT	Global Distance Test
GDT_TS	Global Distance Test Total Score
KB	Knowledge Based
KB/MM	Knowledge Based/Molecular Mechanics
L-BFGS	Limited memory Broyden Fletcher Goldfarb and Shanno
LCS	Longest Continuous Segment
lDDT	local Distance Difference Test
LGA	Local Global Alignment
MD	Molecular Dynamics
MM	Molecular Mechanics
NNSM	Near Native Structure Models
NS	Native State
PDB	Protein Data Bank

PEM	Potential Energy Minimization
PI	Percent Improvement
PMF	Potential of Mean Force
PSI-BLAST	Position Specific Iterated BLAST
RAPDF	Residue-specific All-atom Probability Discriminatory Function
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	Root Mean Square Deviation
RoS	Region of Similarity
SCOP	Structural Classification of Proteins
TM-Score	Template Modelling Score

Abstract

Proteins are the fundamental machinery that enables the functions of life. It is critical to understand them not just for basic biology, but also to enable medical advances. The field of protein structure prediction is concerned with developing computational techniques to predict protein structure and function from a protein's amino acid sequence, encoded for directly in DNA, alone. Despite much progress since the first computational models in the late 1960's, techniques for the prediction of protein structure still cannot reliably produce structures of high enough accuracy to enable desired applications such as rational drug design. Protein structure refinement is the process of modifying a predicted model of a protein to bring it closer to its native state. In this dissertation a protein structure refinement technique, that of potential energy minimization using hybrid molecular mechanics/knowledge based potential energy functions is examined in detail. The generation of the knowledge-based component is critically analyzed, and in the end, a potential that is a modest improvement over the original is presented.

This dissertation also examines the task of protein structure comparison. In evaluating various protein structure prediction techniques, it is crucial to be able to compare produced models against known structures to understand how well the technique performs. A novel technique is proposed that allows an in-depth yet intuitive evaluation of the local similarities between protein structures. Based on a graph analysis of pairwise atomic distance similarities, multiple regions of structural similarity can be identified between structures independently of relative orientation. Multidomain structures can be evaluated and this technique can be combined with global measures of similarity such as the global distance test. This method of comparison is expected to have broad

applications in rational drug design, the evolutionary study of protein structures, and in the analysis of the protein structure prediction effort.

Keywords: Bioinformatics; Protein Structure Prediction; Protein Structure Refinement; Statistical Energy Functions; Protein Structure Comparison; Graph Analysis

Chapter 1

1. Introduction

Proteins are the molecular machines that enable and facilitate the functions of life. From neurons firing, to oxygen circulating throughout organisms, to DNA replication and cell reproduction, proteins are integral in allowing these processes to occur. Not only are they critical for the biological processes within our bodies, but they are also key for the mechanisms that allow many viruses and diseases to afflict us. For example, it is a protein complex on the surface of HIV that allows it to select and attack the vital CD4+ T cells of the human immune system, and it is a misfolded protein due to a single genetic mutation that causes sickle cell anemia. Whether for the purposes of better understanding our basic biology or for the purposes of treating diseases and designing medicines, it is crucial to understand the proteins involved.

A well-known biological adage states, "form follows function" although, in the case of structural biology, it is more practically understood as "function follows form". If one wants to understand the function of a protein, one needs to understand its structure [1]. The way that structures have historically been determined has been through x-ray crystallography, a technique developed in the early 20th century with the first atomic resolution structure, that of table salt, solved in 1914 [2] . The first structures of proteins, myoglobin and haemoglobin, shown in Figure 1.1, were determined in this way by Kendrew and Perutz in the late 1950s [3]-[5]. This technique is

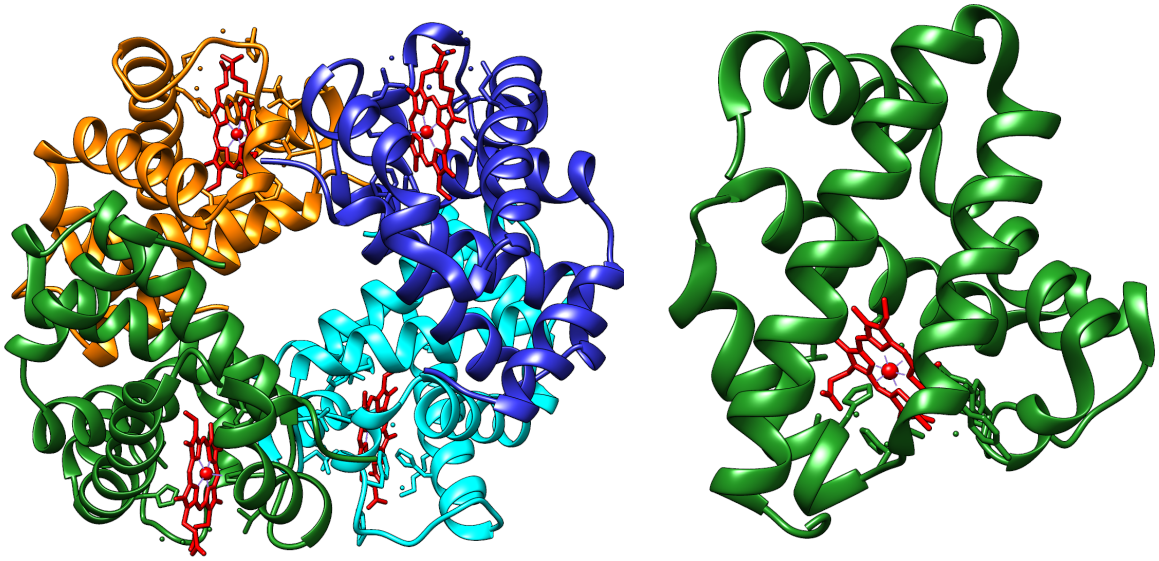

Figure 1.1: The first proteins whose structures were determined. Left, Haemoglobin, determined via x-ray crystallography. Haemoglobin is composed of four subunits, colored orange, green, cyan, and blue, each of which contains a single haem group that binds with oxygen for ferrying red blood cells. PDB accession code 1A3N. **Right**, Myglobin, determined by x-ray crystallography. Myglobin consists of a single unit which contains a single haem group for bind oxygen. PDB accession code5ZZF

widely used today and is joined by other empirical techniques such as NMR spectroscopy [6], [7] and electron microscopy [8]. While these techniques have provided tens of thousands of structures [9], they are time and labor intensive, and there remain cases such as disordered proteins and membrane proteins that are still difficult or even infeasible with modern empirical techniques.

The field of protein structure prediction is concerned with developing computational techniques to determine the structures of proteins. The goal is to provide a quicker, cheaper, and more flexible analysis of new protein structures than empirical methods can provide and to enable the study of proteins that are difficult or infeasible with those methods. This field dates to 1967 when Levitt, Lifson, and Warshel wrote the first computer program representing a protein energy force field and used it to refine the structures of two proteins: myoglobin and lysozyme [10], [11]. Their work follows from Anfinsen's thermodynamic hypothesis: all the information necessary to

determine the structure of a protein is encoded in its amino acid sequence [12]. The intuition is that a protein is a collection of atoms and as such should obey physical laws. A computer program that characterizes these laws and applies them to the atoms of a protein via a numerical optimization process should be able to "fold" that protein from a disordered state into the precise 3-dimensional structure, its native state, that it is drawn towards in nature.

The protein folding problem has turned out to be non-trivial and remains unsolved. It can be argued that there are two major difficulties in the computational protein folding problem. The first is that the interactions within a protein and between a protein and its surrounding solvent are inherently quantum mechanical and that simulating even small systems using quantum mechanics remains infeasible. As Feynman pointed out, quantum mechanical simulations require exponential growth in space and time based on the number of particles, and an exact simulation may not even be possible [13]. A classical simulation on the other hand grows quadratically, and as a result, a large effort in computational physics and chemistry has gone into characterizing classical and statistical energy functions that approximate the true quantum mechanical energy functions as closely as possible.

As Levinthal famously pointed out, another major difficulty in the protein folding problem is that the conformational search space of even a small protein is astronomically large [14]. Given that every amino acid in a polypeptide chain has two flexible backbone torsion angles (ϕ and ψ) that define its local backbone geometry, a chain of 100 residues would have 198 such angles. Assuming that each angle has three stable conformations, this modestly sized chain would have a total of 3^{198} different conformations, a number of conformations greater than the age of the universe in picoseconds. Brute force sampling is not an option, and efficient algorithms to sample and explore the conformational space are a prerequisite to solving the protein folding problem.

As a result of these two major difficulties, a myriad of different algorithms and different energy functions have been developed for the folding of protein structures. In order to evaluate progress in the field, a biannual Critical Assessment of Protein Structure Prediction [15]-[17] experiment is held. In this experiment, protein structure predictors are given the sequences of proteins whose structures have been empirically determined but not yet published. CASP is a blind test of predictors' ability to accurately predict these structures, and it allows predictors to be ranked based on their performance and the best methods to be presented and discussed. In CASP, it is therefore of critical importance to be able to compare predicted models against the native structures and identify their similarity or lack thereof.

1.1 Dissertation Contributions

In this dissertation, the formulation of hybrid molecular mechanics/knowledge-based potentials used for protein structure refinement, specifically the knowledge-based portion of these potentials, is examined in detail. Two questions are asked. The first is: can the performance of the potential be improved by modifying the starting database by either having more strict requirements on the structures included and/or increasing the size of the database to improve the statistics? In the latter case, the hypothesis explored is that a larger body of statistics will smooth out the energy surface, allowing structures easier access to energetic minima. The second question explored is whether or not the classification of atomic interactions within a protein structures into the default 167 atom types as defined by the residue-specific all-atom probability discriminatory function (RAPDF) [18] is the optimal classification scheme for potentials of mean force (PMF) [19]. A rigorous computational approach was taken by defining a measure of atom type similarity and then iteratively combining similar atom types into "merged" atom types under the hypothesis that the combined statistics of atom types with similar characteristics can be leveraged to produce a better

performing potential. The resulting potentials are tested and analyzed. It is shown that combining atom types does result in improved refinement using potential energy minimization, and, in the end, a potential that is a modest improvement over the original, KB_0.1 [20], is presented.

In the formulation and testing of dozens of potentials for structure refinement, it is natural to ask what the practical differences between two protein structure predictors are. That is, does one (for example, a potential with a smoother energy surface generated from a larger statistical database) better form missing hydrogen bonds than another, and what would that look like in the resulting structures? Would large scale, consistent changes be noticeable, such as secondary structures being brought together or, more generally, the formation of difficult structural motifs like beta-sheets? How would one identify these differences between sets of produced models and their natives? Can local similarities and differences between pairs of structures and patterns in the similarities of sets of structures be identified? This thought experiment led to the second project presented in this dissertation.

A novel technique has been developed that allows for the identification of all regions of local similarity between two protein structures, irrespective of changes in global similarity such as domain shifts or conformational changes in disordered regions of those structures. This technique allows structures to be ranked according to their overall local similarity and can be combined with measures of global similarity such as GDT_TS [21] to identify structures that are both globally and locally similar. It allows for regions of local similarity to be visualized either at the sequence level or on the 3D structural representations of the proteins. Sequence level visualization allows for quick and easy analysis of sets of structures. For example, a set of models produced of some native can be analyzed. Likewise, three-dimensional structural representations allow for detailed looks into the similarities and differences of individual pairs of structures. A tool to identify and visualize

regions of similarity is freely available on GitHub[1], and this work is expected to have applications in the analysis of evolutionarily related proteins, in drug-design, and in the evaluation of protein structure predictors.

[1] https://github.com/amaus/jProt

Chapter 2

2. An Introduction to Protein Structure, Prediction, and to Protein Structure Refinement using Hybrid KB/MM Potentials

2.1 The Structure of Proteins

Proteins are composed of one or more polypeptide chains, each composed of a sequence of amino acids. The sequence of these amino acids alone determines the structure of a protein [12] as it is their interactions within the protein and between them and the solvent surrounding the protein that cause it to fold into its natural or "native" state. While we say that a protein has a native state, reality is more complex. A protein is flexible and, *in vivo*, can shift between multiple stable conformations [22] as it interacts with other proteins, substrates, or ligands.

2.1.1 Protein Structure Hierarchy

Protein structures are complex and as first proposed by Linderstøm-Lang, they are often described in a hierarchical fashion [23], [24]. There are four levels of protein structure: primary, secondary, tertiary, and quaternary, shown in Figure 2.1.

2.1.1.1 Primary Structure

The lowest level in the structure hierarchy, primary structure, refers to the amino acid sequence of a polypeptide chain. A protein's sequence is directly encoded by a segment of base pairs in an organism's DNA, and mutations in DNA can cause mutations in the encoded proteins. As a result

of the genomic sequencing, the amino acid sequence of any protein in an organism can be determined.

2.1.1.2 Secondary Structure

At the next level in the structure hierarchy, secondary structures are regularly repeating local structural motifs within polypeptides. The two most common forms of secondary structure (first described by Corey and Pauling before the first structures of proteins had been determined [25]) are alpha helices, which are helices characterized by having 3.6 residues per turn in the helix, and beta sheets, although there are other rarer forms of secondary structure including the 3_{10} and π helices and alpha sheets. Secondary structures are formed and stabilized by networks of hydrogen bonds and they form spontaneously on the pathway to the final stable conformation of a protein.

2.1.1.3 Tertiary Structure

Secondary structures come together to form the tertiary structure of a polypeptide chain. The formation of tertiary structure is guided and stabilized by a variety of forces and inter-residue bonds acting on and within the polypeptide. These include the hydrophobic effect, where hydrophobic residues will naturally form the core of a structure where they are "protected" from water by outer hydrophilic residues, and include hydrogen bonds, disulfide bonds, and ionic bonds between residues separated in sequence within the structure

2.1.1.4 Quaternary Structure

Many proteins consist of multiple polypeptide chains. The quaternary structure of a protein is defined by the arrangement of the tertiary substructures of that protein. For example, as Figure 2.1 shows, hemoglobin is an oligomer consisting of four subunits that non-covalently group together to

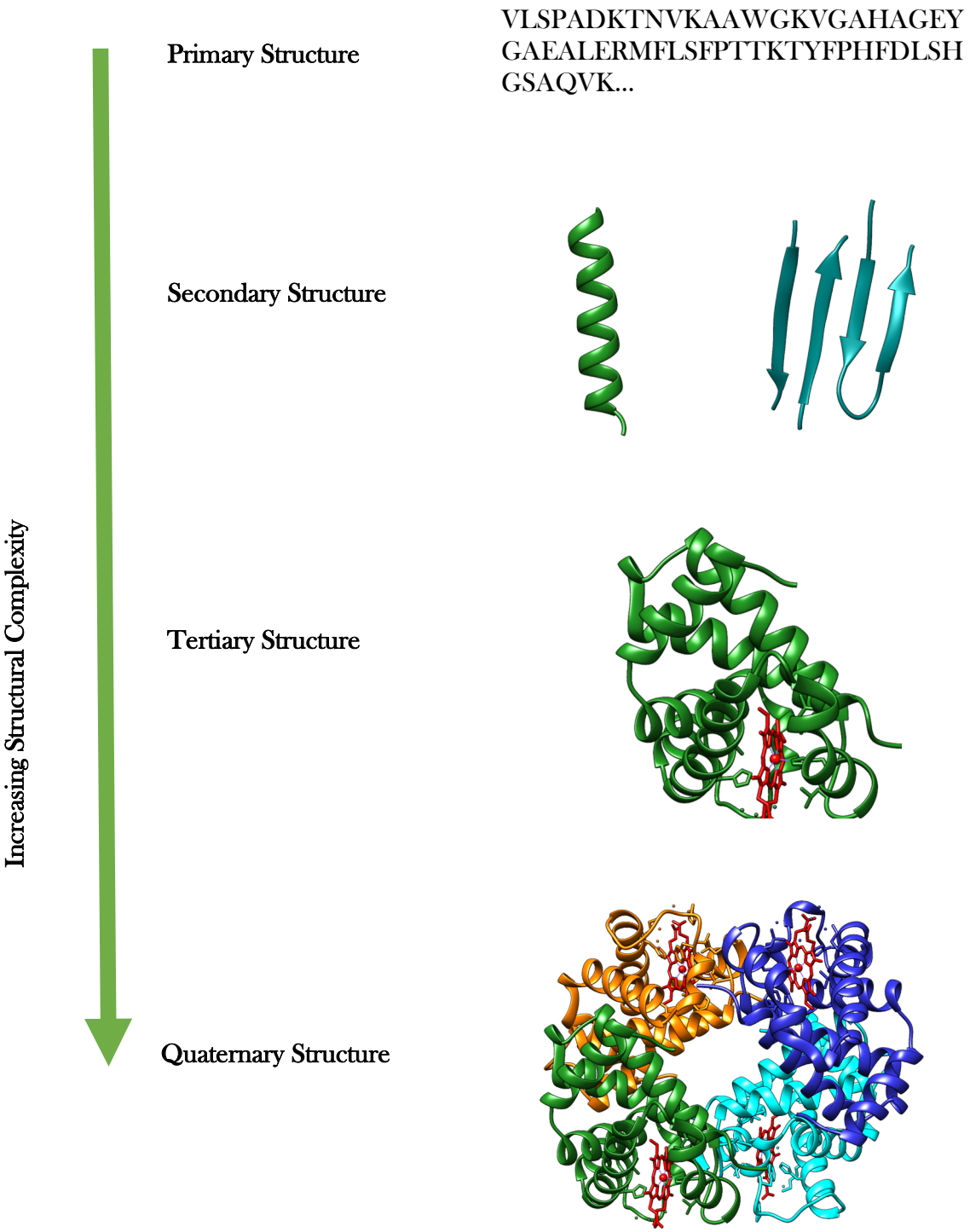

Figure 2.1: Protein Structure Hierarchy. Protein structure is classified into a hierarchy of increasing complexity. The primary structure consists of a polypeptide sequence. Secondary Structures are regularly repeating motifs that form spontaneously during the folding process. They include alpha helices and beta sheets (left and right). Tertiary structure consists of the arrangement of the secondary structures of a single polypeptide. Shown is a subunit of hemoglobin. In red is the haem group containing iron. Quaternary structure is the arrangement of the tertiary components of a protein. Shown is the whole hemoglobin consisting of four subunits which noncovalently group together forming its quaternary structure.

form its quaternary structure. Another example is the envelope glycoprotein hemagglutinin, the oligomer responsible for the selection and membrane fusion of influenza with target cells, shown in Figure 2.2.

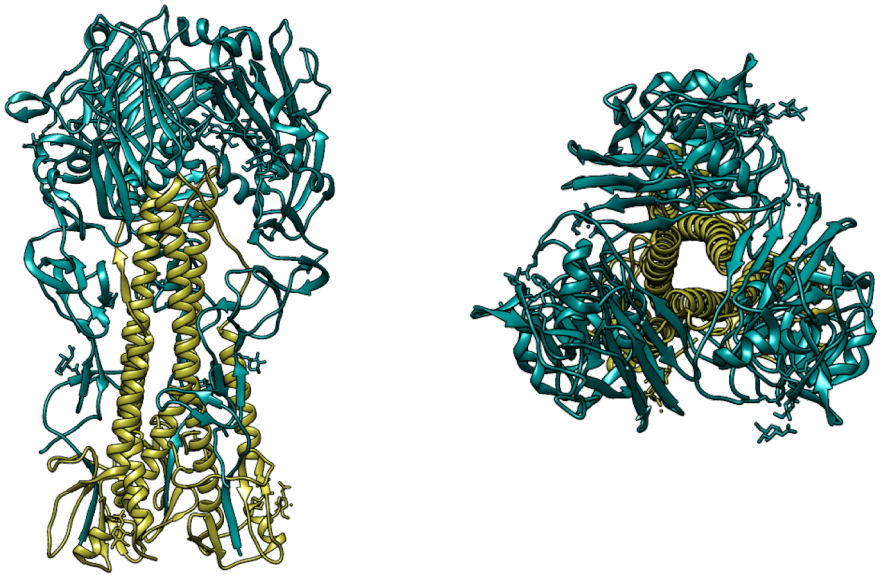

Figure 2.2: The Structure of Hemagglutinin. Shown from the side (left) and top (right), Hemagglutitin is an oligomer consisting of six units of tertiary structure arranged in three-fold symmetry. On top are three identical globular tertiary components responsible for target selection, and in the center are three helical tertiary components responsible for membrane fusion. Shown is hemagglutinin H1 responsible for the 1918 pandemic. PDB ascension code 1RUZ.

2.1.2 Protein Structure Classification

Despite the fact that the number of unique protein sequences is large and that each sequence has an astronomically large number of possible conformations, the number of actual conformations expressed is relatively small, and, in fact, it has been shown that a number of sequences can still result in the same structure [26], [27]. In other words, the sequence space for all possible proteins is larger than the structural space and any given structure may be producible from a number of different sequences. It is therefore not surprising that protein structures tend to have common patterns, and the same "folds" crop up again and again in protein structure analysis.

There are two major projects which have taken to classify and organize proteins into hierarchies of similar structures: the Structural Classification of Proteins (SCOP2) project [28] and the Class Architecture Topology Homologous fold (CATH) database [29]. They both classify structures at the highest respective level based on secondary structure composition, i.e., all alpha helices, all beta sheets, a mix of both, or mainly disordered. From there, structures are classified into various folds: conformations that share similar secondary structure arrangements and topologies. Both databases also take evolutionary information into account, classifying structures by their evolutionary relationship.

2.2 Protein Folding Techniques

The goal of protein structure predictors is simple: given a protein sequence, determine its native state, the conformation it is drawn towards *in vivo*. While this problem seemed insurmountable a few decades ago, there has been much progress in recent years [15], [30], [31]. In general, there are two classes of techniques for protein structure prediction: template-based modelling and *ab initio* prediction.

2.2.1 Template-Based Modelling

With the curation of large datasets of known sequences and structures such as the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) database [9], SCOP [28], and CATH [29], along with powerful sequence alignment tools such as the basic local alignment search tool (BLAST) and position-specific iterated BLAST (PSI-BLAST) [32], [33] it is possible to use this existing information to guide structure prediction. Comparative, or homology modelling [34], [35] and protein threading [36] both make use of this existing information.

2.2.1.1 Homology Modelling

To predict the structure of a sequence via homology modelling, a sequence alignment [37], [38] against a database of known structures is performed to find a homolog of the sequence, relying on the assumption that proteins with significant sequence similarity will generally share the same fold since evolution preserves protein structure and function even though the sequence may change through genetic mutations. If a homolog is found, it is then used as a starting template, and a model is built on that template using one of several possible techniques: rigid-body assembly, segment matching, or through the satisfaction of spatial constraints [39]. Leveraging the accumulated data of decades of structural biology, as long as a reliable template is found, that is, one with sufficient sequence similarity, homology modelling regularly produces accurate predictions and as a class of methods, remains the most accurate used in CASP [40].

2.2.1.2 Fold Recognition aka Protein Threading

If a sequence with sufficiently high sequence identity for homology modelling is not found, then fold recognition, or, protein threading, may be used[36]. The goal of protein threading is to identify a template for a sequence that shares the same fold even though the sequence identity may be low. A set of possible templates from a variety of folds is identified by selecting structures with low sequence identity to the target sequence. Then for each structure, the target sequence is "threaded" onto it and its fit is evaluated via a scoring function. The structure with the best fit for the target sequence can then be used as the starting template for a model to be built using the technique from homology modelling.

2.2.2 *Ab Initio* Prediction

If either of the techniques above are not applicable, that is, if there are no homologous sequences with known structures in existing protein databases, then a sequence's structure must be predicted using *ab initio* techniques[41]-[44]. *Ab initio* techniques fold a protein from first principles and remain among the most difficult techniques for protein structure predictions. They involve searching a protein's conformational space to identify stable, low energy conformations[45]. One possibility is the exploration of the conformational space via monte-carlo sampling [46]-[48] combined with energy minimization or molecular dynamics (MD) simulation [49]-[54] [55]. Alternatively, Dill proposes a "zipping" and assembly method based on the idea that as a protein folds, local metastable structures will form which will then subsequently fold into larger structures [30]. CASP has shown that in the past few years, much progress in *ab initio* techniques has been made by restricting the conformational search space using inter-residue contact predictions from the analysis of residue coevolution by machine learning algorithms[56]-[58].

2.2.3 Protein Structure Refinement

Whether structures are produced via template-based modelling techniques or through *ab initio* prediction, the resulting models are not consistently of native quality. Furthermore, even the most reliable technique, homology modelling, still cannot reliably produce models of sufficiently high accuracy (< 1.0 Å RMSD) for the target applications of protein structure prediction such as rational drug discovery [59]-[62]. In order to move resulting models of any modelling process closer to the native, protein structure refinement is applied.

Refinement processes tend to use one or both of two techniques [59], running MD simulations to allow a near native structure model (NNSM) to explore the conformational space around it, or performing potential energy minimization (PEM) [11], [63], [64] to bring a NNSM to

the nearest local minimum in its energy function landscape. These methods both rely on the assumption that a starting structure is close to its native state. Under this assumption, when using MD simulations, the conformation space to sample is small and, when performing PEM, the nearest minimum is likely the native. Chapter 3 will focus on structure refinement using PEM.

2.2.3.1 Potential Energy Minimization

In potential energy minimization, the energy of a protein structure, as a function of the three-dimensional coordinates of the atoms of that structure, is minimized using numerical optimization. Structure refinement using PEM goes back to the earliest days of protein structure prediction [11], [63]. There are two general classes of energy functions for PEM: traditional physics-based molecular mechanics (MM) potentials and statistically derived "knowledge-based" (KB) [65] potentials. An example of a traditional MM potential can be given as the Energy Calculation and Dynamics (ENCAD) potential [64], [66] which takes the following form:

$$U_{potential} = \sum \frac{1}{2} K_b (b - b_o)^2 + \sum \frac{1}{2} K_\theta (\theta - \theta_o)^2 + \sum \frac{1}{2} K_\phi [1 - \cos(n\phi + \delta)] + \sum \varepsilon [(r_o/r)^{12} - 2(r_o/r)^6] + 332 \sum q_i q_j / r \quad (2.1)$$

The first three terms quantify the energetic contributions for bonded interactions: bond stretches, bond angle bends, and torsion angle twists, respectively. The last two terms represent nonbonded interactions: van der Waals interactions (represented by a Lennard-Jones style function), and electrostatic interactions. The potential energy of a structure is calculated as the sum over all energetic terms, over all bonded and nonbonded interactions. By perturbing the coordinates of a structure's atoms via a numerical optimization method such as the limited memory Broyden-

Fletcher-Goldfarb-Shanno (L-BFGS) [67] technique, its potential energy can be minimized bringing it ideally closer to its native state.

2.2.3.2 Potentials of Mean Force

KB potentials take the form of potentials of mean force (PMF) [19]. Rather than deriving potential functions from physics, they are derived from the statistics of a large set of known protein structures. PMFs are based on Boltzmann's principle, which can be interpreted as saying that states of a system that are seen with high frequency correspond to the low energy states. Given a set of native structures, it should be possible to identify the patterns within them which correspond to low energy states and build energy functions from these patterns. The intuition behind the formulation of PMFs is that they quantify how the patterns that exist within protein structures differ from what would be expected if no consistent forces were at play (i.e., if the atoms existed as an ideal gas). The process of generating a PMF can be outlined as follows.

For the purposes of gathering statistics for a PMF, atoms within proteins are classified into a set of atom types. Most commonly, atoms are categorized into 167 different residue-specific heavy atom types defined by Samudrala and Moult for their RAPDF potential [18]. Other possibilities include categorizing atoms into their basic heavy atom type, e.g., Cα, Cβ, N, O, etc., grouping sets of atoms within residues into virtual atoms, or grouping chemically and functionally similar atoms into virtual atoms. Using the RAPDF schema, atom types are denoted using the following convention: the residue is specified, followed by the atom, followed by its side chain position. Side chain positions are specified using the Greek alphabet from α to ζ. If the atom is on the backbone, no position is specified. For example, AN indicates the backbone nitrogen of alanine, and FCζ indicates the zeta carbon in phenylalanine. For convenience, atom types will be specified using Romanized script. E.g., FCζ will be written as FCZ.

To generate a PMF, given a database of known protein structures and an atom type categorization schema, for each pair of atom types, their contact distances within all structures are counted and sorted into a set of distance bins for that pair. Figure 2.3 gives an example of such counts for a pair of atom types AN and ACB. At the end of this process, each atom pair will have its own set of distance bins where each bin contains the number of pairs of atoms of those two types that were found to be X distance apart in the database. The number of bins and their width are parameters chosen during the design of a PMF. Once all the counts are determined, they are then converted into energy values using one of several currently used derivations [18], [68]-[70]. Figure 2.4 shows the corresponding energy function for the counts shown in Figure 2.3 as calculated using Lu and Skolnick's formalism [71]. This process is performed for all pairs of atom types, and the entire collection of energy curves constitutes the PMF.

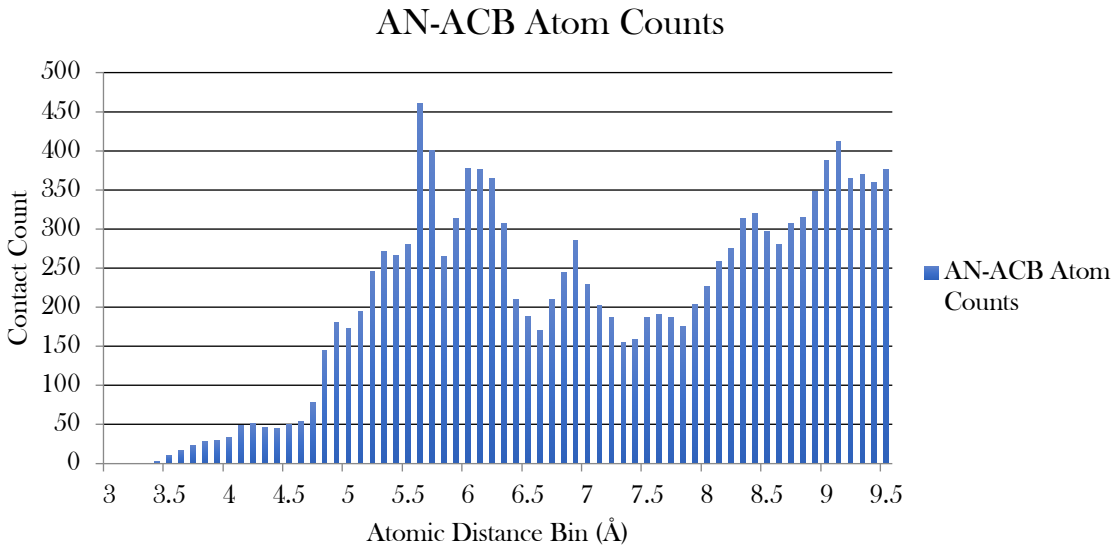

Figure 2.3: Histogram of the contacts between the atom types AN and ACB. Generated from the Top500 Structure Database from the Richardson Lab. The histogram shows, for example, that approximately 450 AN-ACB atom pairs were observed at a distance of 5.6 Å in this database.

Whether an energy function takes the form of a MM potential or a PMF, its use as a potential energy function for PEM is the same. In either case, the energy of a structure is calculated as the sum over all energetic terms. For a MM potential, the sum includes all bonds, angles, and

torsions along with pairwise non-bonded van der Waals and electrostatics interactions, and in a PMF the energy is the sum over all pairwise atomic interactions. The calculated energy can then be minimized via a quasi-Newton optimization method such as l-BFGS algorithm to refine the 3D-coordinates of the structure.

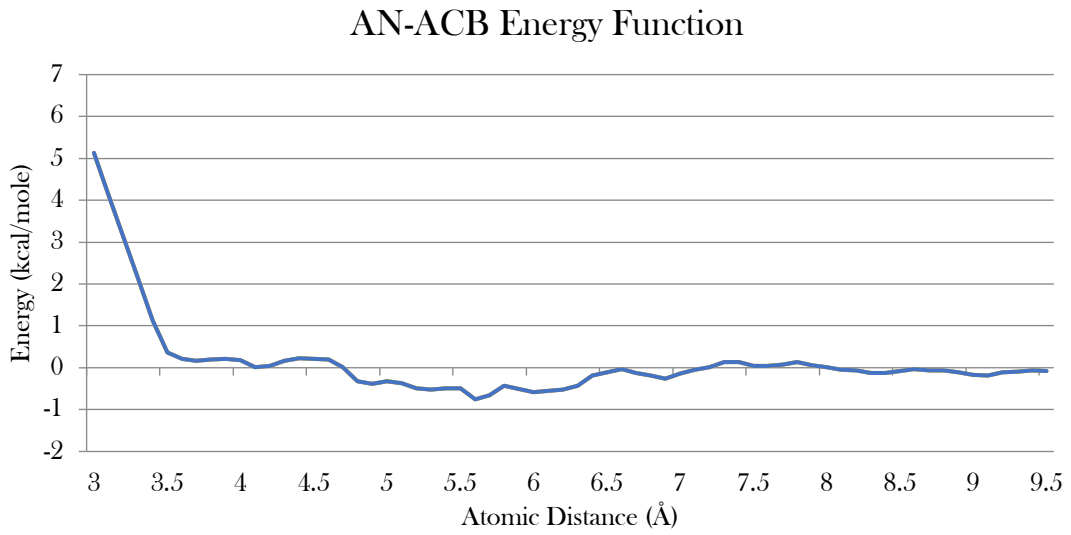

Figure 2.4: Energy function derived from the contact counts shown in Figure 2.3. Energies were calculated using Lu and Skolnick's formalism.

2.2.3.3 Hybrid KB/MM Potentials for *in vacuo* Structure Refinement

As Summa and Levitt showed [20], a KB potential can be combined with a MM potential, and the resulting hybrid potential performs better in protein structure refinement than purely MM potentials alone. In a MM potential, the energetic contributions can be broken up into two broad categories, bonded and non-bonded interactions. While the bonded interactions are the stronger interactions, they are relatively few. The non-bonded interactions on the other hand are many and though they are weaker than the bonded interactions, they are more likely to contain systematic errors due to the neglect of quantum mechanical interactions between atoms. The hybrid potential uses the energetic terms for the bonded interactions from the ENCAD MM potential and represents the nonbonded interactions using a PMF. Since PMFs are built from databases of

known protein structures, quantum mechanical effects are implicitly accounted for. Likewise, the effects of surrounding solution on the structures is also implicitly accounted for. Not only does this free refinement from having to be performed via MD simulations (to explicitly model all water molecules), but PEM using KB/MM potentials with implicit solution results in a greater percent improvement in model distance to the native [72].

2.2.3.3.1 Generating Hybrid KB/MM Potentials

Summa and Levitt generated three different KB/MM potentials. The potentials differed in the generation of their PMFs. Each PMF was generated using proteins from the Top500 Database from the Richardson lab, using all 167 atom types defined in Samudrala and Moult's RAPDF [18]. All atomic interactions less than 20 Å, excluding those from within the same residue or neighboring residues, were included. The PMFs differed in the width of the distance bins into which the statistics were gathered: 0.1, 0.2 and 0.5 Å.

For each PMF, the pairwise counts for each atom type pair were converted into energies using the method of Lu and Skolnick [71], as defined by Eq. 2.2 and Eq. 2.3, with an included repulsive close-contact portion at low distances increasing monotonically to a plateau of 80 kcal/mol. Using Lu and Skolnick's formalism, the energy for atom types i and j for distance bin d is calculated as

$$\varepsilon(i,j,d) = -RT \ln\left[\frac{N(i,j,d)_{obs}}{N(i,j,d)_{exp}}\right] \tag{2.2}$$

where $N(i,j,d)_{obs}$ is the number of observed contacts and $N(i,j,d)_{exp}$ is the number of expected contacts for those two atom types in that distance bin within the database of known structures. $N(i,j,d)_{exp}$ is defined as

$$N(i,j,d)_{exp} = N(d)X_iX_j \tag{2.3}$$

X_i and X_j are the mole fractions of the two atom types in the database and $N(d)$ is the total number of observed contacts in that distance bin over all atom type pairs.

Within each PMF, each pairwise energy curve was fitted to a quintic spline, and these atom type pairwise differentiable potentials were combined with the bonded terms of the ENCAD potential to form the three KB/MM potentials, named KB_0.1, KB_0.2, and KB_0.5 respectively for the width of the distance bins used in the generation their component PMFs. The KB/MM potentials were smoothly truncated to 0 kcal/mol between 9 and 11 Å.

2.2.3.3.2 Evaluating the Performance of the Potentials in Refinement

KB_0.1, KB_0.2, and KB_0.5 were tested against four MM potentials: AMBER99 [73], [74], OPLS-AA[75], GROMOS96 [76], and ENCAD [64], [77]. All seven potentials were tested on a dataset of 75 native protein structures, chosen to each represent a different fold from the Structural Classification of Proteins [78]. For each native, 729 NNSMs were generated using Tirion-style quasielastic normal mode perturbation [79]. An example of a native and a NNSM generated in this way is given in Figure 2.5. Structures are minimized *in vacuo* using the L-BFGS optimization protocol in either GROMACS[80]-[82] or ENCAD.

All potentials were evaluated based on two criteria: their ability to not significantly perturb the nativeand their ability to move NNSMs closer to the native state. Of the seven potentials, KB_0.1 was the best performing with respect to both criteria, followed in second place by AMBER99. For the first criteria, when applied to the natives, KB_0.1 resulted in a mean RMSD deviation of 0.38 ± 0.14 Å and AMBER99 a mean RMSD deviation 0.41 ± 0.20 Å. For the second criteria, performance was measured in the double mean RMSD over all 75 NNSM sets, with

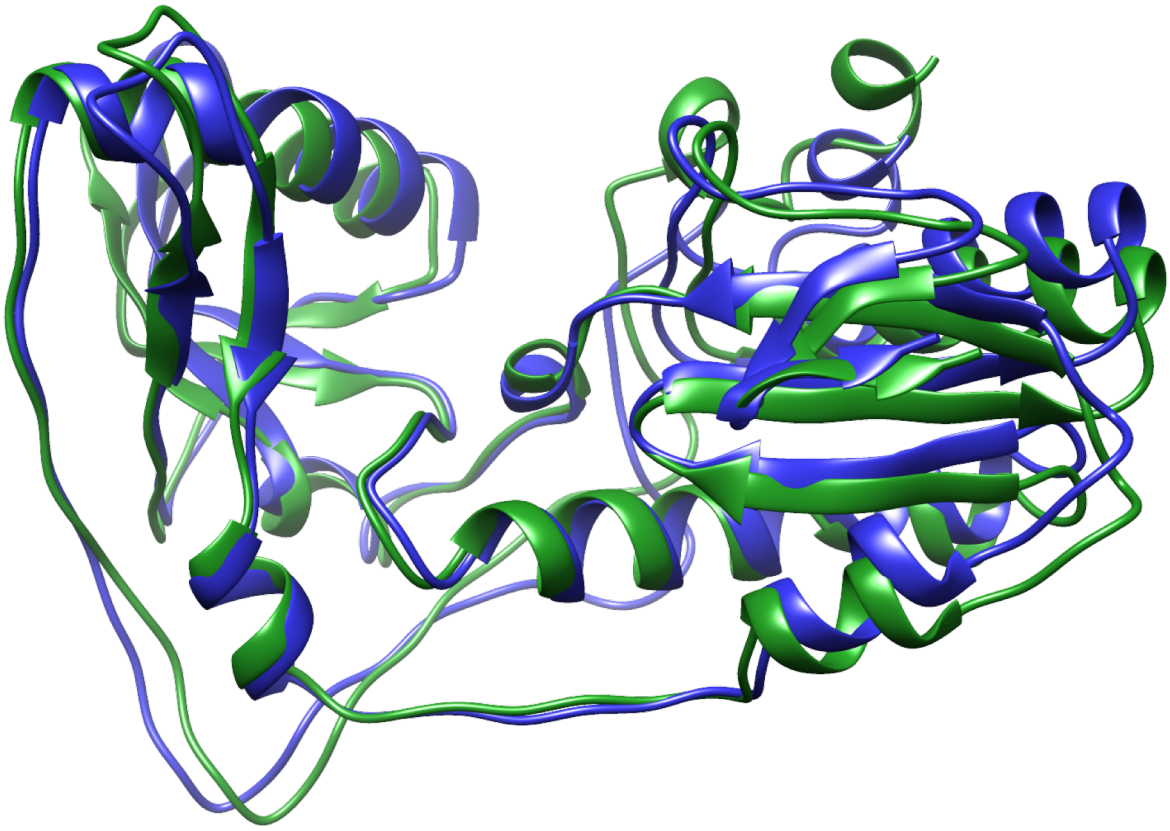

Figure 2.5: Native and a decoy generated via quasielastic normal mode perturbation. In green is the native 1mml and in blue is the decoy. The decoy's RMSD from the native is 2.75 Å.

<RMSD> indicating the mean RMSD of all 729 NNSMs from the native for a particular set and <<RMSD>> indicating the double mean over all 75 sets. Before minimization, <<RMSD>> was 1.06 Å. After minimization, AMBER99 resulted in a <<RMSD>> of 1.03 Å and KB_0.1 resulted in a <<RMSD>> of 0.95 Å, the best improvement of any tested potential.

2.2.3.3.3 Potential Avenues of Improvement in the Hybrid Potential

PMFs have widely adopted the atom type classification scheme of all 167 heavy atom types defined by the RAPDF potential. It is possible that a classification scheme consisting of all possible atom types is not optimal. Some atom types may share similar chemical and/or functional characteristics and defining them as separate type may be redundant. The structural database used to gather the

statistics for the PMF can also be examined. Since the publication of the hybrid KB/MM potential, the Richardson lab has curated a new structural database, an order of magnitude larger than the Top500 used to generate KB_0.1. It is possible that the greater statistics of a larger database can be leveraged to produce an improved potential for protein structure refinement.

2.2.3.3.4 Application of KB_0.1

The KB_0.1 has been used with success as part of a structure refinement protocol in CASP experiments [83] and is used in the KoBaMIN structure refinement web server.

2.3 Summary

An introduction to both protein structure and classification, and to protein folding techniques, including protein structure refinement has been given. Potential energy minimization has been presented along with a method for generating hybrid KB/MM potentials for use in structure refinement. This material serves as a foundation for Chapter 3 in which the formulation of the PMFs used in the hybrid KB/MM potentials is explored and for Chapter 4 in which a novel technique for comparing protein structures is proposed. This comparison technique allows for the exact identification of all regions of local similarity in a pair of structures even if components of secondary or tertiary structure are shifted relative to each other

Chapter 3

3. Refining the Hybrid KB/MM Potential for Potential Energy Minimization

- Exploring the formulation of the knowledge-based force fields

3.1 Introduction

The goal of protein structure predictors is to produce models as close to the true native protein structure as possible. Models can be produced through homology modeling, fold recognition (also known as protein threading), or *ab initio* techniques, and while protein structure predictors have become increasingly accurate, they have not yet reached the accuracy that can be achieved through empirical methods such as x-ray crystallography [15], [16]. The goal of protein structure refinement is to move models produced by protein structure predictors from their near native structure models (NNSM) as close as possible to the native structure (NS), defined as moving NNSMs to < 0.80 Å backbone Cα RMSD from the NS. As Eyal et al. show, 0.80 Å is the accuracy limit for structures determined through X-ray crystallography [84].

3.2 Towards Improving KB/MM Potentials for Protein Structure Refinement

Following the same method as Summa and Levitt [20], outlined in section 2.2.3.3.1, potentials for the purpose of *in vacuo* protein structure refinement using PEM are derived as hybrid KB/MM potentials with the first three terms of ENCAD's MM potential (Eq. 2.1), representing the energetic terms for the bonded interactions of the potential, and a differentiable KB potential representing the nonbonded interactions.

In pursuit of improving the hybrid KB/MM potential, the formulation of its KB component, its PMF, has been examined. Three main questions were asked. First, can the refinement performance of the hybrid potential be improved by selecting a larger starting structure database for the statistics of the PMF? The motivation for selecting a larger database is that it was noticed that the energy curves in the KB PMF portion of KB_0.1 were rough (Figure 3.1). A large database should provide a more robust set of statistics, allowing for smoother energy curves to be

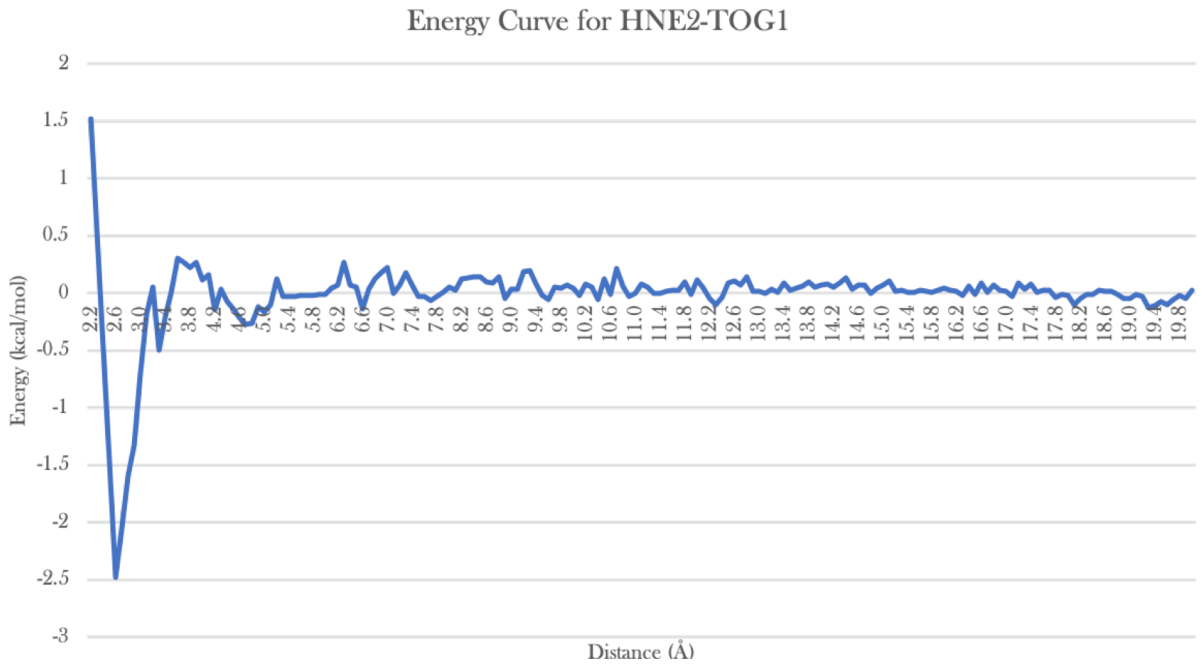

Figure 3.1: Energy curve for the atom type pair HNE2-TOG1. Generated from a database of 500 structures using Lu and Skolnick's formalism for the energy calculations. This energy curve is rough, possible causing PEM to get trapped in local minima that would not exist if a larger statistical database had been used.

generated which will in turn allow PEM to more easily navigate the energy surface to find the global minimum without getting trapped in local minima. Second, can performance be improved by using a stricter database? A stricter database should eliminate artifacts due to clashes in the structures. Third, are all 167 heavy atom types in the PMF required for optimal performance or can performance be improved by combining similar atom types to better leverage the statistics of both?

3.2.1 Generation of PMFs

3.2.1.1 Re-evaluating Low Distance Bin Counts

In a KB/MM potential, the purpose of the PMF is to evaluate the non-bonded interactions. A major goal in its derivation is to accurately represent the critical energetics of close contact interactions. To avoid taking the log of zero, in the statistics gathering phase for KB_0.1, all contact counts were initialized to one. This solved the problem of how to handle an undefined energy where zero counts are observed, but it had a side effect of lessening the energetic bonus for crucial close contact interactions such as hydrogen bonds and disulfide bonds.

Recall from Eq. 2.2 that Lu and Skolnick's energy calculation requires the ratio of the observed number of counts in a distance bin to the expected number of counts in that bin. While having a minimum count of one in a distance bin has a minimal effect on the number of observed counts, but it does have a cumulative effect on the number of expected counts since, as Eq. 2.3 shows, the calculation for the number of expected counts requires the sum of all counts in that distance bin across all atom type pairs. The ones across all distance bins add up, contributing to an artificially high expected value, lowering the energetic bonus for moving these atom type pairs to ideal distances, and affecting the performance of the potential. Figure 3.2 shows the difference in generated energy curves for the atom type pair NOD1-TOG1, where, if the minimum value for a

count is one, the energetic bonus for a desirable contact distance of 2.6 Å is eliminated. All PMFs generated in this work differ from KB_0.1 in that distance bins with no contacts use a count of zero in the energy calculation.

In the energy calculation, if the observed or expected number of counts in a distance bin is zero, the energy is set to zero. The repulsive close contact portion of the energy function starts at the furthest distance bin with no counts where there are no distance bins at smaller distances containing any counts. As a result, the majority of distances bins assigned an energy of zero will be replaced with the repulsive close contact portion of the energy function. It is important to note that it is possible for bins assigned an energy of zero to remain in the PMF. This is possible right

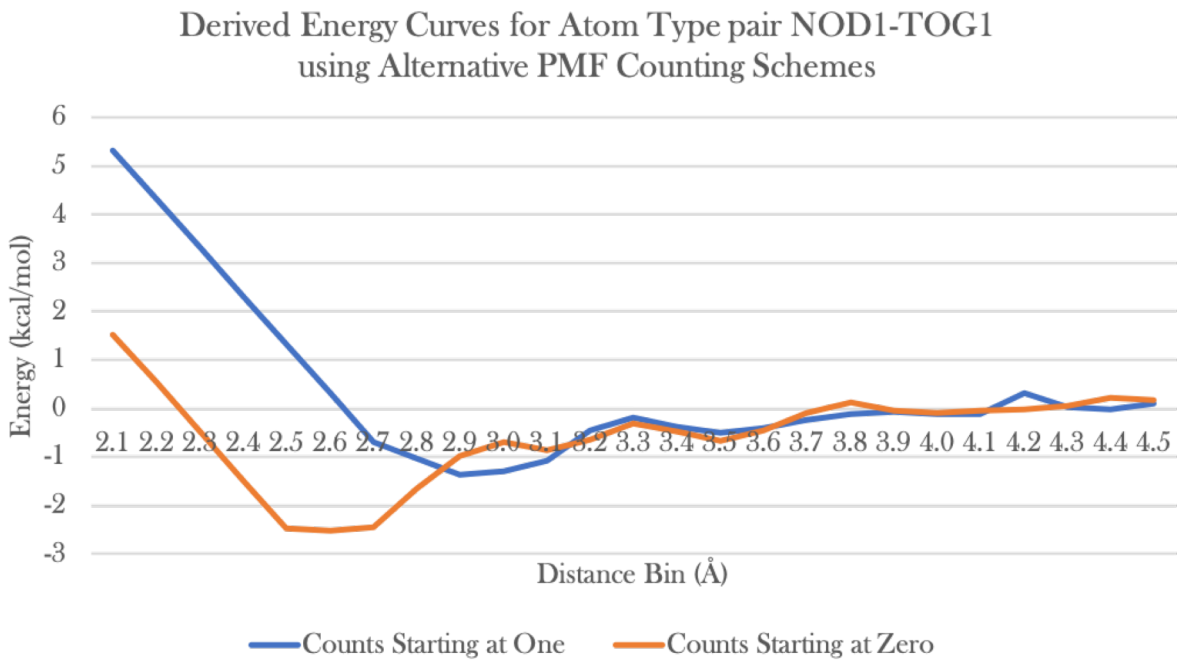

Figure 3.2: Energy curves for atom type pair NOD1-TOG1 derived using alternative counting schemes. If counts start at one, then the energetic bonus of a potential hydrogen bond is eliminated from NOD1-TOG1's energy curve.

outside of hydrogen bond lengths where there may be zero counts since hydrogen bonds are strong attractors.

3.2.1.2 Structure Databases for PMF Generation

Four different databases have been used in the generation of PMFs: Top500, Top8000, Top500_1.00vdw, and Top8000_1.00vdw. The first two databases, Top500 and Top8000 consisting of 500 and 7957 protein structures respectively, are from the Richardson lab. Hydrogens are built into all PDBs in both databases using the Reduce program [85]. Both of these databases apply filters to ensure that only high-quality structures are included. The Top500 database requires all structures have a resolution of 1.8Å or better, a clashscore [86] of < 22/1000 atoms, and < 10/1000 atoms with main chain bond angles outside of 5σ of Engh and Huber's parameters [87]. A structure's clashscore is defined as the number of "serious clashes". Serious clashes are non-hydrogen-bond van der Waals overlaps of 0.4 Å or greater per 1000 atoms. The Top8000 database is similar. It requires all structures have a resolution < 2.0 Å, a MolProbity [88], [89] score of < 2.0, ≤ 5% of residues with bond length or angle outliers of $> 4\sigma$, and ≤ 5% of residues with Cß deviation outliers of > 0.25Å. The MolProbity score includes a structure's clashscore as a component of the overall score.

The Top500 database is the database that was used to generate KB_0.1. The Top8000 database, released after the publication of KB_0.1, consists of an order of magnitude more structures than Top500. The Top8000 database is included in this work to test the hypothesis that a larger database providing a larger set of statistics will generate a PMF with smoother pairwise energy curves. This should allow the minimization process, via numerical optimization, to better find the global minimum by not getting trapped in the local minima of the potential.

The other two databases, Top500_1.00vdw and Top8000_1.00vdw, are subsets of the Top500 and Top8000 databases. While both the Top500 and Top8000 databases are strict on the structures that they allow to be included, both allow serious clashes to be included if their proportion is small. This is an important consideration when designing a PMF. In a PMF, any count from a clash at a low distance, where there are no other interactions, will introduce an unnatural artifact into the resulting energy curve. Figure 3.3 shows the difference between the energy curves generated for the atom pair YCD2-YCD2 using the Top500 database and that same database filtered for clashes greater than 1.00 Å. YCD2 indicates the second carbon delta of tyrosine and the curves represent the interaction energy for this atom type pair. The energy curve for this atom type pair generated from the Top500 database has a dip in energy at 2.5Å due to a

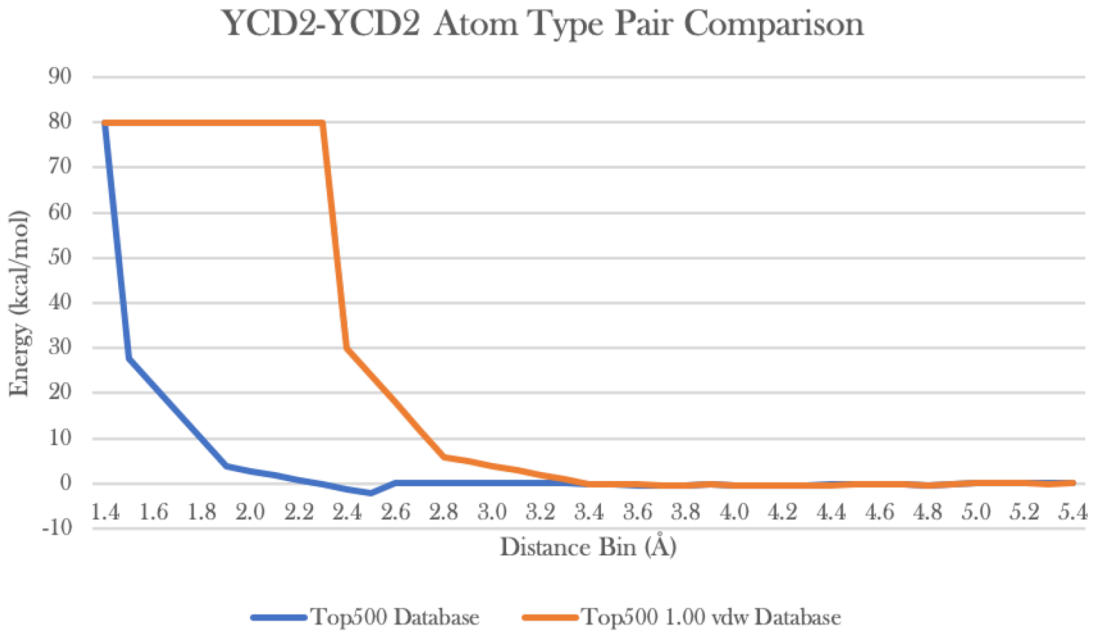

Figure 3.3: The effect of eliminating clashes from structure databases on PMF energy curves. A comparison of the energy curves for atom pair YCD2-YCD2 as generated from the Top500 database and that same database filtered for structures with non-bonded clashes greater that 1.00 Å vdw overlap. The dip in energy in the Top500 curve at 2.5 Å is due to a clash between residues 146 and 151 in structure 1a6mH.pdb, adding an unnatural attractor basin to this energy curve.

single count originating from a clash between residues 146 and 151 in structure 1a6mH.pdb, resulting in an unnatural energy basin at that distance in this energy curve.

To test whether removing these artifacts would improve the performance of the resulting MM/KB potentials, databases Top500_1.00vdw and Top8000_1.00vdw have been generated from the original Top500 and Top8000 databases by removing all structures with non-bonded heavy atom clashes ≥ 1.00 Å van der Waal overlap. Formally, two non-bonded atoms a and b are clashing if their distance $d(a,b) \leq (a_r + b_r) - 1.00$, where a_r and b_r are standard atomic radii for atoms a and b, and atoms are non-bonded if they are separated by five or more covalent bonds and are neither involved in hydrogen nor disulfide bonds. Values for standard atomic radii have been pulled from ENCAD's van der Waal potential's parameters: $C_r = 1.85$ Å, $N_r = 1.65$ Å, $O_r = 1.60$ Å, $S_r = 1.85$ Å. The Top500_1.00vdw and Top8000_1.00vdw databases consist of 449 and 7489 structures respectively.

PMFs were generated for each of the four structure databases, listed in Table 3.1 with the number of structures and pairwise interactions used in PMF generation given. With the exception of the structure databases, the parameters and procedure for the generation of the PMFs are identical. Pairwise interactions are classified using 167 residue-specific heavy atom types, each PMF uses a cutoff distance of 20.0 Å with bins of width 0.1 Å, contacts between atoms in adjacent residues in the sequence are omitted, the Lu and Skolnick formalism is used for the calculation of

Table 3.1: The four PMFs generated from the four structure databases: Top500, Top500_1.00vdw, Top8000, and Top8000_1.00vdw. The number structures and pairwise interactions used in the generation of each PMF is given.

	KB_Top500	KB_Top500_1.00vdw	KB_Top8000	KB_Top8000_1.00vdw
Structures	500	449	7957	7489
Pairwise Interactions	280,653,907	251,680,166	4,809,056,116	4,479,873,427

the energies, and a repulsive close contact portion is added that scales to a plateau of 80 kCal/mol. Since the only difference between the PMFs is the structure database, the four PMFs are named after their databases: KB_Top500, KB_Top8000, KB_Top500_1.00vdw, and KB_Top8000_1.00vdw.

3.2.1.3 Reducing the Set of Atom Types via an Atom Type Merging Process

Samudrala and Moult showed when developing their residue-specific all-atom probability discriminatory function (RAPDF), that of the three PMFs they tested, the best performing was that using 167 residue-specific heavy atom types[90]. Other schemas for classifying atom-pair interactions included a residue-specific virtual atom representation (where groups of atoms within the same residue are combined into 105 virtual atoms) and a non-residue-specific virtual atom representation (where all possible heavy atom types across all residues are combined into 21 virtual atoms). They found that the detail inherent in using all 167 possible residue-specific heavy atoms allowed the RAPDF to be the most accurate native structure discriminator among the three they tested and that their successive atom type approximations of residue-specific virtual atoms and non-residue-specific virtual atoms yielded successively worse performance. In their distance dependent knowledge-based potential, Lu and Skolnick use the 167 residue-specific atom types, but they discuss the idea of grouping similar atom types based on chemical and functional similarities. They discuss, for example, grouping EOE1 and DOD1 into a combined atom type, but they do not propose or test a set of groupings [71].

This work examines in detail the question broached by Lu and Skolnick [71]. Is the classification of atoms into all 167 residue-specific heavy atom types optimal for the performance of the PMF, or are there atom type groupings that can improve performance by leveraging the statistics of one or more atom types? To examine this question systematically, an iterative approach

was taken to identify chemically and functionally similar atom types and merge them into combined atom types at varying levels of similarity.

To identify similar atom types in a PMF, their energy profiles are compared. A single energy profile consists of the interaction energies for each distance bin of a single pair of atom types. Figure 3.2 shows the energy profile for the AN-ACB atom type pair. Each atom type has a set of energy profiles, one for every atom type, including itself, in the PMF. Two atom types that have similar functional characteristics will have a similar set of energy profiles. Given two atom types *a* and *b*, their similarity is defined as the average RMSD of their energy profiles.

$$Similarity(a,\ b) = \frac{1}{N}\sum_{i}^{N} RMSD\big(\varepsilon(a, c_i), \varepsilon(b, c_i)\big) \tag{3.1}$$

N is the number of atom types, and $\varepsilon(a, c_i)$ is the energy profile for atom type *a* and atom type c_i in the set of atom types.

An iterative procedure is used to generate a set of PMFs, based off of some base PMF, containing combined atom types for successively looser thresholds of similarity. Given a PMF and a similarity threshold *t*, all atom type pairs whose similarity is less than *t* are combined into merged atom types, and given a set of thresholds *T*, for each *t* in *T*, a PMF containing merged atom types is generated and then used as the starting PMF for the next threshold. The starting PMF for the first threshold is the base PMF. Two atom types are combined into a single merged atom type by summing their counts across all their bins and then generating an energy curve for these combined counts. This new energy curve now represents both atom types in the combine type. At the end of a single atom type merge, the PMF has one less atom type and each atom type has one less pairwise interaction. These procedures are given in Figure 3.4.

The thresholds for merging are determined empirically for each base PMF. A starting

threshold is chosen such that at least one pair of atom types will be merged, and an increment is selected such that not too many pairs of atom types will be merged in a single iteration. The goal is to end up with a set of PMFs spanning the space of reasonable atom type merges. If no atom types were merged in an iteration, no new PMF is produced and the same starting PMF is passed on to the next iteration.

Each of the four generated PMFs, KB_Top500, KB_Top8000, KB_Top500_1.00vdw, and KB_Top8000_1.00vdw, were subjected to the merging process, producing a new set of PMFs for each original PMF. The starting thresholds and threshold increment for each of these merging processes are given in Table 3.2.

```
Algorithm 1: generateMergedAtomTypePMFs(𝒫)
Input: A Potential of Mean Force 𝒫
Output: A set of PMFs 𝒫_t, one for each
threshold t in
 1  begin
 2  for t in T
 3     𝒫_t = identifyAndMergeAtomTypes(𝒫_t-1, t)
 4  end
```

```
Algorithm 2: identifyAndMergeAtomTypes(𝒫, t)
Input: A Potential of Mean Force 𝒫, and a
similarity threshold t.
Output: A new Potential of Mean Force 𝒫′ where
atom types whose similarity is less than t have
been merged.
 1  begin
 2   𝒫′ = copy(𝒫)
 3   while atomTypesMerged == true
 4       atomTypesMerged = false
 5       for every pair of atom types a,b in 𝒫
 6           if similarity(a,b) < t
 7               merge a and b
 8               atomTypesMerged = true
 9   return 𝒫′
 10 end
```

Figure 3.4: Atom Type Merging Algorithms. They iteratively generate a set of PMFs whose similar atom types have been merged. For each iteration, every pair of atom types whose similarity (as defined in eq. 4) is less than a threshold *t* are merged. The PMF generated at the end of one iteration is used as the starting PMF for the next iteration.

Table 3.2: Similarity thresholds and increments for the atom type merging process. Given for each of the original PMFs: Top500, Top500_1.00vdw, Top8000, and Top8000_1.00vdw.

	KB_Top500	KB_Top500_1.00vdw	KB_Top8000	KB_Top8000_1.00vdw
Initial Similarity Threshold (kcal/mol)	2.58	2.45	1.80	1.85
Threshold increment	0.01	0.01	0.05	0.05

3.2.2 Methods for Evaluating the Performance of the Potentials

3.2.2.1 The Refinement Protocol

Protein structures are refined via *in vacuo* PEM. The L-BFGS minimizer in ENCAD is used, running for 10,000 steps of minimization or until energy convergence to machine precision.

3.2.2.2 Evaluation Criteria

To test the performance of each PMF, a hybrid KB/MM potentials is generated with it as the KB component and that potential is evaluated in protein structure refinement against two criteria.

1. Refinement should not significantly perturb the native.
2. Refinement should move NNSMs closer to the native.

The first criterion ensures that the potential has an energy well at the native. For a concrete criterion, a potential should not move the native by > 0.80 Å RMSD since that is the threshold by which natives are indistinguishable from each other in x-ray crystallography experiments [84].

In evaluating a potential's ability to move NNSMs closer to the native, the following notation will be used. Given a dataset consisting of natives and a set of NNSMs for each, the mean RMSD of a set of NNSMs with respect to the native will be denoted as <rmsd>. The average of

<rmsd> over all sets of NNSMs in the complete dataset will be denoted as <<rmsd>>. It is useful to calculate the percent improvement of a refinement process on a set of NNSMs.

$$PI = \frac{<rmsd>_{min} - <rmsd>_{start}}{<rmsd>_{start}} \quad (3.2)$$

A negative PI indicates improvement.

3.2.2.3 Structure Datasets for Testing

Two datasets were used for testing purposes. The first is a decoy dataset generated using the same method as outlined by Summa and Levitt [20] and the second is a selection of targets and submitted models from CASP experiments 8-13. The decoy dataset consists of 71 natives, selected to be representative of the SCOP [78] folds, each with a set of decoys generated by perturbing the natives using the method of Tirion [79], yielding a total of 21519 decoys in the dataset. This dataset is not identical to the dataset used by Summa and Levitt, but it was regenerated using quasi-elastic normal mode perturbation as was the origin set. Of the 75 original natives, four were not used in this study due to minimization errors. The four omitted natives are 1cem00, 1fh2, 1ge8a01, and 1kfn_3. For this decoy dataset, <<rmsd>> before minimization is 1.872 ± 0.223 Å.

The CASP dataset was built as follows. All submitted models and natives for the split domain regular targets for CASP experiments 8 – 13 were downloaded as a starting dataset. Then all models whose RMSD from the native were less than 0.50 Å or greater than 5.00 Å were removed from the dataset. This was done because the focus of this work is on the performance of the potentials as near-native structure minimizers. Starting models that are too close or too far from the native do not fall in the experimental test case. Finally, all target sets with 100 or more remaining structures were selected as the testing dataset. The CASP dataset consists of 234 natives with a total of 59,527 models. For the CASP dataset, <<rmsd> is 2.951 ± 0.847 Å.

Both datasets serve a purpose in the evaluation of the potentials. The decoy dataset was generated to be a general test of structure refinement ability by being representative of a diverse set of folds. A potential that is suitable for general structure refinement should perform well across the whole of the dataset as opposed to working well for some types of folds but not others. It was also generated to specifically provide a set of near native structures. As a result of the method of generation, quasielastic normal mode perturbation [20], the decoys should be in an energetically accessible conformation with respect to minimization back to the native. That is, there should be no serious energy barriers caused by side-chain packing issues or grossly misfolded conformations of a structure.

The CASP dataset was selected as a real-world test of structure refinement. Structure refinement is performed after a model is generated, whether it is generated via homology modelling, protein threading, or ab initio techniques, and these models may have energetic barriers between them and the native. In an ideal world, structure refinement would only be performed on structures close to the native and somewhere on an energetic pathway to the native, but in practice this cannot be guaranteed, and the CASP dataset provides a realistic set of models that are provided as input to a refinement process. Of the two datasets, the CASP dataset is the more difficult test for structure refinement.

3.3 Results

3.3.1 Atom Type Merging Process

The atom type merging process resulted in a set of PMFs generated from each original PMF giving, for each, a set of potentials spanning the range from using the full 167 atom types to using approximately 100 atom types. A total of 61 PMFs were generated. A list of these PMFs is given in Table 3.3. The difference in the number of PMFs produced for KB_Top500 and

Table 3.3: PMFs generated via the atom type merging process. This process was applied to each original PMF. The similarity threshold used to generate each merged atom types PMF and the number of atom types in that PMF are listed.

KB_Top500		KB_Top500_1.00vdw		KB_Top8000		KB_Top8000_1.00vdw	
Threshold	# Atom Types	Threshold	# Atom Types	Threshold	# Atom Types	Threshold	# Atom Types
2.58	165	2.45	166	1.80	166	1.85	164
2.61	164	2.51	165	1.85	165	1.90	163
2.63	159	2.59	162	1.95	161	1.95	159
2.67	154	2.63	161	2.00	160	2.00	158
2.70	150	2.64	156	2.05	155	2.05	154
2.71	148	2.70	148	2.10	150	2.10	147
2.73	147	2.73	142	2.15	146	2.15	137
2.76	146	2.78	140	2.20	139	2.20	133
2.78	140	2.80	139	2.25	117	2.25	124
2.80	136	2.82	137	2.30	115	2.30	117
2.90	134	2.87	126	2.35	107	2.35	105
2.91	131	2.96	123	2.40	104	2.40	104
2.93	130	3.01	109				
2.94	127	3.04	107				
2.96	125	3.06	105				
2.98	124	3.09	103				
2.99	123	3.14	97				
3.00	122						
3.02	121						
3.03	100						

KB_Top500_1.00vdw is a result of when the various atom types were combined. Not every threshold resulted in atom type merges and the different energy curves between those in KB_Top500 and KB_Top500_1.00vdw resulted in atom type combinations clustering at different thresholds.

3.3.1.1 Merged Atom Types

It is important to ask whether or not the atom type merges are reasonable. Tables 3.4, 3.5, 3.6, and 3.7 give the atom type combinations resulting from the merging processes for each base PMF: KB_Top500, KB_Top500_1.00vdw, KB_Top8000, KB_Top8000_1.00vdw.

For KB_Top500, the first atom types merged into combined atom types are the hydroxyl groups of serine and threonine, and the backbone oxygens of threonine and lysine. Both

combinations make chemical and functional sense. Atom types of the same element and position in the amino acid tend to be combined. For example, carbons at the α and β positions tend to be merged. Likewise, backbone oxygen atoms are commonly merged. For the merge process run on KB_Top500, by threshold 2.91, the backbone oxygens of thirteen of the amino acids have been combined into a single type, suggesting that distinguishing between the majority of the backbone oxygens may not be important in a PMF.

Similar patterns are visible in the rest of the tables. Atom types tend to be merged by element and position in the amino acid. Backbone atoms of the same element tend to group together. Likewise, carbon atoms from hydrophobic residues tend to be combined. Their similarity is evidence of both the importance of the hydrophobic effect and of these KB potentials' ability to implicitly characterize it. Another notable combination is that of the aromatic carbons of phenylalanine with those of tyrosine, a combination that happens in all four merge processes. Given their chemical similarity, this combination is a good sign that the merging process is correctly identifying and combining similar atom types. Complete graphs generated using the open source program GRAPHVIZ [91] of all atom type merges for the Top500 and Top500_1.00vdw PMFs are given in the appendix.

Table 3.4: Results of the atom type merging process on KB_Top500. Merges at later iterations encompass those from earlier iterations. E.g., at threshold 2.61 atom types CO and MO are combined, and at threshold 2.73, that combined type is merged with atom type FO to form a combined atom type representing CO, MO, and FO.

Threshold	Atom Types Merges
2.58	SOG, TOG1 TO, KO
2.61	CO, MO
2.63	LCA, FCA AO, LO VO, YO VCG1, LCD2 FCE1, YCE2
2.67	FCD1, FCD2, FCE2, LCD2, VCG1 QN, RN, KN
2.70	EC, QC, LC AO, LO, DO, TO, KO
2.71	NO, QO, SO
2.73	CO, MO, FO
2.76	VCG1, LCD2, FCD1, FCE1, FCE2, FCD2, YCE2
2.78	VCA, ICA, RCA YCD1, YCE1 ACA, LCB AC, LC, EC, QC, TC
2.80	LCA, FCA, SCA, TCA VO, CO, MO, FO, YO, EO
2.90	LCA, PCA, FCA, SCA, TCA LCD1, FCZ
2.91	AO, VO, LO, CO, MO, FO, YO, NO, EO, QO, SO, TO, KO DCA, NCA
2.93	LCD1, ICD1, FCZ
2.94	ACB, VCG1, LCD2, FCD1, FCE1, FCE2, FCD2, YCE2, HCE1, GCA
2.96	ACA, LCA, LCB, PCA, FCA, SCA, TCA AN, QN, RN, KN
2.98	VN, IN
2.99	IO, RO
3.00	NN, EN
3.02	VCG2, TCG2
3.03	AC, LC, EC, QC, SC, TC VCB, LCG, ICG1, MCG, FCB, NCB IO, PO, HO, RO WCZ2, WCH2 VC, FC, KC YCB, RCG ACA, VCB, LCA, LCB, LCG, ICG1, MCG, PCA, PCB, FCA, FCB, NCB, SCA, SCB, TCA DC, NC DCA, NCA, ECA, TCB KCG, KCD LN, FN, YN

Table 3.5: Results of the atom type merging process on KB_Top500_1.00vdw.

Threshold	Atom Types Merges
2.45	SOG, TOG1
2.51	TO, KO
2.59	VCG1, LCD2 FO, YO LCA, FCA
2.63	SO, GO
2.64	AO, VO, LO, FO, YO VN, IN NO, QO
2.70	FCE1, FCD2, YCD1, YCE2, YCD2 ACA, SCA LN, NN, RN LCD1, ICD1
2.73	ACB, VCG1, VCG2, LCD2, ICG2, FCD1, FCE1, FCE2, FCD2, YCD1, YCE2, YCD2 ACB, VCG2
2.78	SO, TO, KO, GO AO, VO, LO, FO, YO, SO, TO, KO, GO
2.80	LCA, FCA, KCA
2.82	DO, NO, QO FC, RC
2.87	AC, LC, FC, TC, RC VCA, LCA, ICA, FCA, RCA, KCA VC, NC, EC IO, CO, MO, EO
2.96	LCD1, ICD1, FCZ, YCE1 [ACB, VCG1, VCG2, LCD1, LCD2, ICG2, ICD1, FCD1, FCE1, FCZ, FCE2, FCD2, YCD1, YCE1, YCE2, YCD2]
3.01	ACA, LCB, PCB, NCB, SCA HCE1, GCA LN, NN, SN, RN VCA, LCA, ICA, FCA, DCA, ECA, RCA, KCA DOD1, DOD2 AC, VC, LC, FC, DC, NC, EC, TC, RC AO, VO, LO, IO, CO, MO, PO, FO, YO, EO, SO, TO, RO, KO, GO YCA, QCA
3.04	VCA, LCA, ICA, PCA, FCA, YCA, DCA, ECA, QCA, RCA, KCA
3.06	HCA, TCA VN, IN, TN
3.09	HCE1, TCG2, GCA VCA, LCA, ICA, PCA, FCA, YCA, DCA, ECA, ECB, QCA, RCA, KCA
3.14	YOH, SOG, TOG1 QCG, KCB KCG KCD ACA, LCB, PCB, NCB, SCA, SCB QC, HC LCG, WCB

Table 3.6: Results of the atom type merging process on KB_Top8000.

Threshold	Atom Types Merges
1.80	FCE1, YCE1
1.85	FCE1, YCE1, YCE2
1.95	FCD2, YCD1 FCB YCB NCG, GCA VCA, ICA
2.00	FCD2, YCD1, YCD2
2.05	ACA, SCA FN, RN FCE2, WCZ2, WCH2 NO, KO
2.10	LN, FN, RN DC, SC TCA, RCA VCA, LCA, ICA VO, GO
2.15	VCA, LCA, ICA, QCA, KCA FCE1, FCZ, YCE1, YCE2 NCA, TCA, RCA
2.20	ACA, VCA, LCA, ICA, PCA, QCA, SCA, KCA YOH, TOG1 AO, QO, RO VO, LO, GO FCB, YCB, HCB
2.25	AC, DC, SC AO, MO, FO, WO, HO, RO PCB, PCG WCZ3, WCE3 VO, LO, DO, EO, SO, GO LN, FN, NN, RN, KN FCG, YCG HND1, HNE2 LCB, RCB LC, FC, EC, RC, KC FCE1, FCZ, FCE2, YCE1, YCE2, WCZ2, WCH2 YOH, SOG, TOG1
2.30	FCD1, FCD2, YCD1, YCD2 PO, TO
2.35	VCB, LCG PCB, PCG, DCB, NCB, NCG, ECB, HCD2, SCB, GCA NCG, SCB, GCA ICG2, FCD1, FCD2, YCD1, YCD2
2.40	EN, TN ACA, VCA, VCB, LCA, LCG, ICA, PCA, QCA, SCA, KCA IN, YN

Table 3.7: Results of the atom type merging process on KB_Top8000_1.00vdw.

Threshold	Atom Types Merges
1.85	FCE1, YCE2 FCD2, YCD1, YCD2
1.90	FCE1, YCE1, YCE2
1.95	FCB YCB VCA, QCA SCB, GCA DC, SC
2.00	YOH, TOG1
2.05	FCE1, FCZ, FCE2, YCE1, YCE2 NO, RO AO, LO
2.10	VCA, QCA, KCA WCZ2, WCH2 AO, LO, NO, RO, GO ACA, SCA, TCA, RCA
2.15	ACB, FCD2, YCD1, YCD2 ACA, VCA, PCA, ECA, QCA, SCA, TCA, RCA, KCA ICG1, FCB, YCB MO, WO, HO LN, FN, RN WCZ3 WCE3
2.20	MO, FO, WO, HO AC, LC, FC DC, SC
2.25	HND1, HNE2 HCD2, SCB, GCA SO, TO ACB, FCD2, YCD1, YCE2 AO, VO, LO, MO, FO, WO, NO, HO, RO, KO, GO FCD1, WCZ2, WCH2 DCB, ECB HCE1, HCD2, SCB, GCA
2.30	ACB, FCE1, FCZ, FCE2, FCD2, YCD1, YCE1, YCE2, YCD2 LCD1, LCD2 YOH, SOG, TOG1 AO, VO, LO, MO, FO, WO, DO, NO, EO, HO, SO, TO, RO, KO, GO DCG, ECD
2.35	ACA, VCA, VCB, LCB, LCG, PCA, DCB, ECA, ECB, QCA, SCA, TCA, RCA, KCA PCB, PCG YCG, WCG HCB, KCB FCD1, WCZ2, WCH2, WCZ3, WCE3 AC, LC, FC, YC, DC, SC, TC, RC, KC
2.40	ACB, FCE1, FCZ, FCE2, FCD2, YCD1, YCE1, YCE2, YCD2, HCE1, HCD2, SCB, GCA

3.3.2 Performance of the Generated Hybrid MM/KB Potentials in PEM

Each generated potential has been applied in PEM on both the CASP and decoy datasets. The performance with respect to the two criteria (Section 3.2.2.2) of all generated potentials is compared against KB_0.1. The performance of the base four potentials, KB_Top500, KB_Top500_1.00vdw, KB_Top8000, and KB_Top8000_1.00vdw will be presented, followed by the performance of the merged atom types PMFs derived from them. The impact of the starting database selection and of the atom type merging process will be presented.

3.3.2.1 The Baseline: KB_0.1's Performance

Before discussing any modifications to the hybrid KB/MM potential, a control must first be established. The decoy dataset has a starting <<rmsd>> of 1.872 Å. For this dataset, KB_0.1 improves <<rmsd>> to 1.637 Å. Its mean PI over all of the decoy sets is -12.69%. When applied to the natives, KB_0.1 results in a mean perturbation of 0.36 ± 0.12 Å. It is useful to evaluate potentials based on their ability to refine structures at various distances from the native. Figure 3.5 shows KB_0.1's performance in PEM on sets of NNSMs in the decoy dataset that fall in increasing ranges of starting RMSD from the native.

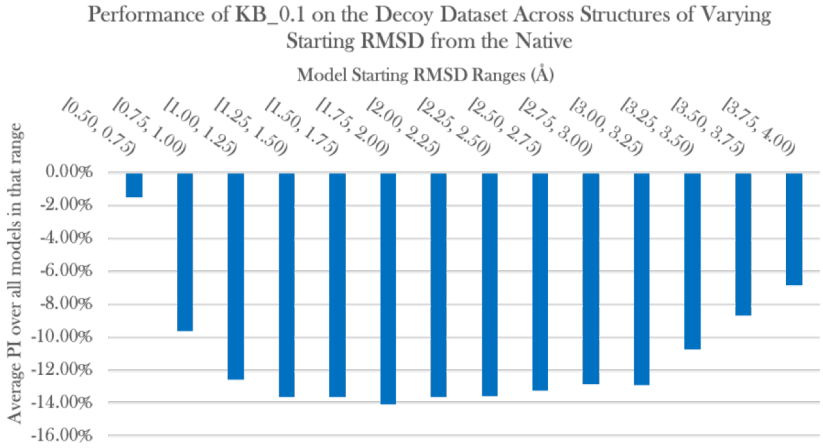

Figure 3.5: KB_0.1's ability to minimize the decoy dataset relative to starting RMSD from native. Performance measured by average PI on the models that fall within each RMSD range.

The CASP dataset has a starting <<rmsd>> of 2.951 Å. KB_0.1 improves <<rmsd>> to 2.918 Å. Its mean PI over all CASP target sets is -1.78%, and when applied to the native, it perturbs them by an average of 0.39 ± 0.20 Å. KB_0.1's performance as a function of starting model RMSD is given in Figure 3.6. The number of models in each starting RMSD range for both datasets is given in Table 3.9.

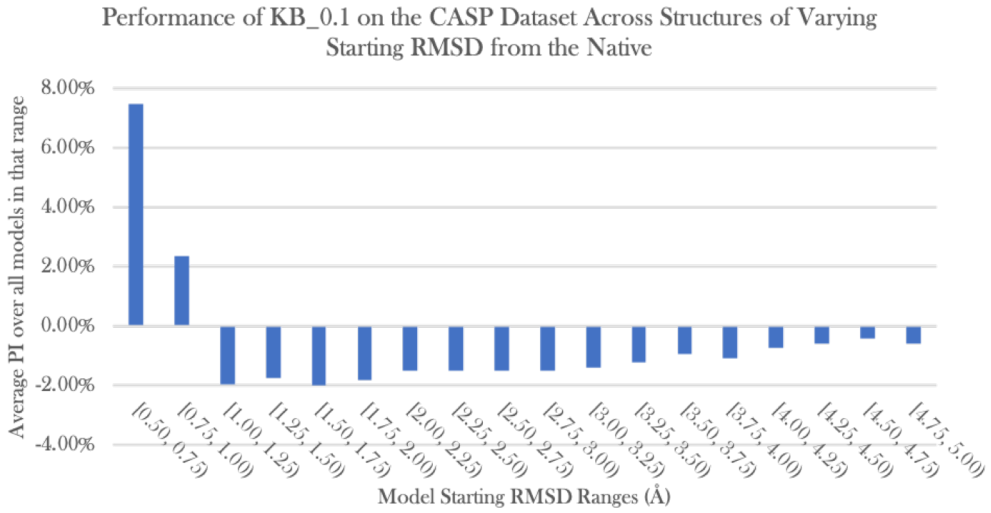

Figure 3.6: KB_0.1's ability to minimize the CASP dataset relative to starting RMSD from native. Performance measured by average PI on the models that fall within each RMSD range.

3.3.2.2 The Performance of the KB_Top500, KB_Top500_1.00vdw, KB_Top8000, and KB_Top8000_1.00vdw Potentials

The four KB potentials generated from the different starting database were evaluated based on their performance according to both evaluation criteria and were compared against KB_0.1. Table 3.8 summarizes the results. A couple of factors are immediately noticeable. First, choosing a larger starting database does not increase performance. By both evaluation criteria, the Top8000 potentials perform worse. First, they significantly alter the atomic coordinates of the native structures, greater than the criterion tolerance threshold of 0.80 Å RMSD. They also perform

Table 3.8: Performance summary of KB_0.1 and four base PMFs. Native perturbations are given as the average RMSD (Å) with standard deviation over all natives in the set. Mean PI is the average of a potential's PI across all of a dataset's decoy or model sets. Starting <<rmsd>> of the Decoy Dataset is 1.87 ± 0.22 Å and of the CASP Dataset is 2.95 ± 0.85 Å.

	Decoy Dataset			CASP Dataset		
	Native Perturbation	Mean PI	<<rmsd>>	Native Perturbation	Mean PI	<<rmsd>
KB_0.1	0.36 ± 0.12	-12.69 %	1.637	0.39 ± 0.20	-1.18 %	2.918
KB_Top500	0.34 ± 0.11	-12.89 %	1.633	0.38 ± 0.14	-1.26 %	2.916
KB_Top500_1.00vdw	0.34 ± 0.14	-12.50 %	1.640	0.53 ± 0.15	-1.21 %	2.917
KB_Top8000	0.92 ± 0.34	-8.52 %	1.715	0.97 ± 0.35	1.81 %	2.980
KB_Top8000_1.00vdw	0.94 ± 0.36	-8.89 %	1.708	0.96 ± 0.33	1.66 %	2.976

worse overall in minimization. Over the CASP dataset, their use results in a net degradation of model quality. Reasons for its performance will be discussed. The performance differences between KB_0.1 and both Top500 potentials is slight. KB_Top500 is the best performer across both datasets, but its advantage, 0.004Å and 0.002 Å is so slight that it is negligible. For the Top500 PMFs, eliminating clashes did not improve their performance, although for the Top8000 PMFs, it did, but once again, by a small amount. Comparing all five potentials with respect to their performance as minimizers of models at varying levels of starting RMSD to the native shows an interesting trend (Figure 3.7). On the decoy dataset, while the

Table 3.9: Decoy and CASP Dataset model counts and starting RMSD distribution.

Starting RMSD Range	# NNSMs	
	Decoy Dataset	CASP Dataset
[0.50, 0.75)	2006	623
[0.75, 1.00)	2048	1355
[1.00, 1.25)	2042	2039
[1.25, 1.50)	2000	2426
[1.50, 1.75)	1962	3474
[1.75, 2.00)	1936	3686
[2.00, 2.25)	1866	4354
[2.25, 2.50)	1826	4506
[2.50, 2.75)	1713	5097
[2.75, 3.00)	1514	4751
[3.00, 3.25)	1163	4668
[3.25, 3.50)	674	4285
[3.50, 3.75)	338	3882
[3.75, 4.00)	189	3174
[4.00, 4.25)	107	2607
[4.25, 4.50)	67	2391
[4.50, 4.75)	34	2004
[4.75, 5.00)	23	1839

Top8000 potentials degrade models at low starting RMSD, they outperform the other potentials at higher levels of starting RMSD.

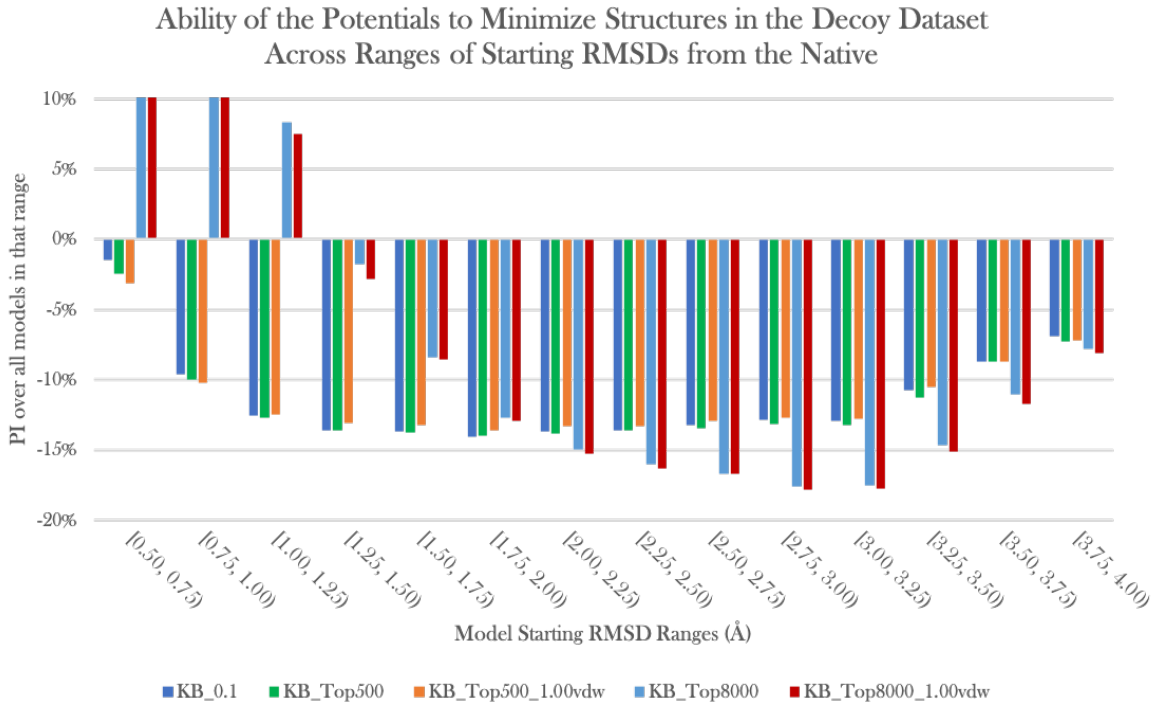

Figure 3.7: The performance KB_0.1 and four base potentials in minimization with respect to model starting RMSD. The chart has been truncated to a maximum PI of 10%. The PI values for KB_Top8000 and KB_Top8000_1.00dw for the first and second bins are 65% and 26%, and 63% and 26% respectively. The Top8000 potentials perform worse than the other potentials for models close to the native but outperform the other potentials when the model is further from the native. Ranges above 4.00 Å were omitted due to a lack of decoys at those distances.

3.3.2.3 Performance of the Merged Atom Types Potentials

All 61 potentials generated from the atom type merging process (listed in Table 3.3) were evaluated against both criteria. Figures 3.9, 3.10, 3.11, and 3.12 give the native perturbation and performance in minimization on the decoy dataset of the potentials using the merged atom types PMFs generated from KB_Top500, KB_Top500_1.00vdw, KB_Top8000, and KB_Top8000_1.00vdw respectively. First, it can be noted that potentials containing merged atom types PMFs derived from the KB_Top500 performed better than KB_0.1 and KB_Top500 in structure refinement (Figures 3.9 top and 3.10 top). They also perturbed the natives more (Figures

3.9 bottom and 3.10 bottom). The maximum mean RMSD over the minimized natives was 0.50 ± 0.17 Å (KB_Top500_1.00vdw_3.14), still well within the acceptable tolerance of 0.80 Å.

As expected, based on the performance of KB_Top8000 and KB_Top8000_1.00vdw, the merged atom types potentials generated from these PMFs did not perform well by either criterion. They significantly perturbed the natives, and they performed worse than KB_0.1 for minimization. Merging atom types for the Top8000 PMFs did not result in improvement in PEM for structure refinement.

Applied to the decoy dataset, of all potentials tested, the best performing is KB_Top500_2.98 with 124 atom types. Figure 3.8 gives its list of combined atom types. Its mean deviation in RMSD of the natives is 0.44 ± 0.14 Å, and it minimized the structures in this dataset to a <<rmsd>> of 1.617 Å from the starting <<rmsd>> of 1.872 Å, an improvement in <<rmsd>> of 0.02 Ã over KB_0.1 (<<rmsd>> = 1.637 Å). Figure 3.14 compares this potential against KB_Top500 and KB_0.1 in PEM of each of the sets in the decoy dataset and shows that, as expected, KB_Top500_2.98 outperforms both KB_Top500 and KB_0.1 as a structure minimizer.

- AO VO LO CO MO FO YO DO NO EO QO SO TO KO
- AC LC EC QC TC
- AN QN RN KN
- VN IN
- ACA LCA LCB PCA FCA SCA TCA
- VCA ICA RCA
- DCA NCA
- LCD1 ICD1 FCZ
- YCD1 YCE1
- ACB VCG1 LCD2 FCD1 FCE1 FCE2 FCD2 YCE2 HCE1 GCA
- SOG TOG1

Figure 3.8: The combined atom types in KB_Top500_2.98. This potential contains 13 combined atom types. For a combined type, the counts for the individual atom types have been summed across bins and a single energy curve generated from these combined types that represents all component atom types

Figure 3.9: Performance of KB_Top500 and its merged atom types PMFs. KB_0.1 is included for reference. Each merged atom types PMF is denoted by its merge threshold. **Top.** Evaluating Criterion 1, the ability of the potentials to minimize NNSMs. While the difference in minimized <<rmsd>> between potentials is small, a trend is observed. Combining atom types results in a net improvement in the ability to minimize structures. The best performing potential is KB_Top500_2.98 which achieves a <<rmsd>> 0.02 Å better than KB_0.1. **Bottom.** Evaluating Criterion 2, that the potentials should not significantly perturb the natives. The mean RMSD over all refined natives is given. As the number of combined atom types increases, the resulting potentials perturb the natives more, but all within the acceptable tolerance of < 0.80 Å.

Figure 3.10: Performance of KB_Top500_1.00vdw and its merged atom types PMFs. KB_0.1 is included for reference. Each merged atom type PMF is denoted by its merge threshold. **Top.** Evaluating Criterion 1, the ability of the potentials to minimize NNSMs. While the difference in minimized <<rmsd>> between potentials is small, a trend is observed. Combining atom types results in a net improvement in the ability to minimize structures. The best performing potential is KB_Top500_1.00vdw_2.87. **Bottom.** Evaluating Criterion 2, that the potentials should not significantly perturb the natives. The mean RMSD over all refined natives is given. As the number of combined atom types increases, the resulting potentials perturb the natives more, but all within the acceptable tolerance of < 0.80 Å.

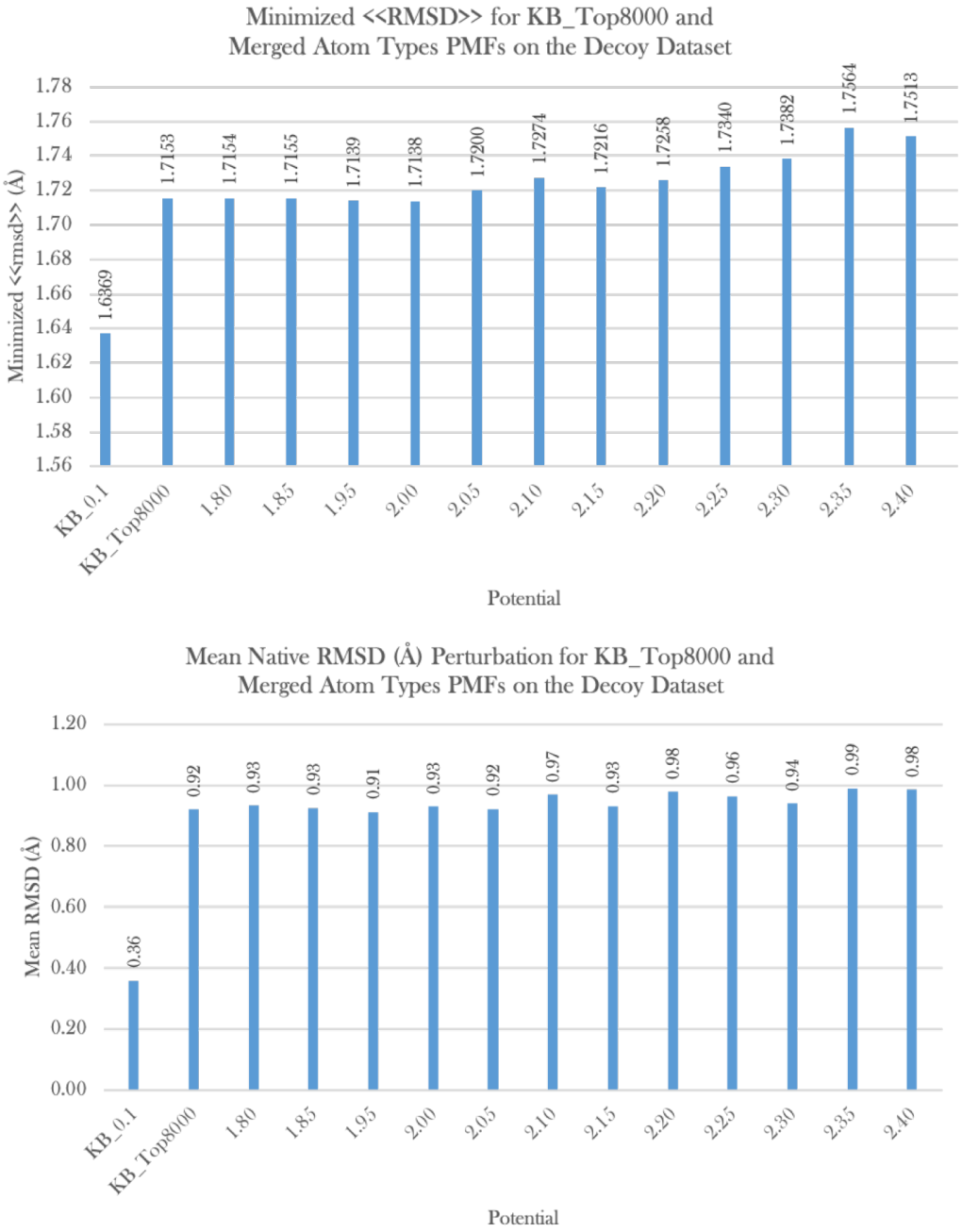

Figure 3.11: Performance of KB_Top8000 and its merged atom types PMFs. KB_0.1 is included for reference. Each merged atom type PMF is denoted by its merge threshold. **Top.** Evaluating Criterion 1, the ability of the potentials to minimize NNSMs. Merging atom types on PMFs derived from the Top8000 database does not result in improved performance in structure refinement. **Bottom.** Evaluating Criterion 2, that the potentials should not significantly perturb the natives. The mean RMSD over all refined natives is given. The KB_Top8000 potential and all potentials containing merged atom types PMFs derived from it significantly perturb the native.

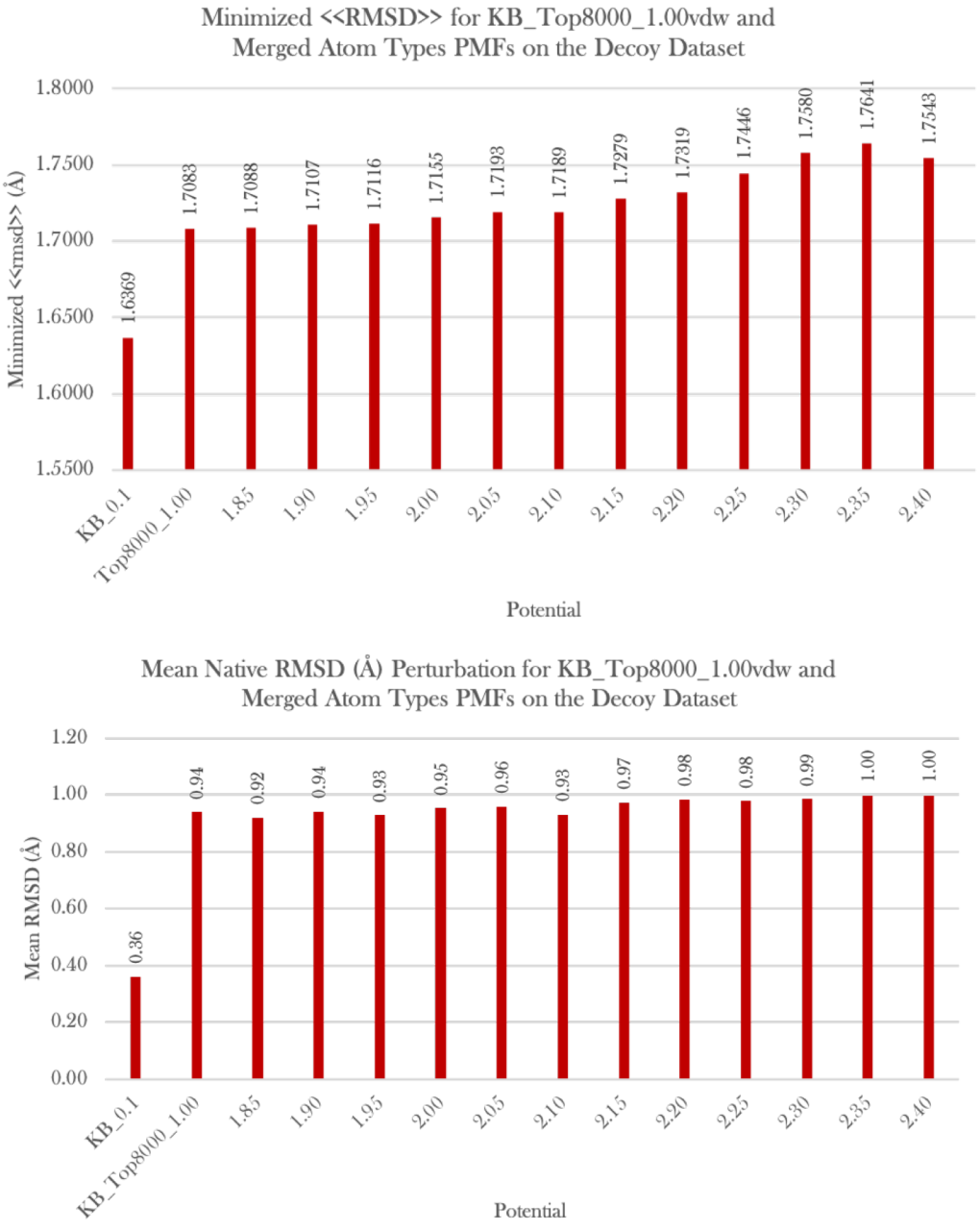

Figure 3.12: Performance of KB_Top8000_1.00vdw and its merged atom types PMFs. KB_0.1 is included for reference. Each merged atom type PMF is denoted by its merge threshold. **Top.** Evaluating Criterion 1, the ability of the potentials to minimize NNSMs. Merging atom types on PMFs derived from the Top8000_1.00vdw database does not result in improved performance in structure refinement. **Bottom.** Evaluating Criterion 2, that the potentials should not significantly perturb the natives. The mean RMSD over all refined natives is given. The KB_Top8000_1.00vdw potential and all potentials containing merged atom types PMFs derived from it significantly perturb the native.

Applied across the CASP dataset, the merged atom types potentials do not result in significant improvement in the refinement of the structures (Figure 3.13). The potential with the best <<rmsd>> for this dataset (KB_Top500_2.70, <<rmsd>> = 2.914 Å) only improves <<rmsd>> by 0.004 over KB_0.1 (<<rmsd>> = 2.918Å). The results for the KB_Top500_1.00vdw potentials were similar with the best improvement in <<rmsd>> over KB_0.1 being 0.002 Å (KB_Top500_1.00vdw_2.70, coincidentally of the same threshold). The merged atom types potentials derived from KB_Top8000 and KB_Top8000_1.00vdw resulted in net degradation of the structures.

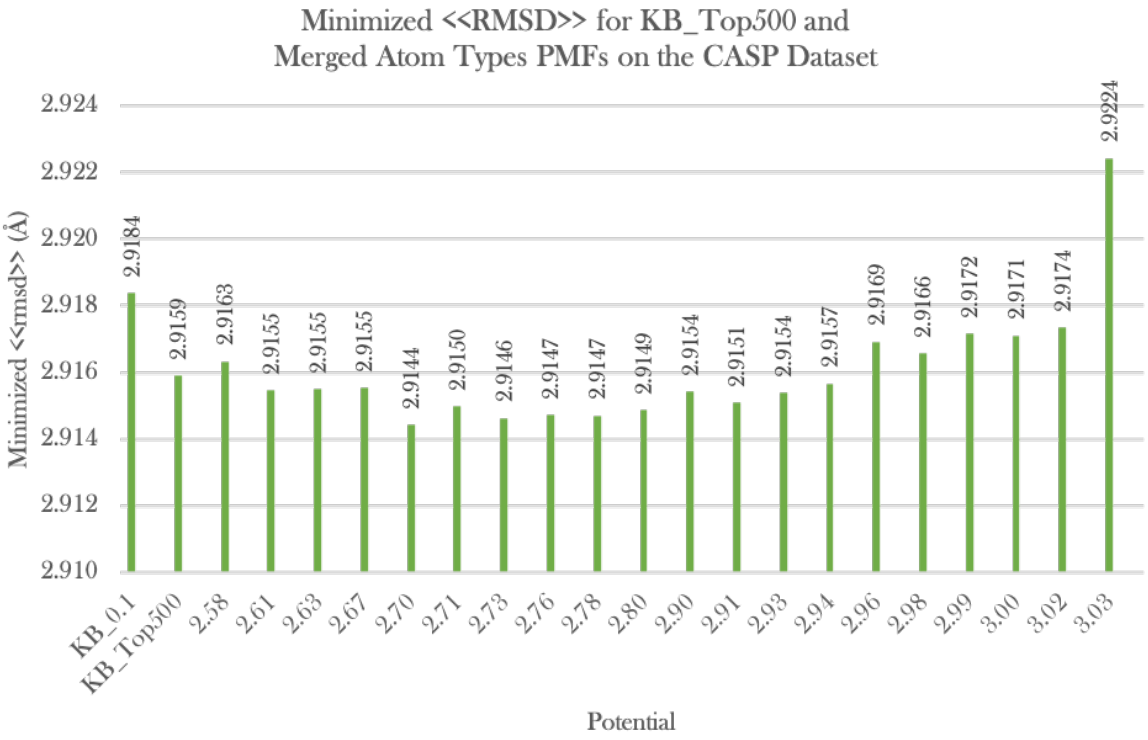

Figure 3.13: Performance of KB_Top500 and its merged atom types PMFs on the CASP dataset. On this dataset, minimizing structures with KB_500, KB_500_1.00vdw, and their derived merged atom types PMFs did not result in significant improvement of <<rmsd>> relative to KB_0.1. The best performing potential of the KB_Top500 set of PMFs improved <<rmsd>> by 0.004 Å relative to KB_0.1.

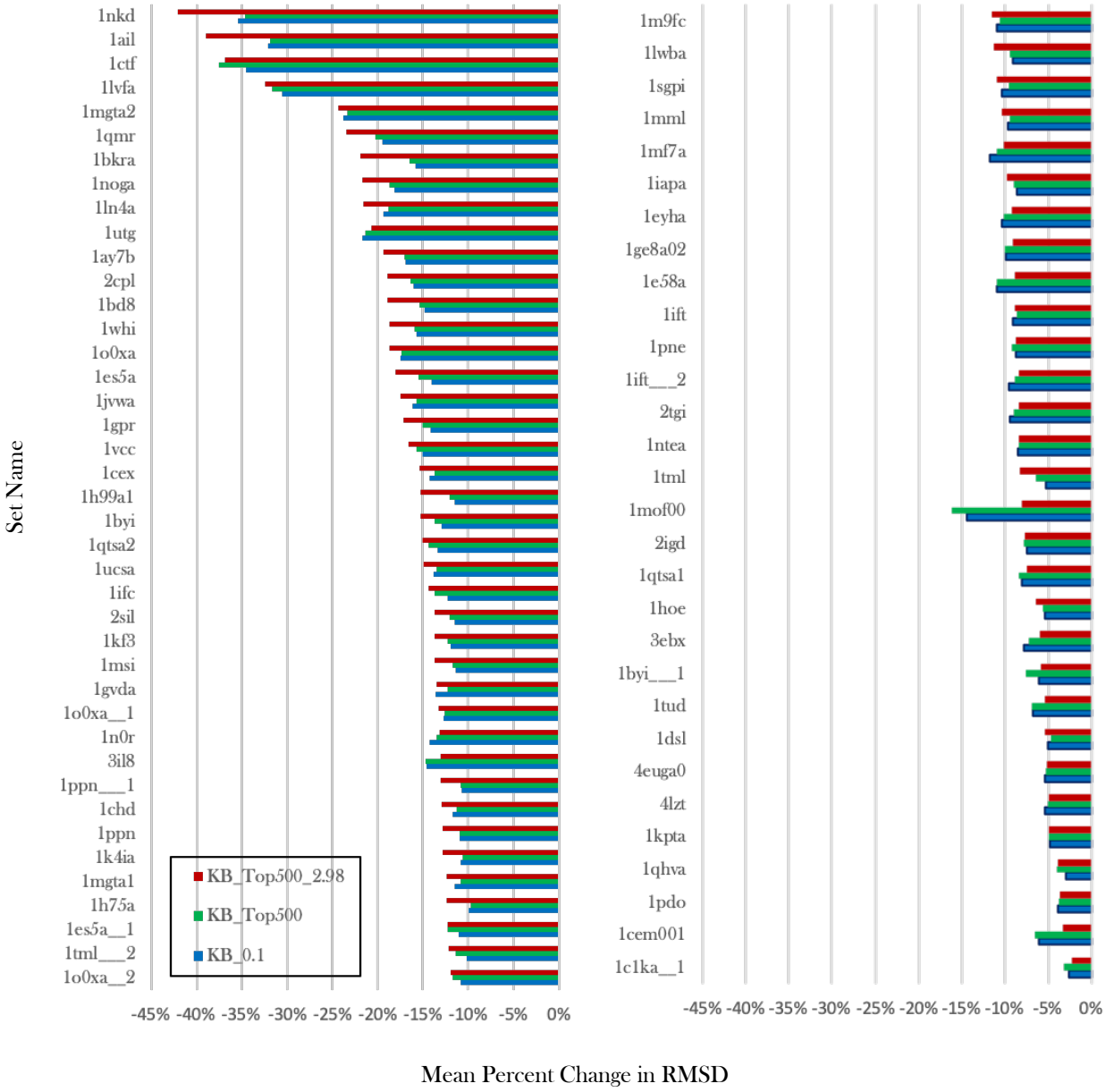

Figure 3.14: PEM using KB_0.1, KB_Top500, and KB_Top500_2.98 on the decoy dataset. The PI of each set with respect to starting and ending mean NNSM RMSD from the native is given. The performance of three potentials is presented: KB_Top500_2.98, KB_Top500, and KB_0.1. KB_Top500_2.98 is the best performing potential of all tested with respect to PEM over this decoy dataset. It was derived via the atom type merging process applied to KB_Top500.

3.3.2.4 Summary

PMFs generated from the four databases (Top500, Top500_1.00vdw, Top8000, and Top8000_1.00vdw) along with PMFs derived from these original four potentials via the atom type merging process were all used as the KB components of hybrid MM/KBs potential. Each hybrid potential was tested in PEM over the CASP and decoy datasets. Their performance was evaluated against two criteria: first, their ability to not perturb the coordinates of the native structures, and, second, their ability to improve the <<rmsd>> of the datasets as compared against KB_0.1.

The potentials generated from the Top8000 and Top8000_1.00vdw databases performed poorly with respect to both criteria (Table 3.8, Figures 3.12 and 3.13). They significantly perturbed the natives, and they did not improve <<rmsd>> with respect to KB_0.1. On the CASP dataset, they overall degraded the structures. The potentials generated from the Top500 and Top500_1.00vdw databases performed on average better than KB_0.1, with the atom type merging process further improving their ability to refine structures (Table 3.8, Figures 3.10 and 3.11). With respect to database selection, filtering out all structures with clashes did not result in improved performance for the potential derived from the Top500 database, but did for the Top8000 database (Table 3.8).

The potential that performed best in PEM over the decoy dataset was KB_Top500_2.98 with 124 atom types (combined types listed in Figure 3.9) in the KB component. It reduced the decoy dataset's <<rmsd>> from 1.872 Å to 1.617 Å.

3.4 Discussion

This work set out to address several questions pertaining to the generation of knowledge-based potentials of mean force. First, does generating PMFs from a larger structural database allow for

smoother pairwise energy curves to be produced which would in turn allow PEM to more easily move across the energy surface in search of the global minimum without getting trapped in local minima? Likewise, would removing structures from structural databases that have steric clashes, thus eliminating unnatural artifacts clashes produce in the pairwise energy curves of PMFs, improve the performance of the potentials in PEM? Lastly, are all 167 atom types necessary in the formulation of a PMF? Are some of them redundant? That is, are there atom types that are so characteristically similar within proteins that they can be merged into a single atom type, leveraging the combined statistics of two or more atom types to better represent them all? Furthermore, would doing so produce PMFs that when used in hybrid KB/MM potentials for PEM, allow the refinement process to better minimize structures?

3.4.1 Generating KB Potentials from a Larger Structure Database

Conclusively, the KB/MM potentials containing PMFs generated from the larger structure database (Top8000) led to worse PEM performance. On the CASP dataset, their use resulted in net degradation of the structures. This result is the most instructive result of this set of experiments. Generating PMFs from the larger database did result in smoother energy functions (Figure 3.15), but that did not allow for structures to better be minimized toward the global minimum. Instead, it allowed PEM to make large changes to structures, potentially moving them away from the native. If we let $<< rmsd >>_{delta}$ indicate the double mean RMSD (as defined in section 3.2.2.2) over all sets in a testing dataset with respect to the minimized vs starting state of the models, then $<< rmsd >>_{delta}$ indicates how much refinement alters the models in a dataset during minimization. PEM with KB_Top500 resulted in $<< rmsd >>_{delta}$ of 0.70 ± 0.18Å on the decoy dataset and 0.52 ± 0.15Å on the CASP datasets, whereas PEM using KB_Top8000

resulted in $<< rmsd >>_{delta}$ of 1.42 ± 0.30 Å and 1.28 ± 0.32 Å for the decoy and CASP datasets respectively. KB_Top8000 significantly altered the structures.

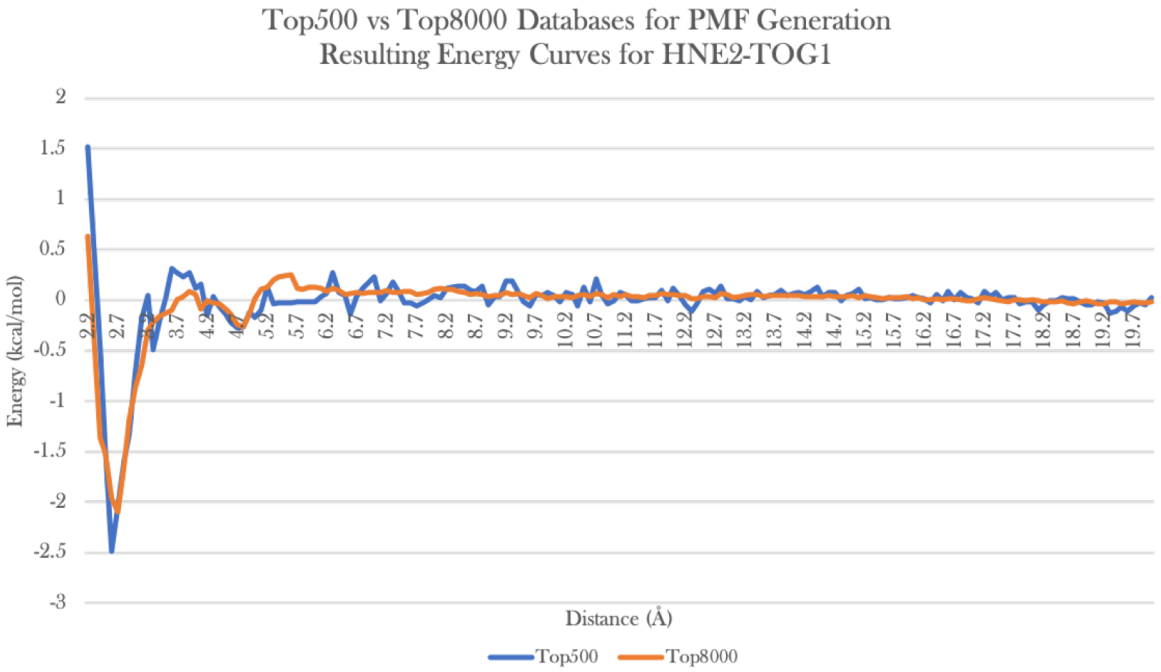

Figure 3.15: Comparison of HNE2-TOG1 energy curves generated from the Top500 and Top8000 databases. The energies are calculated using Lu and Skolnick's formalism. The curve generated from the Top8000 database is significantly smoother than the energy curve generated from the Top500 Database, indicating that the larger set of statistics will result in smoother energy surfaces.

To summarize these results, KB_Top500 makes small changes to structures, consistently improving them, whereas KB_Top8000 makes large changes to models and is much more volatile in its minimizations. Furthermore, consistently large perturbations of the natives (Figures 3.11 and 3.12) by KB_Top8000 indicate that it does not have a strong attractor basins around the natives. Interestingly, as Figure 3.7 shows, KB_Top8000 favors structures that are further from the native. On the decoy dataset, for structures in the 3.00 – 3.25 Å range for model starting RMSD, KB_Top8000 on average improves them by 17.50%, a significant improvement. For a structure with an RMSD from the native the middle of that range, KB_Top8000 on average moves it from 3.12 Å RMSD to 2.57 Å RMSD towards the native. While that is an impressive improvement,

starting model distance from the native is not foreknown, and with large possible reward comes large risk as is evidenced by KB_Top8000's general poor performance. This dynamic between risky refinement methods with a large possible improvements and conservative methods with consistent but small improvements has been observed and documented in recent CASP experiments [59].

It's possible that with too large a statistical dataset, the energy curves became too generalized and featureless, embodying large features such as hydrogen bonds, but losing many small but important features of atomic interactions, and it may be these small features that are crucial to a KB potential's performance. It may also be the roughness which prevents structures from moving too far, creating a conservative but consistent potential for refinement. Contrary to expectations, rather than a smoother energy curve allowing for larger improvements in refinement, it may be that the rough energy surface of KB_Top500 allows it to be successful in PEM, consistently making small improvements.

3.4.2 Eliminating Structures with Clashes from the Databases

While the Richardson lab filters structures for clashes when building their databases, they allow some clashes in the database as long as their proportion is sufficiently small. For statistical potentials, any clash will introduce artifacts into the energy surface. Therefore, when selecting structures for as statistical database, a strict policy of no clashes should be enforced. In building the KB potentials in this work, a simple policy was enforced: if any structure had a van der Waals overlap > 1.00 Å (based on standard atomic radii) for any non-bonded and non-hydrogen bonded atom pair, that entire structure was discarded. As a result, for the Top500_1.00vdw and Top8000_1.00vdw databases, 51 and 468 structures were discarded respectively from the original databases. KB_Top500_1.00vdw did not outperform KB_Top500 as a potential for PEM. In fact,

KB_Top500_1.00vdw and its set of merged atom types PMFs performed slightly worse than KB_Top500 and its set of merged atom types PMFs.

For the Top8000 PMFs, eliminating all clashes did result in improved performance. This may be due to two factors. First, 468 structures were removed from the Top8000 database, eliminating 2053 clashes, an order of magnitude more than the 220 clashes that were eliminated by removing 51 structures from the Top500 database. An order of magnitude more clashes may have had a larger negative effect on the Top8000 PMF than the fewer clashes incorporated into KB_Top500. Secondly, the KB_Top500_1.00vdw may have been negatively impacted by the statistical loss of the structures. As already discussed, the larger statistical dataset used to generate the Top8000 PMFs resulted in smoother energy curves, arguably too smooth as it allowed those PMFs to make large inconsistent (with respect to moving towards native) changes to models. Removing some of Top8000 statistical dataset may have been beneficial to the potential's performance in PEM. On the other hand, for KB_Top500_1.00vdw, the statistical loss may have outweighed the benefit of eliminating energetic artifacts due to clashes.

3.4.3 Combining Atom Types in PMFs

The question was posed as to whether or not classifying atoms into all 167 residue-specific heavy atom types is optimal for a KB potential. An iterative atom type merging algorithm based on the similarity of multiple atom types' energy curves was used to derive sets of PMFs containing merged atom types. Each base PMF (KB_Top500, KB_Top500_1.00vdw, KB_Top8000, and KB_Top8000_1.00vdw) was used as a starting PMF on which the merging algorithm was run. Reasonable combinations of atom types were merged: the hydroxyl groups of serine and threonine, carbons from hydrophobic residues, and, as a general pattern, atom types of the same element and position on the backbone or in the side chains of the residues.

In general, the merged atom types PMFs derived from KB_Top500 and KB_Top500_1.00vdw performed better than the PMF they were derived from. For the Top8000 PMF sets, the opposite was true. Combining atom types within those potentials did not yield improved performance in PEM. For the Top500 PMFs, combining similar atom types may have had two benefits. One, it may have allowed similar atom types to leverage their combined statistics into energy curves that better represent them, and, two, the smoother potentials of the combined atom types may have given minimization more latitude to move the models and explore the energy surface. As stated above, for minimization using KB_Top500, $<< rmsd >>_{delta}$ of the decoy set was 0.70 ± 0.18Å. For the best performing merged atom types PMF (KB_Top500_2.98), $<< rmsd >>_{delta}$ was 0.82 ± 0.20 Å. With the only difference between these two potentials being that KB_Top500_2.58 contains a set of merged atom types, the merging process smoothed some of the energy curves, allowing beneficial movement of the structures to achieve better minimization than the base KB_Top500 potential (Figure 3.9).

Taken together, the performance of the resulting potentials from the atom type merging process for both the Top500 and Top8000 PMFs indicate that if the statistical database is large and the energy surface of a potential is already smooth, then combining atom types will not result in net improvement for PEM, but if the energy surface of a PMF is rough, then combining atom types may improve performance.

3.4.4 Conclusions

Taken together, these experiments – using a larger statistical dataset, eliminating clashes, and merging atom types – suggest that there exists some size of a statistical dataset between that of the Top500 and Top8000 databases that will generate PMFs for hybrid KB/MM potentials that can achieve larger improvements in refinement while still maintaining consistency. The Top8000

database is too large a dataset because its use results in energy curves that are too featureless and smooth, allowing for too much freedom of movement and often degrading models in refinement. On the other hand, the negative effect of eliminating structures with clashes from the Top500 database, suggesting sensitivity to the loss of statistics, and the positive effect of merging atom types, resulting in some smoother energy curves and more movement of the structures in minimization, indicate that KB potentials could benefit from having a larger dataset than the Top 500 database. It is not clear whether combining atom types was key to improved performance or if the improvement was due to the better statistical representation and reduced roughness of the energy curves for the combined types. Lastly, even though the removal of structures with clashes had a negative impact on the performance of KB_Top500_1.00vdw, it is more likely that the negative effect was caused by the reduced statistical dataset rather than some missing positive effect of the energetic artifacts caused by the clashes.

Finally, the CASP dataset was chosen as a real-world dataset for testing. The generation of the decoy dataset ensures that the decoys are on an accessible path from the native. Smooth shifts are made to the structures and there should be no major energy barriers on the way to the native. On the other hand, the CASP dataset consists of models with no guarantee that there are no major problems such as issues with side chain packing that must be resolved to get to the native. PEM using potentials such as those explored in this chapter is not intended to make large changes in structures and is designed to evaluate and address issues with side chain packing. The CASP dataset therefore a much harder dataset and this is evident in the results. Whereas the best potential could improve the decoy dataset <<rmsd>> by 0.25 Å, it could only improve the CASP dataset <<rmsd>> by 0.03 Å. This indicates that conservative methods for structure refinement such as PEM are not enough for the current quality of predicted models. The method of PEM

should be included as part of a pipeline for structure refinement, as it done by the KoBaMin server (which uses KB_0.1) [92] and the Feig group [51].

3.4.5 Future Work

Future work includes identifying the optimal size for the statistical database. It is possible that the optimal size is somewhere between the Top500 and Top8000 databases. The goal is to balance a potential's ability to provide an energy surface that can be traversed, yet still consistently move models towards the native.

Another potentially lucrative avenue is in examining the composition of the database. Much progress has been made with using existing homology information and, recently, coevolution residue contact information [56] in the prediction of protein structures. PMFs generated in this work were general and meant to be generally applicable. Given the breadth of the CATH and SCOP2 databases, it should be possible to use sequence and homology information to select and/or seed structural databases for PMF generation with structures from the same fold as the structure that is being minimized. Since a PMF embodies the patterns discovered in native structures, why not use related and similar structures to generate PMFs specialized for individual families and/or folds of proteins? Specialized fold-specific potentials may be better able to refine structures of that fold than more general potentials such as the ones generated and analyzed in this work.

Chapter 4

4. A Novel Graph Theoretical Protein Structure Comparison and Analysis Technique

4.1 Motivation

Protein structure comparison remains a non-trivial task. Whether for analyzing the results of different protein structure predictors, different conformations of the same protein, or similar conformations of related proteins, the comparison and analysis of differing and complex three-dimensional structures is a difficult yet fundamental task.

In the bi-annual Critical Assessment of Protein Structure Prediction (CASP) experiment, methods to compare and analyze protein structures are of critical importance in the evaluation of the experiment [15]. For each CASP, sequences for proteins whose structures have been empirically solved but not yet published are released to protein structure predictors. Predictors generate structure models for these sequences which are then compared against the known structures and ultimately ranked to determine which predictors produce the most reliable structures. Not only are individual predictors judged, but also is the field as a whole in order to determine how well protein structure prediction is advancing from one CASP experiment to the next. The methods used to compare protein structures need to be intuitive yet powerful enough to able to evaluate and rank complex 3-dimensional structures. There are many competing priorities

for protein structure comparisons. Should models be ranked according to their global fit? Or their local accuracy? Are the side-chains packed correctly? Or is backbone geometry the important factor? To account for the often-orthogonal pull of differing comparison priorities, many methods of comparison have been designed.

The organization of this chapter is as follows. The next section discusses important considerations when analyzing and designing techniques for protein structure comparison. After that, a review of prominent techniques for protein structure comparison is given. Following the review, a novel technique is then proposed that allows for a deep analysis of structural similarities. This method identifies exactly all of the parts of two structures that are the same, presenting information about structure pairs that no other technique provides.

4.2 Important Considerations for Methods that Compare Protein Structures

Ideally, methods to analyze the similarities of and differences between protein structures should have certain properties [93]: They should be quantitative and visualizable (i.e. they should produce an overall metric but rely on underlying information that can easily be visualized in a meaningful way). They should not only allow analysis across large data sets, but also allow insightful analysis into individual comparisons. They should be stable against large variations in small parts of the structures (i.e. large swings in variable loops or at the termini of a structure should not result in large leaps in the similarity score). Finally, any new method should provide information that is not easily accessible from other measures, and their assessments should be intuitive to understand.

It is important to note that in protein structure comparison there is a distinction between the global and local accuracy of structures and that these two directions of structure analysis are often orthogonal. Globally accurate structures are those which orient the tertiary components of

structures, such as domains, correctly relative to each other while locally accurate structures are those that get the details of the components correct. Structures which are globally accurate might not be locally accurate and vice versa. For example, domain movements in multi-domain structures will contribute to a poor global score even if the domains themselves are locally accurate. Balancing the orthogonal pull of the analysis global versus local accuracy remains a key difficulty in protein structure analysis.

4.3 Existing Metrics

Given the complexity of protein structures and the reality that desired properties for a protein structures comparison metric can conflict, a many metrics have been developed. At its most basic level, when comparing protein structures, a set of correspondences between reference points (usually the α backbone carbon atoms, or Cαs) in one structure to reference points in the other is required, and it is based on these correspondences that differences and similarities in the two

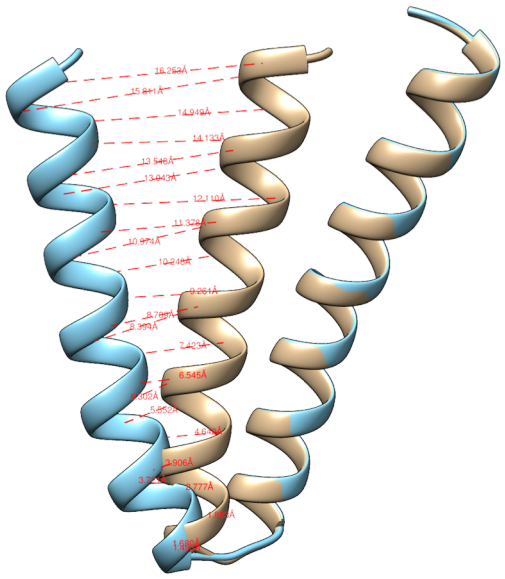

Figure 4.1: Correspondences for structural comparison. 1rop and an artificially modified version of it. In this example, the correspondences are the distances between the Cαs of analogous residues in the superposed structures. In cases where the comparison isn't between identical proteins, analogous residues are determined via sequence alignment.

structures can be assessed (Figure 4.1). Broadly speaking, there are two major categories of methods for protein structure comparison, superposition-based methods and contact-based methods, the difference between them being how the correspondences are determined.

4.3.1 Superposition-Based Metrics

In superposition-based methods, the correspondences between structures are the distances between analogous Cαs following a superposition of one structure onto another. The optimal superposition is determined by finding the transformation of one structure onto the other that minimizes the Root Mean Square Deviation (RMSD) of the corresponding Cαs between the structures. The RMSD can be returned as a score for the two structures but it suffers a couple of drawbacks. The major drawback is that RMSD is calculated by taking the square of the errors. The parts of the structures with the largest errors will dominate the score. Consequently, structures that are similar throughout but have a small part that is very different, such as a loop, will receive poor scores. Figure 4.2 shows human estrogen receptor α in two conformations which differ only in the terminal alpha helix's orientation yet have an RMSD of 6.24 Å. The other major drawback of

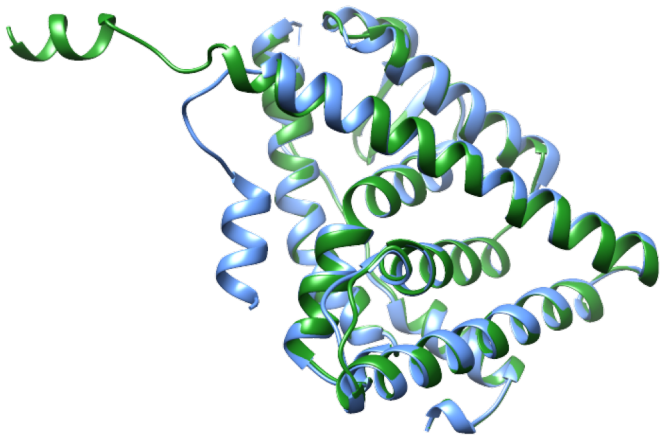

Figure 4.2: Human estrogen receptor α in two conformations. These conformations only differ in the orientation of the terminal alpha helix, yet they have an RMSD score of 6.24 Å. PDB accession codes: 1R5K, 1A52.

RMSD is that the analysis of the errors is difficult. If, for example, one wanted to compare a model against the native to identify which parts were well-modeled and which were not, the superposition errors cannot be used because they are ambiguous. Is any particular error due to an intrinsic difference at that location or an unfavorable superposition? Figure 4.1 is a contrived, but good example of this. The superposition errors of the left helix increase from the hinge to the terminus, but the helix as a whole as well modelled as the right helix. They are both identical to the reference structure. It is the superposition that gives the left helix its large errors.

4.3.1.1 Local Global Alignment: GDT & LCS

The Local Global Alignment (LGA) method was developed to overcome the shortcomings of RMSD [21]. LGA consists of two complementary components, the Global Distance Test (GDT) and the Longest Continuous Segments (LCS) algorithm. The idea behind LGA is that rather than relying on a single global superposition of the two structures, multiple superpositions can be used to identify regions of similarity that could not be identified in a single global superposition.

With the GDT component of LGA, the goal is to find the largest set of residues that can be superimposed under some distance threshold. More specifically, for a given distance threshold d, GDT finds the largest set of residues that can be superimposed where no corresponding pair of residues has a distance greater than the threshold. This in effect finds the largest region of global similarity between the structures under that threshold where "global" refers to sequence. The residues in the region can come from anywhere in the protein sequence. Within LGA, GDT uses thresholds from 0.5 Å to 10.0 Å in increments of 0.5 Å. For each threshold, GDT produces a score, the percent of residues that are in the region under that distance cutoff.

While the GDT component focuses on global regions, the LCS component is designed to identify regions of local similarity. LCS finds the longest continuous – within the sequence –

segments of the structures that can be superimposed under some RMSD cutoff. The cutoff is an important distinction. Whereas GDT finds the maximum number of residues that can be superimposed and whose *distances* all fall under a threshold, LCS finds a largest continuous segment of the sequence that can be superimposed who's total ***RMSD*** falls under a threshold. For a set of residues, the goal isn't to minimize the distances, but to minimize the RMSD of that set. This has a major consequence for LCS. Choosing RMSD as the selection criterion allows for optimal similarity information for a region. In comparison, because GDT's goal is to minimize the distances between all the residues in the region, it cannot guarantee optimal results, only an approximation. Within LGA, LCS is run with default thresholds of 1.0, 2.0, and 5.0 Å. Like GDT, for each threshold, LCS returns the percent of residues in the longest continuous segment under that RMSD cutoff. The LGA program also includes in its output the RMSD of each region.

In order to combine the global information from GDT with the local information from LCS, LGA calculates a total score for a pair of structures as a weighted sum of scores calculated for the GDT and LCS components. Using a weight factor w ($0.0 <= w <= 1.0$), the LGA score is defined as

$$LGA\_S = w * S(GDT) + (1 - w) * S(LCS) \quad (4.1)$$

where S(F) is itself a weighted sum of the percent of residues that can fit under each threshold for that component. Lower valued thresholds are weighted more heavily, and the total sum is divided by a factor based on the number of thresholds used. S(F) is thus defined as follows:

```
X = 0;
for threshold v_i in v_1, v_2, …, v_k {
     Y = (k – i + 1)/k;
     X = X + Y*F_v_i;
}
S(F) = X /((1 + k)* k/2);
```

While LGA_S combines the local information from LCS and the global information from GDT into a single score, it is the GDT component of this score that has made its way into prominent use. The GDT component is used as a key metric in the evaluation of the CASP experiments [15]. From it, a GDT_TS score is calculated as the average of the percent of residues under distance cutoffs (1.0, 2.0, 4.0, 8.0). A high accuracy version, GDT_HA can be calculated using cutoffs (0.5, 1.0, 2.0, 4.0).

GDT works well when comparing structures which have only a single domain but cannot handle structures with multiple domains. If two dual-domain structures are the same except that the domains are shifted relative to one another, GDT will count the residues in the larger domain as matching and omit the residues in the smaller domain because GDT maximizes the number of residues that can be optimally superimposed. If the domains are close in size, GDT will give a poor score for the comparison even though the structures may overall be very similar. The problem is that GDT is not designed to identify multiple regions of similarity in a structure, only the largest one. This limits GDT to working on either single domain structures, or those structures whose domains are known and whose domains can be analyzed one by one. While the issue is phrased in the language of domains, the core problem applies even to single domain structures. If secondary structures within a domain are shifted relative to each other, the same results will occur. Only the largest region will be identified. Whether analyzing multi-domain structures or single domain structures, smaller regions of similarity are omitted from the score and analysis by GDT.

4.3.1.2 TM-Score

The Template Modelling Score (TM-Score), developed by Zhang and Skolnick, was developed as a tool to assess the quality of protein structure threading templates and the predicted models from those templates [94]. Motivated by the metrics that came before it, it was also developed to address two issues common to existing metrics. The first is that metrics such as GDT, which are based on the percent of residues that fit under a sets of distance cutoffs, discard detailed error information by treating all residues within a cutoff band – for example, [4, 8) Å – as identical contributions to the score. The second issue that is that the magnitude of many metrics is dependent on the size of the input structures. The same score for a pair of small proteins and a pair of large ones can have different meanings. For example, as Zhang and Skolnick point out, an absolute GDT score of 0.4 can reflect significant similarity between structures of size 400 residues but could indicate a near random selection from the PDB for small structures of size 40 residues.

Motivated by the Levitt-Gerstein score [95], the TM-score is defined as

$$TMScore = Max\left[\frac{1}{L_n}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \tag{4.2}$$

where L_N is the length of the native, L_T is the length of the aligned residues to the template, d_i is the distance between the i^{th} pair of aligned residues, and d_0 is a normalization factor to eliminate the dependence of the score on the size of the structures. d_0 is defined as

$$d_0 = 1.24\sqrt[3]{L_n - 15} - 1.8 \tag{4.3}$$

It approximates an estimation of the average distance of corresponding residue pairs in random related proteins in the TM-score superposition. Max indicates that the maximum TM-score is selected and returned after an iterative search finds the optimal superposition of the template to the native structure that maximizes the TM-score.

The TM-score search process goes as follows. Starting from an initial fragment of L_{int} neighboring residues aligned onto the native, this fragment is superimposed to the corresponding residues of the native. Then, all the residues in the template with a distance to the native of less than d_0 are included in the fragment and the fragment is superimposed onto the native again. This repeats until the rotation matrix converges. This process is performed for an extensive set of initial fragments, determined from a set of initial fragment sizes – $L_T, L_T/2, L_T/4, L, 4$ – each of which, if less than L_T, are windowed across the template from N- to the C-terminus to give the fragments. The result of the whole procedure is a near-optimal superposition of the template onto the native which maximizes the TM-score. The TM-score cannot be guaranteed to be the maximum, but an experiment performed by the authors showed that tripling the search with additional randomly selected initial fragments improved the TM-score of a small percent of their test set (6%) by only a negligible amount (<0.002).

The inclusion of the normalization factor d_0 successfully eliminates dependence of the score on the size of the structures. Regardless of size, random unrelated protein pairs should have a TM-score of ≈ 0.17. It can therefore be said that TM-scores ≤ 0.17 indicate unrelated proteins, and, by the definition of the score, a value of 1.0 indicates completely identical structures. The TM-score provides a well-defined metric which can be interpreted uniformly no matter the size of the structures. Like GDT, it still suffers the same drawbacks associated with calculating a score

from the superposition of one structure onto another. It is not built to handle multi-domain structures nor to identify separate regions of a model which match a reference structure.

4.3.1.3 Sphere Grinder

Whereas GDT and TM-score evaluate the global accuracy of pairs of protein structures, Sphere Grinder was designed to allow for insight into their local accuracy [96]. Instead of optimizing some global superposition of one structure onto another, it superimposes local environments from throughout the structures. Given a reference, a model, and a radius R_0 (default 6 Å), for each residue in a reference structure, Sphere Grinder identifies all atoms whose distance from that residue (by default determined using Cα) is less than R_0 as that residue's local environment. Then the corresponding atoms in the model are identified and superposed onto the local environment in the reference. The RMSD of this superposition is the accuracy score of that residue. The Sphere Grinder Score is then calculated as the percent of residues with an RMSD less than a cutoff.

By varying R_0, Sphere Grinder can be tuned to focus on local or global accuracy. Small values for R_0 are used to evaluate structures for local accuracy and large values for global accuracy.

4.3.2 Contact-Based Metrics

Contact-based metrics are those which, rather than being based on the distances between the structures after a superimposition, are based on corresponding distances and/or interactions within the structures. A contact can be defined as two atoms that are separated by less than some threshold distance. The major benefit of using intra-structure distances as opposed to superposition errors as the underlying information for a metric is that the information is

unambiguous. Two residues are either the same distance apart in two different structures or they are not, whereas the superposition error for corresponding residues depends on the superposition.

4.3.2.1 CAD

The Contact Area Difference (CAD) score performs structure comparisons by examining the contact areas of pairs of residues within both structures [97]. It was designed to overcome a major issue of RMSD, its inability to rank partially correct models of some reference. RMSD tends to be dominated by the incorrect parts of the structures, and when ranking structures, those parts outweigh the parts that are well modelled. By basing the comparison on the contact areas of residues within the structures, CAD was also inherently designed to account for side chain packing.

To calculate a CAD score, first, contact area matrices are calculated for both a reference and a model structure. For each pair of residues *i* and *j* in a structure, their contact area is calculated by rolling a probe of radius R over residue *i* to determine the area of the surface traced by the center of that probe that is occluded by the van der Waals surface of residue *j*. Doing this for all pairs of residues in both structures results in contact area matrices A^R and A^M for the reference and model structures respectively. The CAD score is then defined as the normalized weighted sum of the absolute differences of the contact area matrices.

$$CAD = \frac{100 \sum_{i,j} W_i W_j \left| \left(A_{ij}^R - A_{ij}^M \right) \right|}{A_{worst}} \tag{4.4}$$

where:

$$A_{worst} = C \sum_{i,j} W_i W_j \frac{1}{2} \left(A_{ij}^R + A_{ij}^M \right) \tag{4.5}$$

and:

$$W_i = e^{\left(-\frac{B_i}{B_{std}}\right)} \tag{4.6}$$

The CAD score is weighted by factors W_i, calculated from the temperature factors of the residues and a standard factor which the author suggests take the value of 20 to balance the weight of residues that have high vs. low temperature factors. CAD is normalized by the weighted average of the elements of both matrices. The author recommends C=0.9 in the normalization factor so that scrambled random structures will have close to 100% CAD difference from their reference structure.

The result of all these considerations is a score that is robust against fractional changes and domain movements and that still accounts for side chain packing within the structures while accounting for the natural propensity of some parts of a structure to be flexible by factoring in residue temperature factors.

4.3.2.2 lDDT

The local Distance Difference Test (lDDT) is designed to address the issue of domain movements between comparable structures [98]. It does so by creating a measure that balances both local and global similarity, referring respectively to environments within a structure and the structure as a whole. It also includes, built into the score itself, the validation of stereochemical plausibility.

lDDT measures the number of contacts within a predefined inclusion radius R_0 that are preserved between the reference and model structures. To calculate the score, the distances for all atom pairs with a distance under R_0 are saved in a set of distances L. In the model, the percent of corresponding atom pairs whose distances are preserved, within a tolerance threshold, those in L

are computed. lDDT is calculated as the average of four fractions of matching atom pairs using tolerance thresholds 0.5 Å, 1.0 Å, 2.0 Å, and 4.0 Å, the same thresholds as used to calculate the GDT-HA score. lDDT can be calculated over all atom pairs, just the Cα atoms, or the backbone atoms. By default, lDDT uses an inclusion radius of $R_0 = 15$Å. The authors determined the inclusion radius empirically by performing an analysis of the CASP9 experiment They examined the correlation between the GDC-all and lDDT scores of the CASP9 models as the value of the inclusion radius varied in the range 2 to 40 Å. GDC-all is an all atom version of GDT with thresholds from 0.5 to 10.0 in steps of 0.5 Å. The authors found that $R_0 = 15$Å produces scores that are a good balance between local and global similarity. Lower values for the inclusion radius focus the metric more on local similarity while higher values shift the balance towards global similarity.

lDDT validates stereochemical plausibility by considering stereochemical violations and steric clashes. Stereochemical violations are bond lengths and angles which diverge from expected values by more than 12 standard deviations. Steric clashes are atom distances distance which are less than the sum of their van der Waals radii, within a default tolerance of 1.5 Å. If side-chain atoms of a residue show stereochemical violations or steric clashes, all distances including any of the side-chain atoms of that residue are considered not preserved. If the backbone atoms exhibit stereochemical violations or steric clashes, any distances that include any of that residue's atoms are considered not preserved.

lDDT can also be calculated using a set of structures as the reference state. Using multiple references, for each atom pair, an acceptable distance range is defined by the min and max observed distance over the set of reference structures. To calculate the score, each atom pair distance in the model which falls within its acceptable range is considered preserved. The

percentage of preserved distances, accounting for stereochemical violations and clashes, is returned as the score.

4.4 Regions of Similarity

All existing protein structure comparison methods return a score for similarity, but few give a deep underlying look at the parts of the structures which match. Zemla's Global Distance Test (GDT) [21] partially does this by identifying the largest region whose superposition errors all fall under some threshold, but the region and its errors are dependent on that superposition, and smaller regions are not identified. By converting the Cα distances matrices of two structures into a graph, a maximum clique analysis can be used to identify the largest non-overlapping regions of similarity between the structures. These regions can easily be visualized, and they lend themselves to a deep analysis of the underlying similarities between structures, complementing existing methods of comparison by providing additional information that is not readily available. Additionally, when applied to an analysis such as that performed for each CASP experiment, models which correctly represent each domain in a multi-domain structure but whose orientations differ from the native will be immediately apparent. A regions of similarity analysis can be performed on multi-domain targets without *a priori* knowledge of the domains.

4.4.1 Methods

4.4.1.1 Definition of Regions of Similarity

A Region of Similarity is a set of aligned residues between two protein structures whose intra-structure Cα distances are all the same – within a tolerance threshold – in both structures and which all form a cohesive unit within the structures. Rigorously defined, given a reference and a model structure whose residues have been aligned, a region of similarity is a set of residues whose:

1. Size is at least 10 residues.
2. Pairwise Cα atomic distances are all the same, within a tolerance threshold, in both structures.
3. Contact map in the model forms a connected graph.

The third condition ensures that the residues in a region all come from some local part of the model. It forces a region to contain contiguous residues in three-dimensional space and enforces the idea that a region should represent a set of residues that take the shape they do because they are strongly interacting with one another. Without this condition, it would be possible to have residues from distant parts of the structures forming a region because they are coincidentally the same distance apart in both structures.

4.4.1.2 Finding Regions of Similarity

To find the largest region of similarity between two protein structures, first their sequences are aligned. Then the distance differences matrix is calculated: $D_{i,j} = R_{i,j} - M_{i,j}$ where i and j are aligned residues, R is the Cα distance matrix for the reference structure, M is the Cα distance matrix for the model structure, and D is the distance differences matrix. A similarity graph is then built from $D_{i,j}$. Every residue is a vertex, and there is an edge between two vertices if their value in $D_{i,j}$ is less than a tolerance threshold, $t = 1.0$Å by default. The maximum clique of this graph reveals the set of potential residues for the region of similarity. The last step is to select only those which form the largest spatially contiguous region in the model. To find this region, a graph is built from the contact map of the model (all residues are vertices and there is an edge between two residues if their Cαs are less than 10.0 Å apart), and the largest component found by a depth-first

search of this graph reveals the final residues in this region. If this region contains at least 10 residues, return it, otherwise there is no region of similarity between the structures.

A disjoint set of regions of similarity (denoted simply as RoS) can be found by iteratively identifying regions on the same similarity graph G. After each region is found, its residues are removed from G to prevent residues from being assigned into multiple regions. This continues until no more regions are found. If the two structures are identical, there will be a single region containing all residues. If the structures consist of two identical domains that are shifted relative to each other, then there will be two regions of similarity, one for each domain.

Regions of similarity can also be used to perform a threshold tiered test inspired by GDT: RoS-GDT. Given a set of thresholds {1.0, 2.0, 4.0, and 8.0 Å}, four regions of similarity are identified: $R_{1.0}$, $R_{2.0}$, $R_{4.0}$, and $R_{8.0}$. Each region is the largest region of similarity in the similarity graph built under its threshold which, for each threshold except the first, completely encompasses the region of similarity found for the previous threshold. To find these regions, four similarity graphs, $G_{1.0}$, $G_{2.0}$, $G_{4.0}$, and $G_{8.0}$, are constructed as described above. To start, the largest region of similarity in $G_{1.0}$ is found. This is $R_{1.0}$. Then, the subgraph in $G_{2.0}$ consisting of the residues from $R_{1.0}$ is identified and all residues which are neighbors of this subgraph and which have an edge to every residue in this subgraph are selected. The maximum clique found within these residues in $G_{2.0}$ is the maximum set of residues which can be combined with those in $R_{1.0}$ and still form a clique in $G_{2.0}$. Within this combined set of residues, the largest connected component in the contact map graph is found, and the residues in this component are returned as $R_{2.0}$. The same process is repeated for $R_{4.0}$ and $R_{8.0}$. The thresholds {0.5, 1.0, 2.0, and 4.0 Å} can be used to perform an RoS-GDT-HA test. The set of regions found by RoS-GDT is called an expanded region of similarity.

The regions found by RoS-GDT show tiers of modelling quality, but they only encompass one part of a pair of structures. Like the original GDT, in a multi-domain structure where separate domains are well modelled but shifted relative to each other, RoS-GDT will identify only the largest domain. To identify multiple areas of a pair of structures that are similar, a disjoint set of Expanded Regions of Similarity (ERoS) can be identified. Each expanded region of similarity has tiers of residues found using the thresholds {1.0, 2.0, 4.0, and 8.0 Å}. To start, a set of disjoint regions of similarity is identified under the first threshold. Then, for each subsequent threshold, each region of similarity, in the order of initial discovery, is expanded to the next threshold using the similarity graph for that threshold omitting all residues found in all other regions so far. At the end of the process, a set of Expanded Regions of Similarity is returned. A score similar to GDT_TS can be calculated from this set: the average of the percent of residues under each threshold. ERoS_score is defined as:

$$ERoS\_score = \frac{1}{4}\left(R_{t_1} + R_{t_2} + R_{t_3} + R_{t_4}\right) \tag{4.7}$$

R_{t_n} is the sum of the fractions of residues that fall under the n^{th} threshold over all the expanded regions of similarity. Each fraction is calculated with respect to the number of residues in the reference structure.

Expanded Regions of Similarity can also be generated using twenty thresholds: {0.5, 1.0, 1.5, ..., 10.0}. The fraction of residues under each threshold can be used to generate plots which show the percent of the structures which match under decreasing levels of accuracy. This technique is denoted as ERoS-Plot.

4.4.1.3 Visualizing Regions of Similarity

Local accuracy maps can be generated from regions of similarity. They show, at the sequence level, which residues in a model are within which region of similarity. Up to five regions can be colored: blue, green, purple, brown, and yellow. If a single threshold is used, such as when finding disjoin regions of similarity, the region with the largest number of residues is colored blue and the region with the smallest number of residues is colored yellow. If expanded regions of similarity are being visualized, the colors are determined in the same order by the size of the regions identified using the most stringent threshold. Residues which are not in any of the top five regions are colored red, and those that are not in the reference or the model are colored white. The colors have been chosen to be visually distinct. If expanded regions of similarity are being visualized, within each color, the shades vary uniformly in saturation and luminosity to indicate under which threshold that residue was added to the region. Darker shades indicate more stringent thresholds. Finally, if RoS-GDT regions are being represented, a divergent color scheme from blue to peach is used. Red residues are not in any of the regions. Examples of local accuracy plots are given in Figure 4.3.

ERoS plots can be generated from the ERoS-Plot data. For each model, the total fraction of residues identified under each threshold is plotted and the result shows how well that model represents the target. Those models which include larger portions of their structure within regions of similarity under tighter thresholds are the better models. Figure 4.6 gives an example of ERoS-Plot.

Regions of similarity can also be visualized on the three-dimensional structural representations of proteins as well. Both PyMol [99] and Chimera [100] scripts can be generated to select and color residues belonging to each region and threshold so that individual structure pairs can be examined in detail. Figure 4.3 shows two structures, 1qvi_A and 1b7t_A superposed with

their regions of similarity colored. Their RMSD is 60.36, but they are actually quite similar. The major difference is a large shifted domain at the bottom. It is also easy to see through the regions that there are two domains in the "body" portion that are shifted slightly relative to each other.

4.4.1.4 Feasibility Study

Identifying regions of similarity relies on solving instances of the NP-complete problem of finding maximum cliques. To ensure the feasibility of the technique, a study was performed on a set of 88,758 pairs of different experimentally determined structures for identical proteins provided by Kufareva[93]. This dataset contains a variety of structures of varying sizes and levels of similarity. The smallest structures contain less than 20 residues and the largest over 1000. Measured by

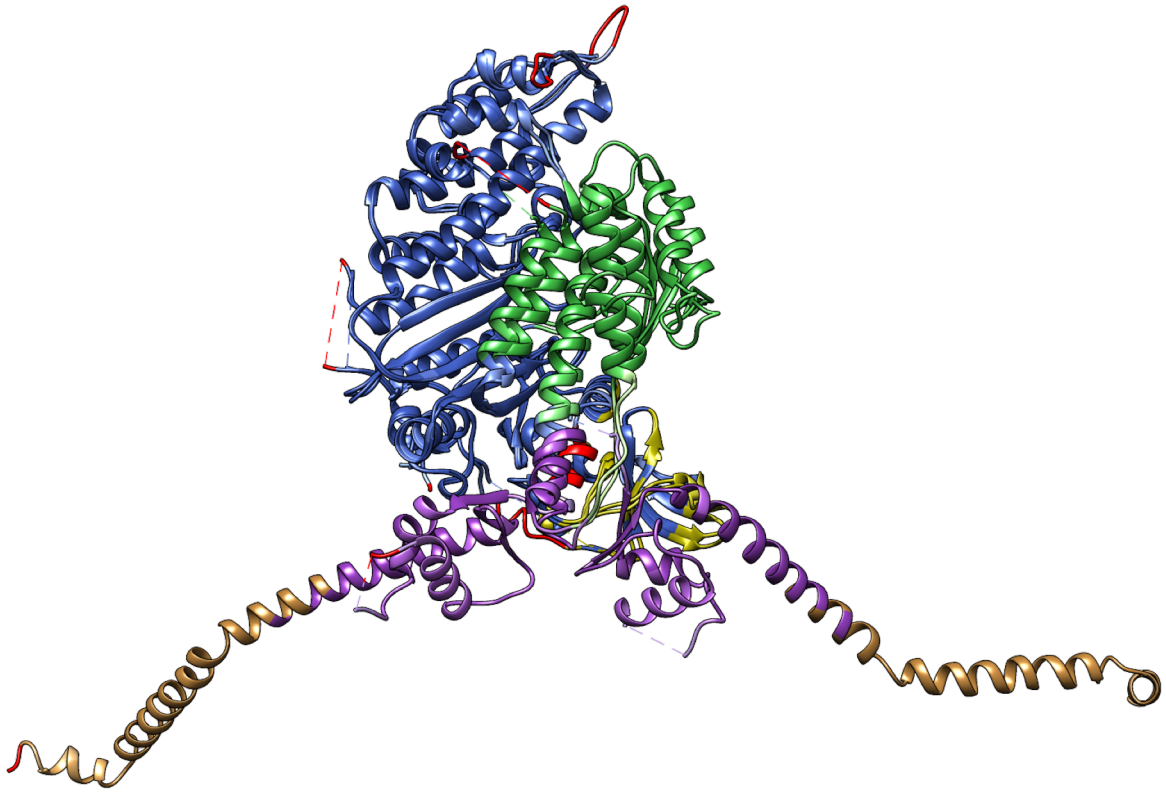

Figure 4.3: Regions of Similarity Colored on Structures 1qvi_A and 1b7t_A. These are two empirically determined structures of the same protein from the Kufareva dataset. They have an RMSD of 60.36 Å yet, as the regions indicate, they are actually quite similar with a significantly shifted domain.

LGA_S, the least similar pairs have scores less than 10 and the most similar have scores of 100. For each pair, RoS, RoS-GDT, RoS-GDT-HA, ERoS, and ERoS-Plot were generated. The runtimes were recorded and are presented below.

4.4.1.5 Software & Hardware

All algorithms for finding regions of similarity have been implemented in jProt, a java protein comparisons library freely available at https://github.com/amaus/jProt. Maximum cliques are found using Li, Fang, and Xu's C program implementation of their IncMaxCLQ algorithm[101]. Local accuracy maps and ERoS plots were generated using gnuplot. The feasibility study was performed on the lee2 cluster at the University of New Orleans. This cluster consists of 36 compute nodes, each with dual XEON X5650 CPUs. Lee2 has a total of 1.1 TB of RAM.

4.4.2 Results

4.4.2.1 Illustrating Regions through Local Accuracy Maps

Local accuracy maps can be generated using each of three major techniques: RoS, ERoS, and RoS-GDT. Figure 4.4 illustrates the differences between them using the two-domain target T0976 from the CASP13 experiment[15]. This target was chosen because most models roughly represent each domain (and some do accurately), but they generally shift the domains relative to each other with respect to the reference structure. In these plots, the top four models ranked according to their *ERoS_Score* are displayed.

The regions identified by RoS and ERoS show that in these structures, there are two large regions, blue and green, that are well-modelled. Since the residues in these regions are not sequential, it is likely that these are elements of secondary structure that are accurately representing parts of the tertiary structure of the reference. Additionally, in the top model, in each half there are

sequential segments of the sequence, brown and yellow, that are likely secondary structures shifted relative to the others. Comparing these plots against the three-dimensional structures illustrated in Figure 4.5, the two large regions correspond to the two domains and the yellow and brown regions are alpha helices shifted relative to their domains.

The information in these maps is information that regions of similarity can present in addition to the information provided by other methods of comparison. For example, while lDDT gives each residue a local accuracy score, regions of similarity can identify the sets of residues that together are all locally accurate as a group. While regions of similarity, like lDDT, is a measure of local accuracy, GDT is a measure of global accuracy. It tends to rank structures favorably that are globally accurate since structures with accurate global orientations are more likely to capture larger parts of the structures in an optimal superposition. In the case of T0976, GDT will rank well the models which have the domains in the same orientation as the reference structure. In conjunction with GDT, regions of similarity can then identify which parts of the structures that are globally accurate are locally accurate as well.

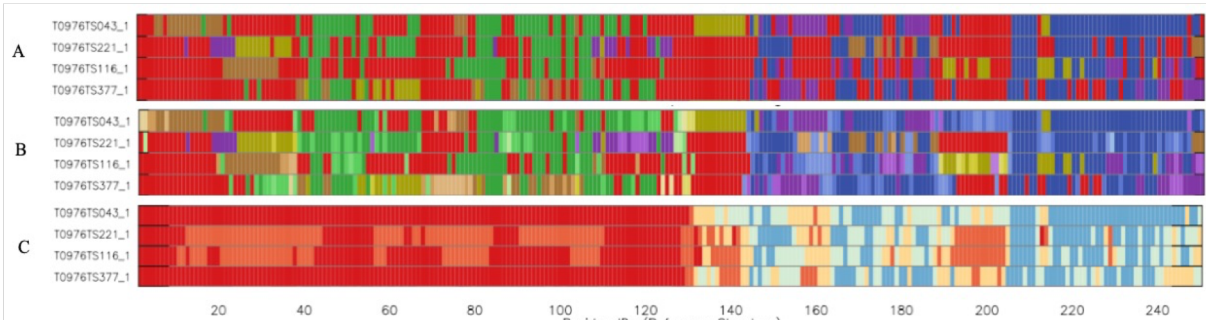

Figure 4.4: Comparison of the three Regions of Similarity methods on target T0976 from CASP13 (A) RoS: A disjoint set of regions of similarity (identified under the default threshold of 1.0 Å), colored in order of largest to smallest: blue, green, purple, brown, then yellow. Red indicates that a residue is not in any of the largest five regions highlighted. **(B)** ERoS: The Expanded Regions of Similarity. Starting from those found by RoS, each region has been expanded in turn to include residues at looser thresholds. The coloring is the same except that different shades indicate under which threshold the residue was added to the region. Darker shades indicate more stringent thresholds. **(C)** RoS-GDT: A test analogous to GDT. The largest region of similarity is identified and expanded through the GDT thresholds. The divergent color scheme indicates decreasing modeling accuracy from blue to light red for this region. Bold red indicates that a residue is not included under any of the thresholds.

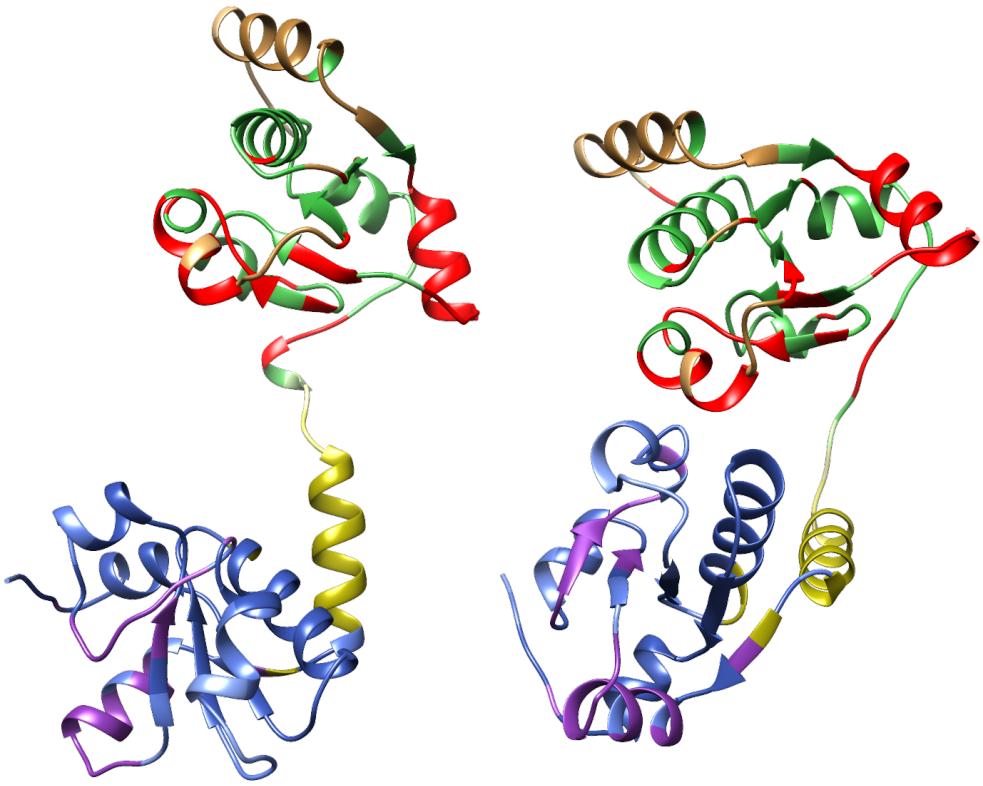

Figure 4.5: Regions of similarity identified for T0976 and T0976TS043_1. Left: T0976 (the reference) and on the right is T0976TS043_1 (the model) colored according to the expanded regions of similarity illustrated in Figure 4.3 **Right:** Despite the fact that the two domains in this structure are oriented differently between the reference and the model, the regions of similarity can still be identified and the overall similarity between the structures is apparent.

4.4.2.2 ERoS Plots

ERoS plots can be generated for one or more models of some reference structure. They show how well each structure models the reference by plotting the percent of residues within all regions of similarity under each of twenty thresholds {0.5, 1.0, 1.5, ..., 10.0 Å}. The larger the fraction of a structure that is included within regions of similarity under each of the thresholds, the better that structure will perform in the plot. Given that the underlying analysis relies on regions of similarity, ERoS Plots illustrate how well each of a set of structures match their reference structure locally across the whole of their structures.

Figure 4.6 shows the ERoS plot for the "first models" submitted for the CASP13 target T0976. In a CASP experiment, each group may submit multiple models for each target. The models plotted in Figure 4 are those each group submitted as their "first model", the model they wish to be included in the default rankings for the experiment. The curves of the models T0976TS043_1, T0976TS472_1, and T0976TS322_1 are highlighted in blue, green, and purple respectively. The first is the top ranked model by ERoS_Score. It should also be noted that this model is ranked first by lDDT as well. This is not surprising given the similarity between these two scores, but the scores are not directly analogous. The next two models are those ranked as the first and second place models respectively according to GDT_TS. The plot shows that while TS472_1 has a better global score, TS322_1 has more of its structure within regions of similarity across the majority of the thresholds. In other words, its local geometries are a better representation of the native.

In any structural comparison, structures with a high degree of global similarity, such as domains being in proper orientations, may not have a high degree of local similarity and vice versa. ERoS plots can be used in conjunction with global measures such as GDT or TM-Score to identify

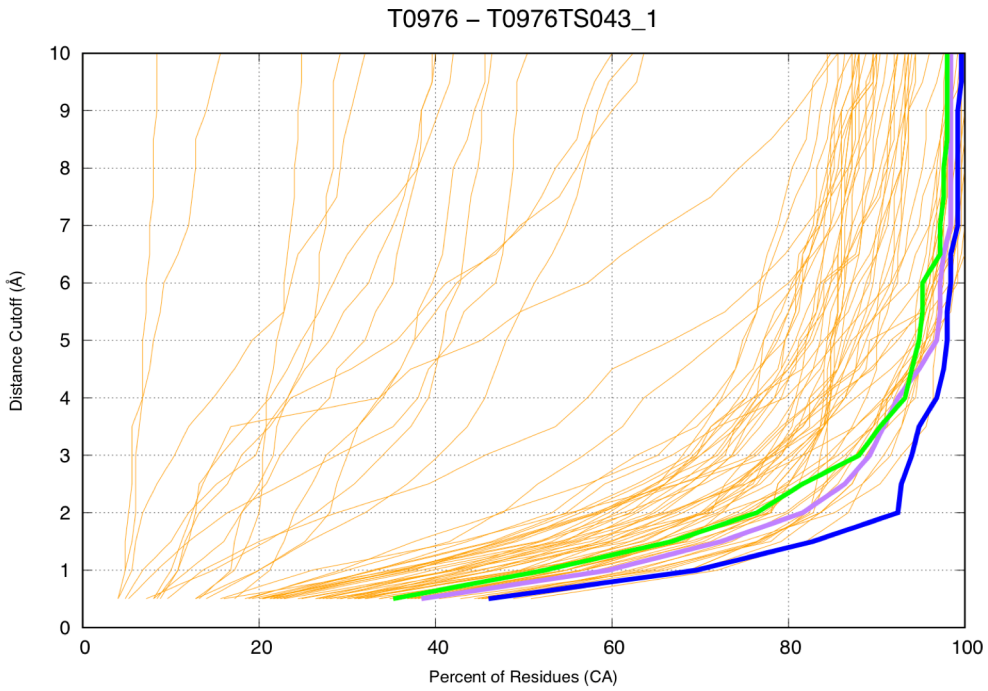

Figure 4.6: ERoS Plot for CASP13 target T0976. T0976TS043_1 (blue), T0976TS472_1 (green), and T0976TS322_1 (purple) are highlighted. The first is the model ranked best by ERoS_Score. The next two are the top two models ranked by GDT_TS. While TS472_1 is a slightly better global representation of the target (GDT_TS score of 59.2 vs 58.2 for TS322_1), the plot shows that TS322_1 is a better local representation.

those structures which not only match globally but locally as well. Combined with local accuracy maps and three-dimensional representations, the structures which exhibit both global and local similarity can then be further analyzed to identify exactly which parts of the structures match.

4.4.2.3 Feasibility Analysis

Since the regions of similarity techniques rely on solutions to instances of an NP-Complete problem (finding the maximum clique of a graph), these techniques were rigorously tested on a set of 88,758 pairs of different structures for identical proteins[93]. Table 4.1 summarizes the results.

Table 4.1: Region of Similarity Techniques Runtimes (ms)

Technique	RoS	RoS-GDT	RoS-GDT-HA	ERoS	ERoS-Plot
Average	1352	964	935	1749	7315
Median	991	620	539	1226	4350
Max	90457	89791	17813	98558	237509

In Table 4.1, the runtime statistics for five different comparison techniques are presented. As the table shows, the most intensive technique is ERoS-Plot. This matches expectations as ERoS-Plot has the largest number of thresholds to evaluate and therefore depends on solving more instances of the maximum clique problem than any other method. Its average runtime is 7.3 seconds. The maximum time recorded for any individual comparison is 238 seconds. This time is for the structure pair 2drd_C and 2j8s_A. Three of the largest runtimes in Table 1, those for RoS, RoS-GDT, and ERoS, are all for the same pair of structures, 3hhm_A and 2rd0_A. These results speak to the nature of instances of NP-Complete problems. For many cases, the solution will be easy, but for some, the solution will be difficult. For the majority of the comparisons, the solutions took on the order of seconds. For a few, the time required was on the order of minutes.

The identical proteins dataset is a rigorous test of these techniques. As an example of a practical application, the most intensive technique, ERoS-Plot, was run on the CASP12 dataset containing 131 targets with a total of 9545 models. The average runtime was 1.5 seconds with a median runtime of 553 ms and a maximum runtime of 23 seconds.

Figures 4.7 and 4.8 shows the ERoS-Plot runtimes for the identical proteins and the CASP12 datasets respectively. In the plots, the structure pairs are ordered by groups of identical proteins and by models for a given CASP12 target in the top and bottom of the figure respectively. In both plots, the outlier runtimes group together. These runtimes come from comparisons within

sets of multiple structures of the same protein in the identical proteins dataset and from within sets of models submitted for some target within the CASP12 dataset. In Figure 4.7, two outlying structures pairs are identified. The runtime for structure pair 2drd_C – 2j8s_A was 238 seconds and the runtime for pair 3hhm_A-2rd0_A was 132 seconds. Both of these pairs are shown in Figure 4.9. Likewise, the CASP model with the longest runtime, T0920TS421_1 with a runtime of 23 seconds, is compared against its reference structure in Figure 4.10.While a full discussion is beyond the scope of this research, it should be noted that there is some feature within the similarity graphs constructed for these structures that make them difficult instances of the max clique problem. No simple correlation was found between the size or the density of the graph and the runtime, but it can be noted that the longest runtimes tend to belong to large structures that are very similar.

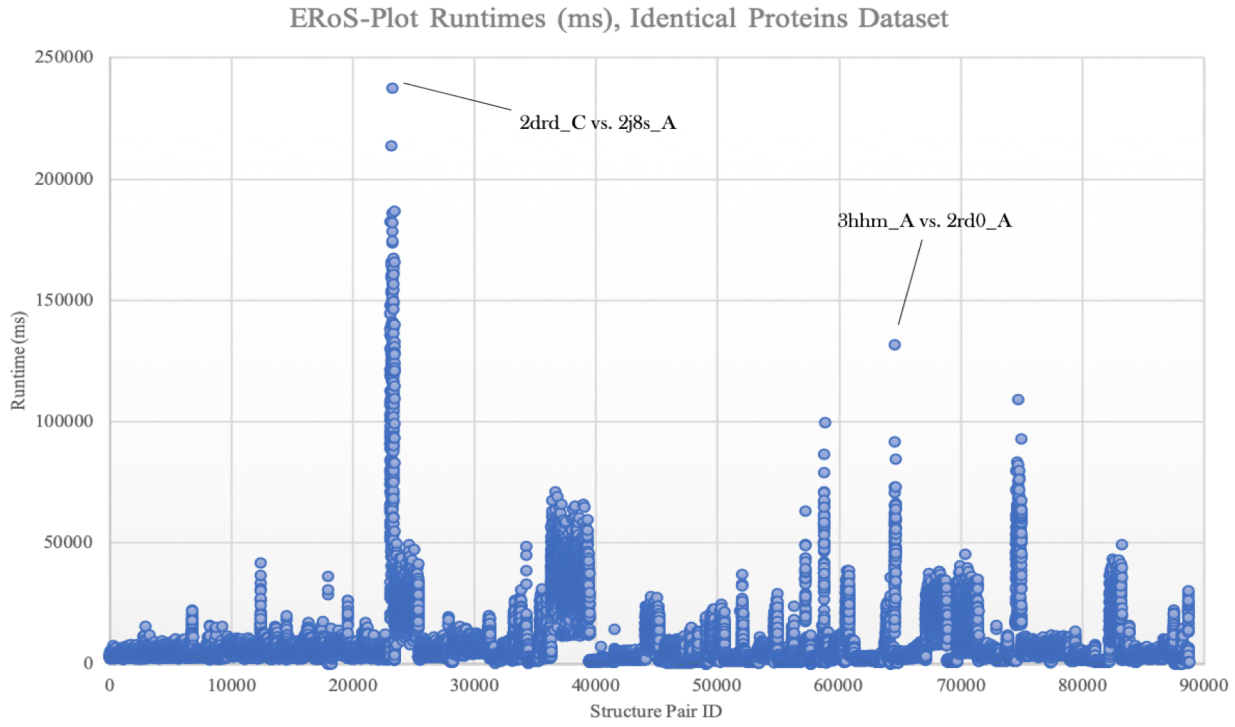

Figure 4.7: ERoS-Plot runtimes for the structure pairs in the identical proteins dataset. Two outlying structure pairs are labeled. The "spikes" are sets of identical structures all pairwise compared with each other. Identical sets tend to have similar runtimes. There is some undetermined property of their underlying similarity graphs that make them difficult instances of the max clique problem.

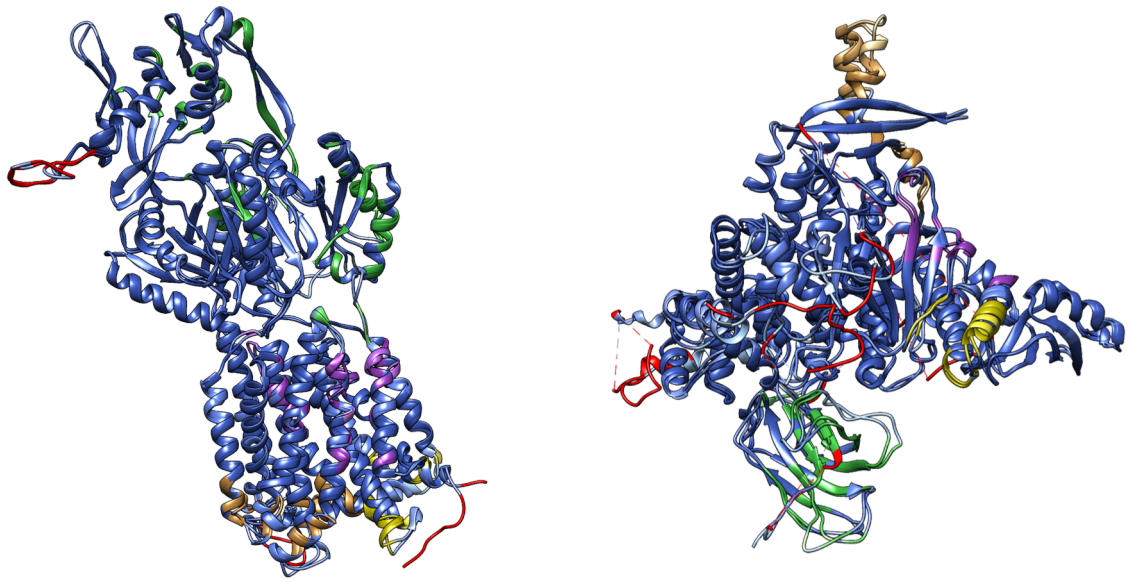

Figure 4.9: The two structure pairs from the identical proteins dataset with the outlier ERoS Plot Runtimes. Left: 2drd_C vs. 2j8s_A, runtime 238 s **Right:** 3hhm_A vs. 2rd0, runtime 132 s. Each structure pair is superposed with the ERoS regions colored. They are both large structures that are very similar, probably contributing to their long ERoS-Plot runtimes.

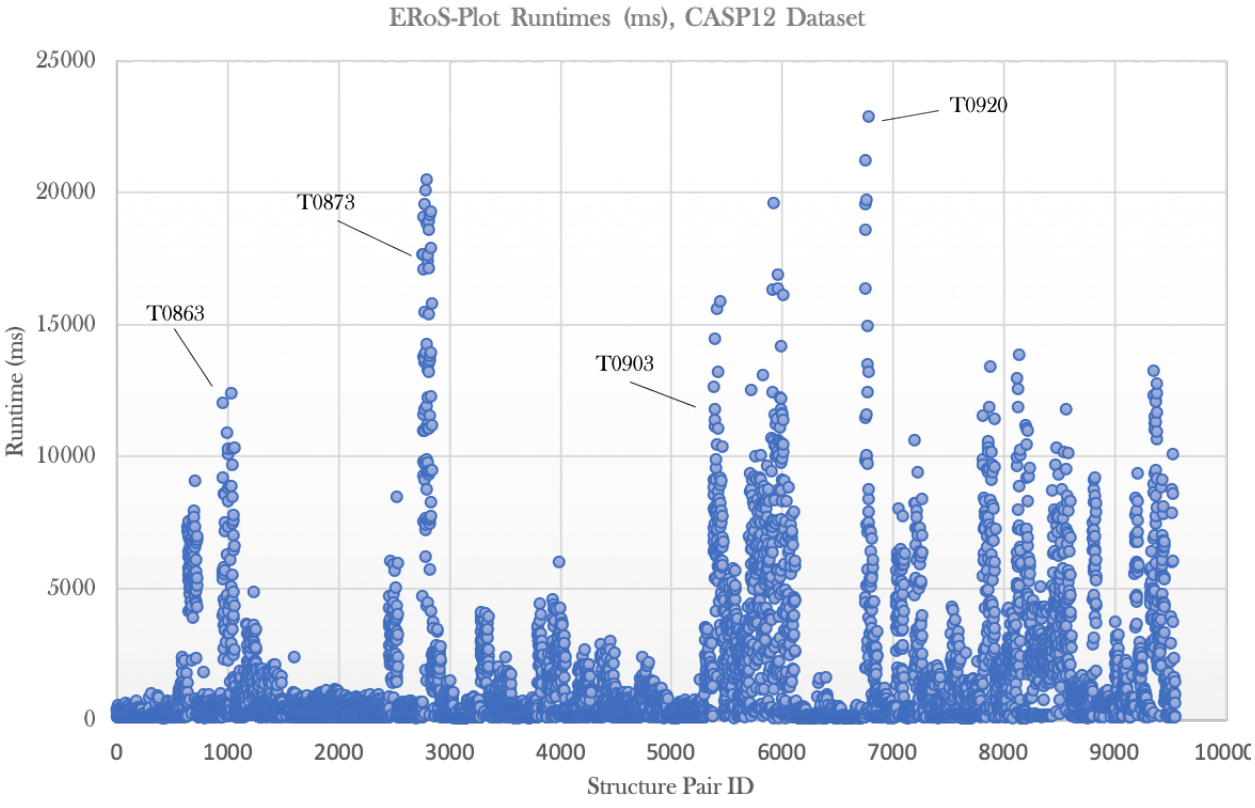

Figure 4.8: ERoS-Plot runtimes for the structure pairs in the CASP12 dataset. The most prominent "spikes" are labeled by the CASP target the structure pairs in it belong to. Note the scale for the runtimes. The range is 0-25 seconds, compared against Figure 4.7 with a runtime range of 0-250 seconds. Evaluation of the CASP12 dataset is feasible with this technique.

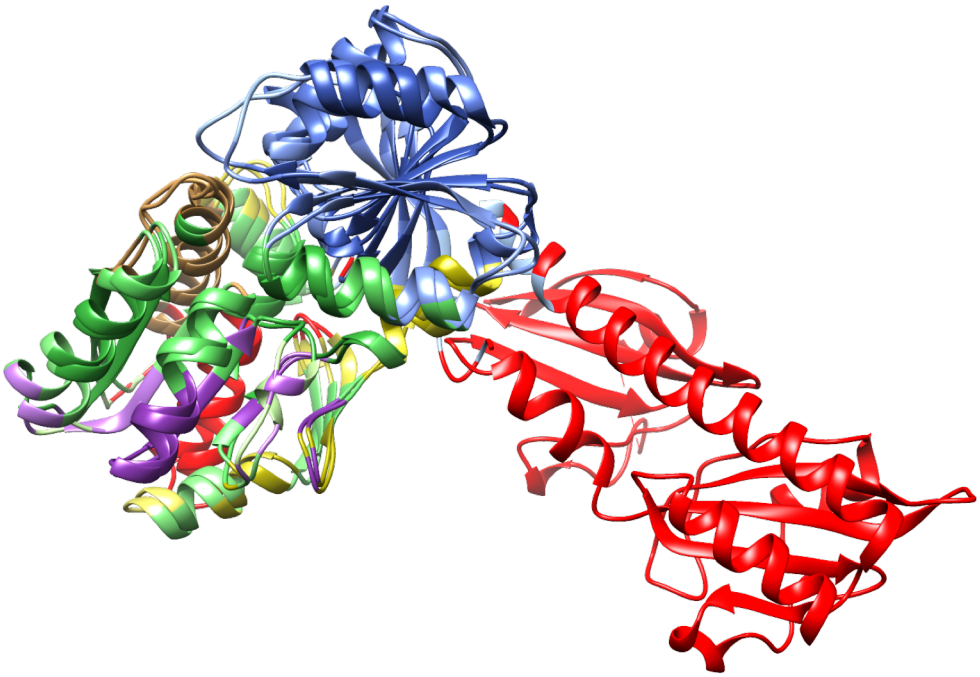

Figure 4.10: CASP12 model T0920TS421_1 compared against its reference T0920. This structure pair had the longest runtime for the ERoS plot technique, 23 seconds. T0920 is a two-domain target. For this model, one of the domains was submitted.

4.5 Discussion

Many protein structure comparison methods provide an overall similarity score for structure pairs, but few take an in-depth look at the underlying information of the comparison. GDT [21] partially does by allowing the largest set of residues from a model whose superposition errors on some reference are all under some threshold, but the set identified depends on the superposition and multiple sets are not identified. lDDT [98] allows for an in-depth look at the residues of the structures. It gives each residue a score, measuring how well its local environment (defined as all atoms within some radius of the that residue) is reproduced in a model by finding the fraction of preserved contacts within that environment. Likewise, Sphere Grinder [96] provides similar information. It also measures the accuracy of the environment around each residue, but instead of using contacts, it superimposes corresponding environments and uses the RMSD of that

superposition as the measure. Both methods provide scores for individual residues, but they do not identify sets of residues whose environment as a whole is reproduced.

Regions of Similarity is a contact-based protein structure comparison suite which performs a graphical analysis on the contacts within the structure to provide a detailed analysis of the similarities between two protein structures. A region of similarity is a set of residues that together are geometrically similar in both structures. That is, all of their inter-residue distances are the same, within some tolerance threshold. Based on a maximum clique analysis on the graph representing pairwise residue contact similarities between a pair of structures, regions are found independently of the superposition of the structures. Disjoint regions of similarity, those which are independent of each other and possibly shifted relative to each other, can be found. As a result, regions of similarity can be identified in multi-domain structures irrespective of domain movements. It must also be noted that while this method relies on solutions to the NP-Complete problem of finding maximum cliques, it has been tested against a rigorous dataset of similar proteins and found to be feasible.

Regions of similarity can easily and meaningfully be visualized. At the sequence level, residues can be colored according to their region and the tolerance threshold at which they were added to that region, showing not only which parts of the sequence form regions of similarity, but also giving an indication of the relative local accuracy of each residue. These local accuracy maps can be generated for sets of structures, allowing a group of models to be compared against some reference structure. These same regions can also be visualized on the individual three-dimensional structures using either PyMol [99] or Chimera [100]. Lastly, overall accuracy plots (ERoS-Plots) can be produced. These plots show, for each structure in some set compared against a reference, how the fraction of residues identified within regions of similarity changes as the tolerance

threshold of similarity is increased from 0.5 Å to 10.0 Å in increments of 0.5 Å. These plots allow for a whole set of structures to be quickly evaluated and for different models within a set to be compared against each other. Those models which are locally accurate over larger portions of the structures will be evident.

Regions of Similarity evaluates the local accuracy of a pair of protein structures. While different use cases may have different requirements, binding site analysis may require high levels of local similarity and conformational analysis may focus more on global similarity, in general, when evaluating models against some reference structure, the best models are those which exhibit both global and local accuracy, two orthogonal modes of comparison. Only by combining both global and local methods can the similarities of and differences between protein structures be fully explored. In conjunction with global measures such as GDT_TS and TM-Score [94], regions of similarity can be used to identify which of the models that are globally accurate are also locally accurate and furthermore, exactly which parts of the models are accurate representations of their corresponding parts in the reference. By providing access to information that was not previously available, regions of similarity allow for a novel and intuitive look into the similarities between protein structures and can be used in concert with existing metrics to provide a complete global and local comparative analysis of proteins structures.

4.6 Future Work

The CAD score works by creating a pairwise calculating a pairwise residue contact area matrix for both structures in a comparison. The difference between analogous pairwise residue contact areas is then used to calculate the CAD score. A regions of similarity analysis could be applied to this data, and if so, it would be possible to determine regions within two proteins that have the same side chain packing. This would add another dimension to local structural analysis. At present, RoS

is limited to analyzing backbone geometry. In the future, RoS will be expanded to calculate and analyze residue contact areas so that it can analyze both backbone geometry and side chain packing.

Chapter 5

5. Conclusion

The understanding of proteins is critical not only for advances in basic biology but also in the discovery of new treatments and cures for diseases. With the major advances made in genomics in the past few decades, it is now possible to determine the amino acid sequence of any protein [102], and as Anfinsen stated, the structure and function of a protein is completely determined by its amino acid sequence [12]. The field of protein structure prediction is concerned with developing computational techniques to determine the structure and function of a protein from its amino acid sequence. Despite much progress in the past several decades [15], [31], protein structure predictors are still not able to consistently produce models of high enough accuracy for desired applications such as rational drug design [59]. Protein structure refinement techniques are therefore being developed to move predicted models closer to the native state [62].

In this dissertation two major projects have been presented. The first is an in-depth examination and analysis of the formulation and generation of hybrid KB/MM potentials for protein structure refinement using potential energy minimization, and the second is a novel graph theoretical technique for protein structure comparison and analysis.

5.1 Hybrid KB/MM Examination and Analysis Summary

In the analysis of the hybrid KB/MM potentials, the generation of the potentials of mean force for the KB portion of the hybrid potential was the focus. Special attention was paid to the pairwise energy curves and the performance of the resulting potentials. In this analysis, several factors affecting the generation of the KB potentials were explored:

1. The effect of the counting scheme on the potentials, especially at critical low distances.
2. The size of the structural database used (either Top500 or Top8000) in the generation of the potentials, affecting the smoothness of the energy curves.
3. The strictness of the starting database, eliminating all structures with clashes to remove energetic artifacts from the energy curves
4. The number of atom types used in the generation of the potentials, identifying and combining similar atom types to improve the statistical representation of those atom types in the potential.

To evaluate performance, all generated potentials were applied in structural refinement against two datasets, a decoy dataset generated using quasi-elastic normal mode perturbation and a CASP dataset collated from the regular target submissions for CASPs 8-13. Every potential was evaluated against two criteria.

1. Refinement should not significantly perturb the native.
2. Refinement should move models closer to the native.

5.1.1 Results and Discussion

It was found that a very modest improvement in potential performance was achieved by altering the contact counting scheme in the statistics gather phase to initialize all PMF bins to zero rather

than one, and it was also found that combining similar atom types within the potentials generated from the Top500 databases resulted in a more significant improvement in performance. On the other hand, combining atom types for potentials generated from the Top8000 databases did not improve performance. Increasing the size of the starting database (generating potentials from the Top8000 database) resulted in potentials that were more volatile and performed worse in refinement. These potentials significantly altered natives and led to a net degradation of the models in the CASP dataset. Finally, removing all structures with clashes from the databases gave mixed results. For the smaller Top500 database, potentials generated from the subset only containing structures with no clashes performed slightly worse than the potentials generated from the full database. For the larger Top8000 database, removing clashes slightly improved the performance of those potentials.

When considering the implications of these results, it is important to note that the energy curves within KB_0.1 [20] (the original potential this work is based on) and within the PMFs generated in this work from the Top500 database (the difference between these and KB_0.1 being only the counting scheme) are rough. See Figure 3.15 for an example. This could be an indication that these potentials are capturing important features of the interactions that are key to refinement performance, or that a larger statistical database is needed to smooth out some of these artifacts. It is most likely the case that both implications are true. In either or both cases, it seems to be the roughness of these curves which prevents refinement from making large changes to structures.

In the case of the potentials generated from the Top8000 database, the curves are much smoother (Figure 3.15), but those potentials significantly perturb the natives and result in worse performance overall. It was expected that removing all clashes (and the energetic artifacts caused by them) would overall improve performance. So why did it not do so for the Top500 potentials?

It may be because removing the 51 structures from the database in order to eliminate all clashes negatively impacted the statistical robustness of the dataset. This would imply that the Top500 database is either just the right size or could be expanded to include more structures. Potentials generated from 500 structures containing no clashes should be tested.

Why did using combined atom types within the Top500 potentials improve performance? Combining similar atom types allows for an improved statistical representation of the combined types. The process resulted in potentials with more freedom to move structures that performed better in refinement. This implies that perhaps the Top500 database should be expanded to improve statistics, and also that there may be an ideal size somewhere between the 500 structures in Top500 and the 7957 structures in the Top8000 database for the generation of potentials of mean force.

The best performing potential generated in this work is one based on the Top500 database (including structures with clashes), with initialized statistical counts starting at zero, and containing 124 atom types with common combinations including backbone atoms of the same element and carbons from hydrophobic residues (Figure 3.8).

Moving forward, databases containing no clashes with sizes between 500 and 8000 structures should be tested, and atom type combinations on these potentials should continue to be determined and tested. Given that combining atom types did not result in improved performance for potentials generated from the Top8000 database, there may be a point at which combining atom types does not improve performance. This may coincide with an optimal statistical database size. Another avenue for improvement may be in using evolutionary data in the generation of potentials. With large databases of known families of proteins (SCOP2[28] and CATH [29]), it may be possible to generate specialized potentials for individual protein folds. If a homologous

family of a structure can be identified via structural or sequence analysis, a potential could be generated from or seeded with homologous structures, and this potential may better embody the patterns within the fold and allow more improved refinement of that structure.

5.2 A Novel Graph Theoretical Protein Structure Comparison Technique

In the process of generating and evaluating the performance of dozens of potentials for structural refinement, it was natural to ask how resulting structures of the potentials differed from one another. For example, does one potential better form hydrogen bond networks, and how would that look in the resulting structures? In general, if different predictors were better or worse at predicting certain structural motifs, could that pattern be noticed and how would one identify such regions of local similarity between structures? These questions led to the development of the Regions of Similarity family of techniques presented in Chapter 4.

These techniques allow for the exact identification of all regions between two structures that are similar, irrespective of changes in global similarity such as changes in relative orientation like domain shifts or conformational changes in disordered regions. It works by performing a graph analysis on the underlying similarities between two structures, the intra-structure Cα distances. If two analogous Cαs are the same distance apart in both structures that is a single point of similarity. By building a graph from these similarities and finding maximum cliques on it, complete regions of similarity, where all Cαs in that region are the same distance apart in both structures, can be identified. Despite relying on solutions to an NP-Complete problem, through rigorous testing, this technique has been found to be feasible.

Regions of similarity allows for a complete and intuitive analysis of the local similarity between two structures and can be combined with global measure of similarity such as GDT [21]

to identify structures that are both globally and locally similar (two orthogonal modes of comparison). Regions of similarity can be visualized in several ways to allow for a robust analysis of pairs or sets of structures. They can be visualized on the sequence level, allowing for a set of models of a native to quickly be analyzed (Figure 4.4). They can also be visualized on the 3D representations of structures (Figures 4.3, 4.5, 4.8, and 4.10), allowing for an in depth look into the similarities between any given pair of structures. Finally, plots relating increasing thresholds of similarity to the percent of residues included in all regions can be provided to give a good indication of overall structural local similarity (Figure 4.6). A tool to identify and visualize regions of similarity is freely available on GitHub[1], and this work is expected to have broad applications in rational drug design, the evolutionary study of protein structures, and in the analysis of the protein structure prediction effort.

An exciting avenue for future work on this project is in leveraging this technique to analyze the similarity data generated in the calculation of the CAD score [97]. CAD operates by generating pairwise residue contact area matrices for two structures. The difference between analogous residue pair contact areas is used to calculate its score. A regions of similarity analysis could be applied to this data to identify regions between proteins that have the same side chain packing. This would add another dimension to the Regions of Similarity project. At present, it can identify backbone similarity. With the addition of residue contact area analysis, side chain packing could be identified as well, allowing for a more complete look and a deeper analysis of structural similarity.

[1] https://github.com/amaus/jProt

Bibliography

[1] J. C. Whisstock and A. M. Lesk, "Prediction of protein function from protein sequence and structure," *Quarterly Reviews of Biophysics*, vol. 36, no. 3, pp. 307–340, Aug. 2003.

[2] W. L. Bragg, "The structure of some crystals as indicated by their diffraction of X-rays," *Proceedings of the Royal Society of London*, vol. 89, no. 610, pp. 248–277, 1913.

[3] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.," *Nature*, vol. 181, no. 4610, pp. 662–666, Mar. 1958.

[4] J. C. Kendrew, *The Thread of Life*. Cambridge, MA, Havard University Press, 1966.

[5] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. North, "Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis.," *Nature*, vol. 185, no. 4711, pp. 416–422, Feb. 1960.

[6] K. Wüthrich, "The way to NMR structures of proteins.," *Nat. Struct. Biol.*, vol. 8, no. 11, pp. 923–925, Nov. 2001.

[7] G. M. Clore, "Adventures in Biomolecular NMR," *Encyclopedia of Magnetic Resonance*, vol. 132, pp. 1–7, 2011.

[8] M. Baker, "Cryo-electron microscopy shapes up.," *Nature*, vol. 561, no. 7724, pp. 565–567, Sep. 2018.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank.," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, Jan. 2000.

[10] M. Levitt, "The birth of computational structural biology.," *Nat. Struct. Biol.*, vol. 8, no. 5, pp. 392–393, May 2001.

[11] M. Levitt and S. Lifson, "Refinement of protein conformations using a macromolecular energy minimization procedure.," *Journal of Molecular Biology*, vol. 46, no. 2, pp. 269–279, Dec. 1969.

[12] C. B. Anfinsen, "Principles that govern the folding of protein chains.," *Science*, vol. 181, no. 4096, pp. 223–230, Jul. 1973.

[13] R. P. Feynman, "Simulating physics with computers," *International Journal of Theoretical Physics*, vol. 21, no. 6, 1982.

[14] C. Levinthal, "How to Fold Graciously," *Mossbauer Spectroscopy in Biological Systems Proceedings of a meeting held at Allerton House, Monticello, Illinois.*, pp. 22–24, 1969.

[15] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)-Round XII," *Proteins*, vol. 86, no. 1, pp. 7–15, Dec. 2017.

[16] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction: Progress and new directions in round XI," *Proteins*, vol. 84, pp. 4–14, 2016.

[17] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) — round x," *Proteins*, vol. 82, no. 2, pp. 1–6, 2014.

[18] R. Samudrala and J. Moult, "An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction.," *Journal of Molecular Biology*, vol. 275, no. 5, pp. 895–916, Feb. 1998.

[19] M. J. Sippl, "Knowledge-based potentials for proteins.," *Current Opinion in Structural Biology*, vol. 5, no. 2, pp. 229–235, Apr. 1995.

[20] C. M. Summa and M. Levitt, "Near-native structure refinement using in vacuo energy minimization.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 104, no. 9, pp. 3177–3182, Feb. 2007.

[21] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, Jul. 2003.

[22] E. I. Juritz, S. F. Alberti, and G. D. Parisi, "PCDB: a database of protein conformational diversity," *Nucleic Acids Research*, vol. 39, no. 1, pp. D475–D479, Nov. 2010.

[23] K. U. Linderstrøm-Lang, *Lane medical lectures: proteins and enzymes*. Stanford, CA, Stanford University Press, 1952.

[24] J. A. Schellman and C. G. Schellman, "Kaj Ulrik Linderstrøm-Lang (1896–1959)," *Protein Science*, vol. 6, no. 5, pp. 1092–1100, May 1997.

[25] R. B. Corey and L. Pauling, "Fundamental dimensions of polypeptide chains.," *Proceedings of the Royal Society of London, Series B, Biological Sciences*, vol. 141, no. 902, pp. 10–20, Mar. 1953.

[26] B. W. Matthews, "Studies on protein stability with T4 lysozyme.," *Adv. Protein Chem.*, vol. 46, pp. 249–278, 1995.

[27] A. Matouschek and A. R. Fersht, "Protein engineering in analysis of protein folding pathways and stability.," *Meth. Enzymol.*, vol. 202, pp. 82–112, 1991.

[28] A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, and A. G. Murzin, "SCOP2 prototype: a new approach to protein structure mining.," *Nucleic Acids Research*, vol. 42, no. Database issue, pp. D310–4, Jan. 2014.

[29] N. L. Dawson, T. E. Lewis, S. Das, J. G. Lees, D. Lee, P. Ashford, C. A. Orengo, and I. Sillitoe, "CATH: an expanded resource to predict protein function through structure and sequence.," *Nucleic Acids Research*, vol. 45, no. 1, pp. D289–D295, Jan. 2017.

[30] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem.," *Annu Rev Biophys*, vol. 37, no. 1, pp. 289–316, 2008.

[31] K. A. Dill and J. L. MacCallum, "The Protein-Folding Problem, 50 Years On," *Science*, vol. 338, no. 6110, pp. 1042–1046, Nov. 2012.

[32] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool.," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[33] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[34] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints.," *Journal of Molecular Biology*, vol. 234, no. 3, pp. 779–815, Dec. 1993.

[35] P. K. Warme, F. A. Momany, S. V. Rumball, R. W. Tuttle, and H. A. Scheraga, "Computation of structures of homologous proteins. Alpha-lactalbumin from lysozyme.," *Biochemistry*, vol. 13, no. 4, pp. 768–782, Feb. 1974.

[36] D. T. Jones, W. R. Taylor, and J. M. Thornton, "A new approach to protein fold recognition.," *Nature*, vol. 358, no. 6381, pp. 86–89, Jul. 1992.

[37] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, Mar. 1970.

[38] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, Mar. 1981.

[39] A. Fiser and A. Sali, "Modeller: generation and refinement of homology-based protein structure models.," *Meth. Enzymol.*, vol. 374, pp. 461–491, 2003.

[40] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, J. Moult, T. Schwede, and A. Tramontano, "Evaluation of the template-based modeling in CASP12," *Proteins*, vol. 86, no. 5, pp. 321–334, Dec. 2017.

[41] R. Bonneau and D. Baker, "Ab initio protein structure prediction: progress and prospects.," *Annu Rev Biophys Biomol Struct*, vol. 30, no. 1, pp. 173–189, 2001.

[42] R. Samudrala, Y. Xia, E. Huang, and M. Levitt, "Ab initio protein structure prediction using a combined hierarchical approach.," *Proteins*, vol. 3, pp. 194–198, 1999.

[43] P. Bradley, L. Malmström, B. Qian, J. Schonbrun, D. Chivian, D. E. Kim, J. Meiler, K. M. S. Misura, and D. Baker, "Free modeling with Rosetta in CASP6.," *Proteins*, vol. 61, no. 7, pp. 128–134, 2005.

[44] C. Hardin, T. V. Pogorelov, and Z. Luthey-Schulten, "Ab initio protein structure prediction.," *Current Opinion in Structural Biology*, vol. 12, no. 2, pp. 176–181, Apr. 2002.

[45] A. Liwo, J. Lee, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, "Protein structure prediction by global optimization of a potential energy function.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 96, no. 10, pp. 5482–5485, May 1999.

[46] A. Kolinski and J. Skolnick, "Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme.," *Proteins*, vol. 18, no. 4, pp. 338–352, Apr. 1994.

[47] U. H. Hansmann and Y. Okamoto, "New Monte Carlo algorithms for protein folding.," *Current Opinion in Structural Biology*, vol. 9, no. 2, pp. 177–183, Apr. 1999.

[48] Y. Zhang, D. Kihara, and J. Skolnick, "Local energy landscape flattening: Parallel hyperbolic Monte Carlo sampling of protein folding," *Proteins*, vol. 48, no. 2, pp. 192–201, Jun. 2002.

[49] M. Levitt, "Molecular dynamics of native protein. I. Computer simulation of trajectories.," *Journal of Molecular Biology*, vol. 168, no. 3, pp. 595–617, Aug. 1983.

[50] M. Levitt, "Molecular dynamics of native protein. II. Analysis and nature of motion.," *Journal of Molecular Biology*, vol. 168, no. 3, pp. 621–657, Aug. 1983.

[51] M. Feig and V. Mirjalili, "Protein structure refinement via molecular-dynamics simulations: What works and what does not?," *Proteins*, vol. 84, pp. 282–292, Aug. 2015.

[52] M. R. Lee, J. Tsai, D. Baker, and P. A. Kollman, "Molecular dynamics in the endgame of protein structure prediction," *Journal of Molecular Biology*, vol. 313, no. 2, pp. 417–430, Oct. 2001.

[53] V. Mirjalili and M. Feig, "Protein Structure Refinement through Structure Selection and Averaging from Molecular Dynamics Ensembles," *J Chem Theory Comput*, vol. 9, no. 2, pp. 1294–1303, Dec. 2012.

[54] H. A. Scheraga, M. Khalili, and A. Liwo, "Protein-Folding Dynamics: Overview of Molecular Simulation Techniques," *Annu. Rev. Phys. Chem.*, vol. 58, no. 1, pp. 57–83, May 2007.

[55] N. A. Bernhardt, W. Xi, W. Wang, and U. H. E. Hansmann, "Simulating Protein Fold Switching by Replica Exchange with Tunneling," *J Chem Theory Comput*, vol. 12, no. 11, pp. 5656–5666, Oct. 2016.

[56] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, and M. D. Peraro, "Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods," *Proteins*, vol. 86, no. 1, pp. 97–112, Nov. 2017.

[57] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, and A. M. J. J. Bonvin, "Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age.," *Proteins*, vol. 86, pp. 51–66, Mar. 2018.

[58] L. Sutto, S. Marsili, A. Valencia, and F. L. Gervasio, "From residue coevolution to protein conformational ensembles and functional dynamics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, no. 44, pp. 13567–13572, Nov. 2015.

[59] L. Hovan, V. Oleinikovas, H. Yalinca, A. Kryshtafovych, G. Saladino, and F. L. Gervasio, "Assessment of the model refinement category in CASP12," *Proteins*, vol. 86, no. 9, pp. 152–167, Nov. 2017.

[60] Y. Zhang, "Protein structure prediction: when is it useful?," *Current Opinion in Structural Biology*, vol. 19, no. 2, pp. 145–155, Apr. 2009.

[61] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, and E. L. Barker, "Computational Methods in Drug Discovery," *Pharmacol Rev*, vol. 66, no. 1, pp. 334–395, Jan. 2014.

[62] M. Feig, "Computational protein structure refinement: Almost there, yet still so far to go.," *Wiley Interdiscip Rev Comput Mol Sci*, vol. 7, no. 3, May 2017.

[63] M. Levitt, "Energy Refinement of Hen Egg-white Lysozyme," *Journal of Molecular Biology*, vol. 82, pp. 393–420, 1974.

[64] M. Levitt, M. Hirshberg, R. Sharon, and V. Daggett, "Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution," *Computer Physics Communications*, vol. 91, pp. 215–231, 1995.

[65] C. M. Summa, M. Levitt, and W. F. DeGrado, "An Atomic Environment Potential for use in Protein Structure Prediction," *Journal of Molecular Biology*, vol. 352, no. 4, pp. 986–1001, Sep. 2005.

[66] M. Levitt, "Molecular dynamics of native protein. I. Computer simulation of trajectories.," *Journal of Molecular Biology*, vol. 168, no. 3, pp. 595–617, Aug. 1983.

[67] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1, pp. 503–528.

[68] M. J. Sippl, "Calculation of Conformational Ensembles from Potentials of Mean Force," *Journal of Molecular Biology*, vol. 213, pp. 859–883, 1990.

[69] H. Lu and J. Skolnick, "Application of statistical potentials to protein structure refinement from low resolution ab initio models.," *Biopolymers*, vol. 70, no. 4, pp. 575–584, Dec. 2003.

[70] H. Zhou and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," *Protein Science*, vol. 11, no. 11, pp. 2714–2726, Apr. 2009.

[71] H. Lu and J. Skolnick, "A distance-dependent atomic knowledge-based potential for improved protein structure selection.," *Proteins*, vol. 44, no. 3, pp. 223–232, Aug. 2001.

[72] G. Chopra, C. M. Summa, and M. Levitt, "Solvent dramatically affects protein structure refinement," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 105, no. 51, pp. 20239–20244, Dec. 2008.

[73] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," *Journal of Computational ...*, 2000.

[74] E. J. Sorin and V. S. Pande, "Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations," *Biophysics*, vol. 88, no. 4, pp. 2472–2493, Apr. 2005.

[75] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, "Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides," *J. Phys. Chem. B*, vol. 105, no. 28, pp. 6474–6487, Jul. 2001.

[76] W. R. P. Scott and E. A., "The GROMOS biomolecular simulation program package," *J. Phys. Chem.*, vol. 103, pp. 3596–3607, 1999.

[77] M. Levitt, M. Hirshberg, R. Sharon, K. E. L. and, and V. Daggett, "Calibration and Testing of a Water Model for Simulation of the Molecular Dynamics of Proteins and Nucleic Acids in Solution," *J. Phys. Chem.*, vol. 101, pp. 5051–5061, Jun. 1997.

[78] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures.," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, Apr. 1995.

[79] M. Tirion, "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis.," *Phys. Rev. Lett.*, vol. 77, no. 9, pp. 1905–1908, Aug. 1996.

[80] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: Fast, flexible, and free," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005.

[81] E. Lindahl, B. Hess, and D. Van Der Spoel, "GROMACS 3.0: a package for molecular simulation and trajectory analysis," *J Mol Model*, vol. 7, no. 8, pp. 306–317.

[82] H. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: a message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, pp. 43–56, 1995.

[83] G. Chopra, N. Kalisman, and M. Levitt, "Consistent refinement of submitted models at CASP using a knowledge-based potential.," *Proteins*, vol. 78, no. 12, pp. 2668–2678, Sep. 2010.

[84] E. Eyal, S. Gerzon, V. Potapov, M. Edelman, and V. Sobolev, "The Limit of Accuracy of Protein Modeling: Influence of Crystal Packing on Protein Structure," *Journal of Molecular Biology*, vol. 351, no. 2, pp. 431–442, Aug. 2005.

[85] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation.," *Journal of Molecular Biology*, vol. 285, no. 4, pp. 1735–1747, Jan. 1999.

[86] J. M. Word, S. C. Lovell, T. H. LaBean, H. C. Taylor, M. E. Zalis, B. K. Presley, J. S. Richardson, and D. C. Richardson, "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms.," *Journal of Molecular Biology*, vol. 285, no. 4, pp. 1711–1733, Jan. 1999.

[87] R. A. Engh, R. Huber, IUCr, "Accurate bond and angle parameters for X-ray protein structure refinement," *Acta Crystallogr., A, Found. Crystallogr.*, vol. 47, no. 4, pp. 392–400, Jul. 1991.

[88] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom contacts and structure validation for proteins and nucleic acids," *Nucleic Acids Research*, vol. 35, no. Web Server, pp. W375–W383, May 2007.

[89] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography.," *Acta Crystallogr D Biol Crystallogr*, vol. 66, no. 1, pp. 12–21, Jan. 2010.

[90] R. Samudrala and J. Moult, "A graph-theoretic algorithm for comparative modeling of protein structure," *Journal of Molecular Biology*, vol. 279, no. 1, pp. 287–302, May 1998.

[91] E. R. Gansner and S. C. North, "An open graph visualization system and its applications to software engineering," *Softw. Pract. Exper.*, vol. 30, pp. 1203–1233, 2000.

[92] J. P. G. L. M. Rodrigues, M. Levitt, and G. Chopra, "KoBaMIN: a knowledge-based minimization web server for protein structure refinement.," *Nucleic Acids Research*, vol. 40, no. Web Server issue, pp. W323–8, Jul. 2012.

[93] I. Kufareva and R. Abagyan, "Methods of protein structure comparison.," *Methods Mol. Biol.*, vol. 857, pp. 231–257, 2012.

[94] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, vol. 57, no. 4, pp. 702–710, 2004.

[95] M. Levitt and M. Gerstein, "A unified statistical framework for sequence comparison and structure comparison.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 95, no. 11, pp. 5913–5920, May 1998.

[96] P. Antczak, T. Ratajczak, J. Blazewicz, and P. Lukasiak, "SphereGrinder-reference structure-based tool for quality assessment of protein structural models," presented at the IEEE International Conference on Bioinformatics and Biomedicine BIBM, 2015, pp. 665–668.

[97] 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, May 1997.

[98] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.," *Bioinformatics*, vol. 29, no. 21, pp. 2722–2728, Nov. 2013.

[99] Schrödinger, LLC, "The PyMOL Molecular Graphics System, Version 1.7.4.4"

[100] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004.

[101] C.-M. Li, Z. Fang, and K. Xu, "Combining MaxSAT Reasoning and Incremental Upper Bound for the Maximum Clique Problem," presented at the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 939–946.
[102] J.-M. Chandonia and S. E. Brenner, "The impact of structural genomics: expectations and outcomes.," *Science*, vol. 311, no. 5759, pp. 347–351, Jan. 2006.

Appendix

A.1 Lists of Omitted PDBs for the Generation of KB_Top500_1.00vdw and KB_Top8000_vdw

The 'H' appended to every PDB name indicates that hydrogens have been built into to them by the Reduce program (see section 3.2.1.2 for further details about these databases). This is true of all PDBs in the Top500 and Top8000 databases. In the Top8000 database, PDBs are split by chain. The chain ID is indicated following an '_'.

A.1.1 Omitted PDBs from Top500 for the Generation of KB_Top500_1.00vdw

1a6mH	1cf9BH	1lkkH	1sluH	2tnfAH
1aayH	1cgoH	1mbaH	1tgsIH	3claH
1aqbH	1cl8H	1mdcH	1ttbAH	3pteH
1b9wH	1fusH	1mmlH	1tudH	3sebH
1babBH	1gaiH	1oncH	1ubpH	3stdAH
1bdmBH	1gciH	1phnAH	2bbkLH	5cytH
1becH	1gd1OH	1qgqH	2bopAH	9wgaAH
1btyH	1gsoH	1qgwBDH	2cbaH	
1bu8H	1guqAH	1qnfH	2hmzAH	
1bueH	1hmtH	1qnjH	2pvbH	
1ceqH	1htrH	1rhsH	2qwcH	

A.1.2 Omitted PDBs from Top8000 for the Generation of KB_Top8000_1.00vdw

1a7tFH_B	1bueFH_A	1deuFH_A	1eltFH_A	1fusFH_A
1ayeFH_A	1bxuFH_A	1dl2FH_A	1eq9FH_B	1g6aFH_A
1b63FH_A	1cjcFH_A	1dpjFH_A	1f7bFH_C	1gaiFH_A
1bsgFH_A	1d5tFH_A	1e25FH_A	1fj2FH_B	1gciFH_A
1bu8FH_A	1dciFH_C	1ejdFH_B	1fljFH_A	1gpiFH_A

1gppFH_A
1gpuFH_A
1gvzFH_A
1h0hFH_K
1h4aFH_X
1hj8FH_A
1hleFH_A
1hmtFH_A
1hp1FH_A
1hpgFH_A
1hx0FH_A
1hzoFH_A
1iuzFH_A
1jd0FH_B
1jltFH_A
1jltFH_B
1k07FH_A
1k3iFH_A
1k75FH_B
1ka1FH_A
1kgcFH_D
1ku1FH_A
1lo6FH_A
1m2xFH_D
1m40FH_A
1m8sFH_A
1mc2FH_A
1mdoFH_A
1me4FH_A
1mexFH_H
1mj5FH_A
1mn8FH_B
1n12FH_A
1n63FH_E
1n9pFH_A
1nlnFH_A
1nrjFH_A
1nu0FH_A
1nxoFH_A
1o0eFH_B
1o7eFH_B
1o82FH_A
1odmFH_A
1ongFH_A
1ox0FH_A
1oxsFH_C
1pa2FH_A
1pfzFH_A
1pzgFH_A
1qnjFH_A
1qouFH_B
1qwgFH_A
1qwoFH_A
1qxyFH_A
1r0rFH_E
1r0uFH_A
1r6wFH_A
1r8hFH_D
1rhcFH_A
1rutFH_X
1rwhFH_A
1rypFH_J
1rypFH_K
1s1fFH_A
1spjFH_A
1syyFH_A
1t0bFH_D
1to4FH_A
1tt2FH_A
1u2bFH_A
1u6eFH_A
1uixFH_A
1ulrFH_A
1ut7FH_B
1v05FH_A
1v0wFH_A
1v54FH_A
1vmeFH_B
1vmhFH_A
1vr5FH_B
1vr8FH_A
1vyfFH_A
1vzyFH_B
1w0nFH_A
1w1qFH_A
1w32FH_A
1w3wFH_A
1w7cFH_A
1wb0FH_A
1wl8FH_A
1wrmFH_A
1x0lFH_A
1x38FH_A
1xdwFH_A
1xiyFH_A
1xsoFH_B
1xx1FH_C
1y2mFH_C
1y63FH_A
1y7tFH_B
1y81FH_A
1yg9FH_A
1ynpFH_B
1yxyFH_A
1z57FH_A
1z76FH_B
1z7aFH_D
1zd0FH_A
1zi9FH_A
1zr0FH_D
1zr6FH_A
1zsxFH_A
1zuuFH_A
1zx8FH_C
1zzkFH_A
2anyFH_A
2apxFH_A
2b6nFH_A
2bbaFH_A
2bcmFH_B
2bezFH_C
2bkrFH_A
2bw0FH_A
2bz6FH_H
2cayFH_A
2cjzFH_A
2cn0FH_H
2d1gFH_A
2e7zFH_A
2eq6FH_B
2ex4FH_A
2f8aFH_A
2f91FH_A
2f9nFH_B
2fdsFH_A
2fgrFH_A
2fhxFH_B
2fm6FH_A
2fosFH_A
2fpqFH_A
2fueFH_A
2gaiFH_B
2gasFH_A
2gauFH_A
2gbwFH_E
2h0uFH_A
2h12FH_A
2h26FH_A
2h4pFH_A
2h5cFH_A
2h6eFH_A
2h8oFH_A
2hbvFH_A
2hc1FH_A
2he2FH_A
2hekFH_B
2heuFH_B
2hl7FH_A
2hlcFH_A
2hlvFH_A
2ht9FH_B
2hy7FH_A
2hyxFH_D
2i0qFH_A
2icrFH_A
2idlFH_B
2ijxFH_D
2in8FH_A
2ip2FH_B
2iw1FH_A
2iwzFH_A
2j97FH_A
2j9cFH_B
2jdfFH_A
2jikFH_A
2jilFH_A
2jisFH_A
2jkhFH_A
2jliFH_A
2nw2FH_B
2oblFH_A
2okmFH_A
2opcFH_A
2oqbFH_A
2ouaFH_A
2oxgFH_Y
2p49FH_B
2p74FH_A
2pfeFH_B
2pi6FH_A
2pltFH_A
2pmqFH_A
2pmrFH_A
2pq8FH_A
2pqmFH_B
2pvbFH_A
2pzeFH_B
2q0uFH_A
2q2hFH_A
2q7wFH_A
2qa9FH_E
2qeeFH_F
2qmjFH_A
2qmqFH_A
2qruFH_A
2qudFH_A
2qvbFH_A
2qvoFH_A
2qwcFH_A
2qxiFH_A
2r16FH_A
2r1bFH_B
2ra3FH_B
2rhfFH_A
2sgaFH_A
2tnfFH_B
2uurFH_A
2uuuFH_C
2uv4FH_A
2uw1FH_A
2uxqFH_A
2uxwFH_A
2v03FH_A
2v5iFH_A
2vacFH_A
2vifFH_A
2vngFH_A
2vo8FH_A
2vo9FH_B
2vphFH_B

2vq8FH_A
2vqpFH_A
2vsvFH_A
2vwrFH_A
2vx5FH_A
2vxtFH_H
2vxtFH_I
2vzmFH_A
2w0iFH_A
2w98FH_B
2wb6FH_A
2weiFH_A
2welFH_A
2wj5FH_A
2wk0FH_A
2wkkFH_C
2wnpFH_F
2wnxFH_A
2wolFH_A
2woyFH_A
2wtgFH_A
2wweFH_A
2wwfFH_C
2wwxFH_B
2wyqFH_A
2x26FH_B
2x49FH_A
2x4jFH_A
2x5pFH_A
2x7bFH_A
2x98FH_B
2xbpFH_A
2xdeFH_A
2xdgFH_A
2xi9FH_B
2xn6FH_A
2xsuFH_A
2xttFH_B
2xu7FH_B
2xvsFH_A
2yrxFH_A
2ywnFH_A
2yxwFH_A
2yzhFH_C
2z66FH_B
2z7fFH_E
2zxyFH_A
2zyaFH_B
3a3dFH_B
3a40FH_X
3a4rFH_A
3aarFH_A
3abdFH_B
3ajoFH_A
3b7eFH_A
3b7sFH_A
3b9tFH_A
3beuFH_B
3bfvFH_A
3bixFH_A
3bj1FH_C
3bn7FH_A
3bvkFH_F
3c5aFH_A
3c5eFH_A
3c9aFH_B
3c9xFH_A
3ccfFH_A
3ccgFH_A
3cecFH_A
3cfcFH_H
3ck6FH_B
3ckmFH_A
3claFH_A
3cmcFH_Q
3cn4FH_B
3coxFH_A
3d0oFH_A
3d4uFH_A
3d8tFH_A
3db7FH_A
3dmeFH_B
3dpkFH_A
3durFH_B
3dz1FH_A
3e6jFH_A
3ed7FH_A
3edgFH_A
3edvFH_A
3ee4FH_A
3eojFH_A
3er6FH_A
3eupFH_B
3ew0FH_A
3ewhFH_A
3eyiFH_A
3f5hFH_B
3f8tFH_A
3fdlFH_A
3fedFH_A
3ff9FH_B
3fg1FH_D
3fo3FH_A
3fw3FH_A
3fzyFH_B
3g0eFH_A
3g5sFH_A
3g6mFH_A
3g8yFH_A
3g9xFH_A
3ggwFH_B
3gkvFH_B
3gpkFH_B
3guyFH_B
3gvoFH_A
3gylFH_B
3h04FH_A
3h34FH_A
3h4nFH_A
3h9uFH_C
3hoiFH_A
3hr6FH_A
3hsrFH_D
3ht1FH_A
3hx8FH_D
3i09FH_A
3i10FH_A
3i2nFH_A
3i94FH_A
3iavFH_A
3iboFH_A
3ie5FH_A
3ie7FH_A
3ihvFH_A
3iofFH_A
3iq0FH_A
3isgFH_A
3iv4FH_A
3jqlFH_A
3js8FH_A
3jszFH_A
3jxoFH_A
3jzyFH_A
3k01FH_A
3kaxFH_A
3kcgFH_H
3kdwFH_A
3keoFH_B
3kkfFH_A
3kkgFH_A
3klkFH_A
3kqrFH_A
3kv1FH_A
3kz5FH_A
3kz7FH_A
3l0lFH_B
3l4rFH_A
3l7oFH_A
3l8aFH_B
3l91FH_B
3la7FH_B
3lgbFH_B
3llpFH_B
3lwkFH_A
3lwxFH_A
3lxpFH_A
3lxyFH_A
3ly7FH_A
3m70FH_A
3m7aFH_A
3m86FH_B
3maoFH_A
3mhsFH_A
3mhwFH_U
3mi4FH_A
3mm6FH_A
3mswFH_A
3mzvFH_B
3n3sFH_A
3n6yFH_B
3n7oFH_A
3nclFH_A
3nepFH_X
3njnFH_C
3nn1FH_A
3no3FH_A
3npdFH_A
3nqxFH_A
3nxgFH_E
3nyyFH_A
3o3uFH_N
3oa2FH_C
3oblFH_B
3obuFH_A
3ol0FH_A
3oseFH_A
3p1gFH_A
3p6lFH_A
3p9pFH_A
3pcvFH_A
3pe7FH_A
3pf2FH_A
3phsFH_A
3pjyFH_B
3pt1FH_A
3q4tFH_A
3q5yFH_A
3qe1FH_A
3qhzFH_M
3qqiFH_B
3qyqFH_C
4ubpFH_A
6rxnFH_A

A.2 Complete Atom Type Merge Graphs for KB_Top500 and KB_Top500_1.00vdw

On the next pages are given the complete atom type merging graphs for the four original PMFs (see Sections 3.2.1.2 and 3.3.1). These graphs are large and are broken into panes for formatting purposes. They were generated using the open source program GRAPHVIZ.

The following key applies to all of them. Each node represents an atom type and each atom type is enclosed in a colored polygon (with a circle used as an additional shape). Each atom type is colored coded to their element according to standard colors: nitrogen – blue, carbon – black, oxygen – red, and sulfur – yellow. Each atom type enclosed in a polygon indication its position in its amino acid. Backbone atoms (except Cαβγδεη) are enclosed in a diamond. For all other atom types, the number of sides of their polygon indicates side chain position. CA is enclosed in a triangle, Cβ a square, Cγ a pentagon, and so on. Atom types at the η level are enclosed in a circle.

Each level of the graph corresponds to a PMF. All 167 atom types on the first level are the atom types in the unmerged PMF. The atom types on the next level are those of the PMF generated after one iteration of merging, and so on. Merged atom type are denoted by a single atom type identifier. For example, after atom types FCA and LCA are merged in KB_Top500 Pane 1, the combined atom type is denoted as LCA.

A.2.1 Atom Type Merge Graph for KB_Top500

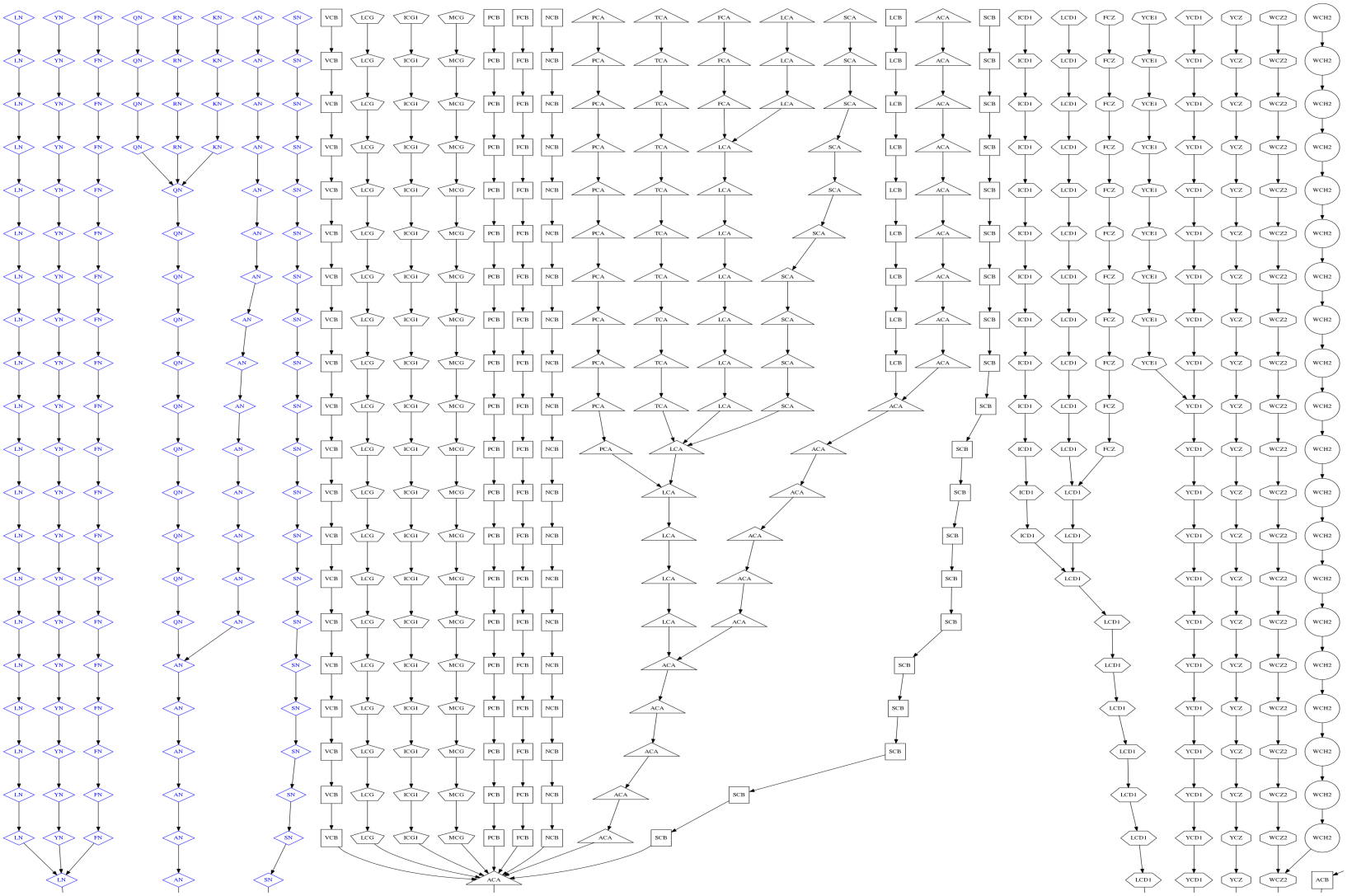

Figure A.1: Atom Type Merge Graph: KB_Top500, Pane 1

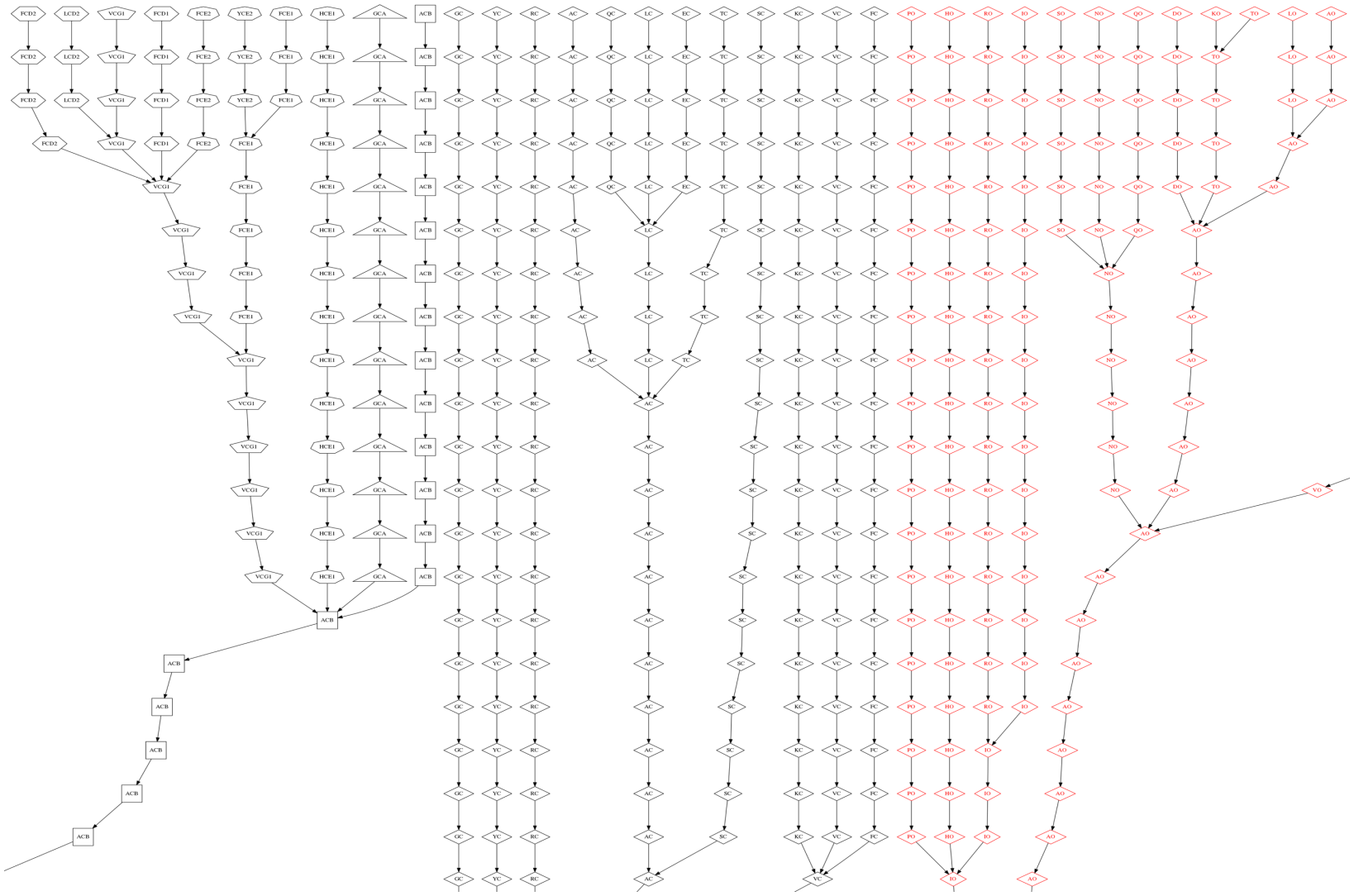

Figure A.2 Atom Type Merge Graph: KB_Top500, Pane 2

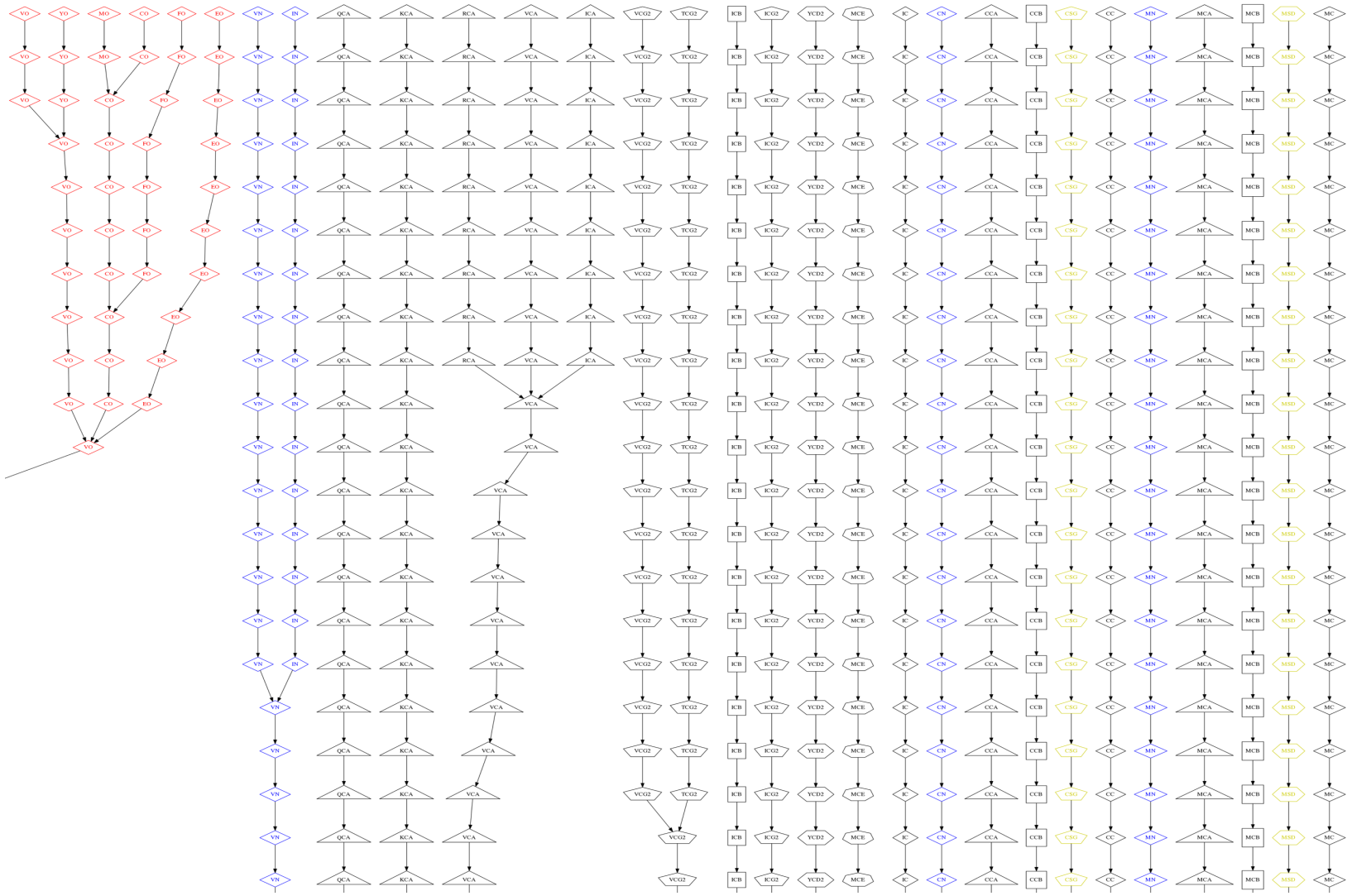

Figure A.3 Atom Type Merge Graph: KB_Top500, Pane 3

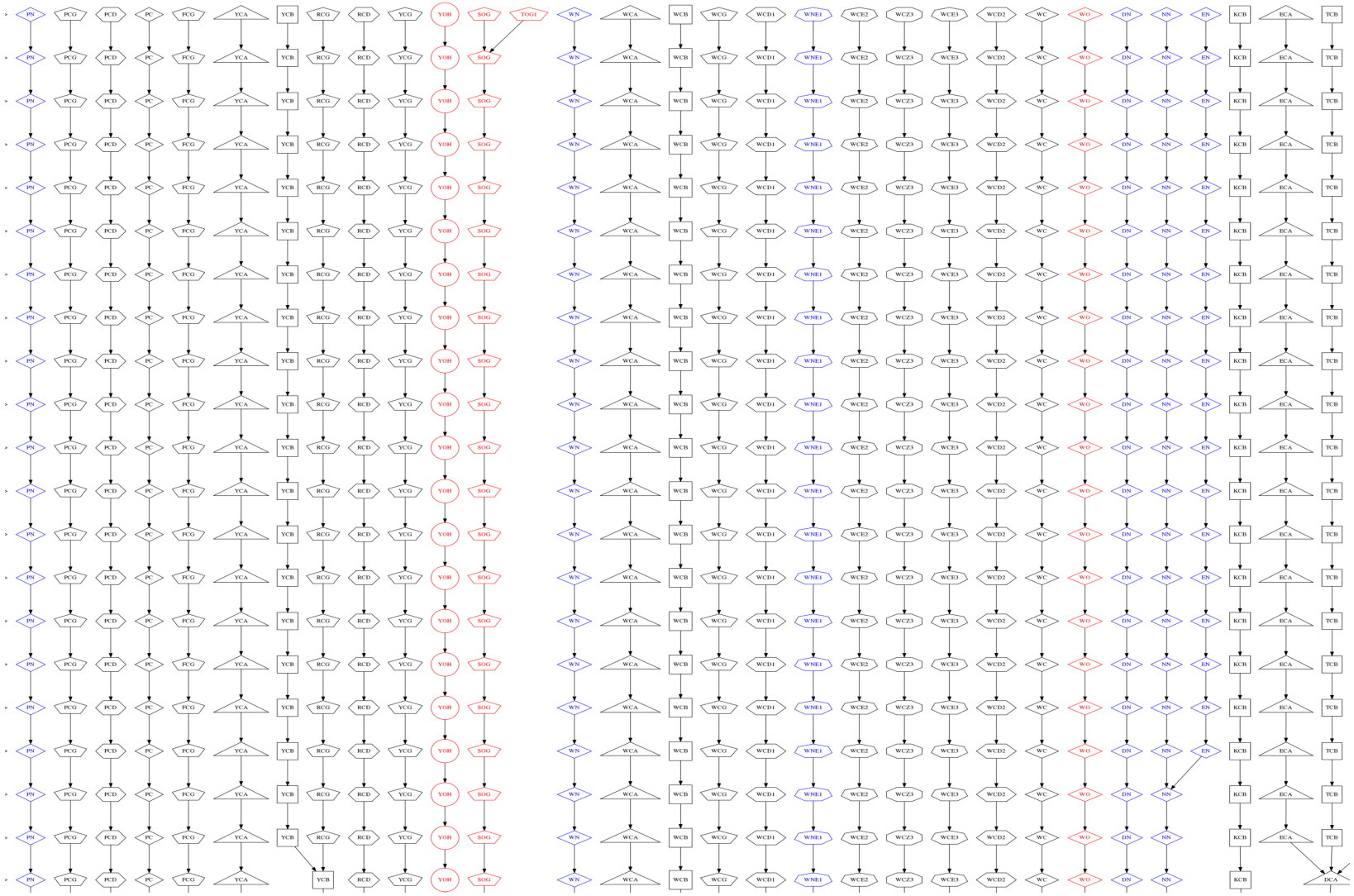

Figure A.4 Atom Type Merge Graph: KB_Top500, Pane 4

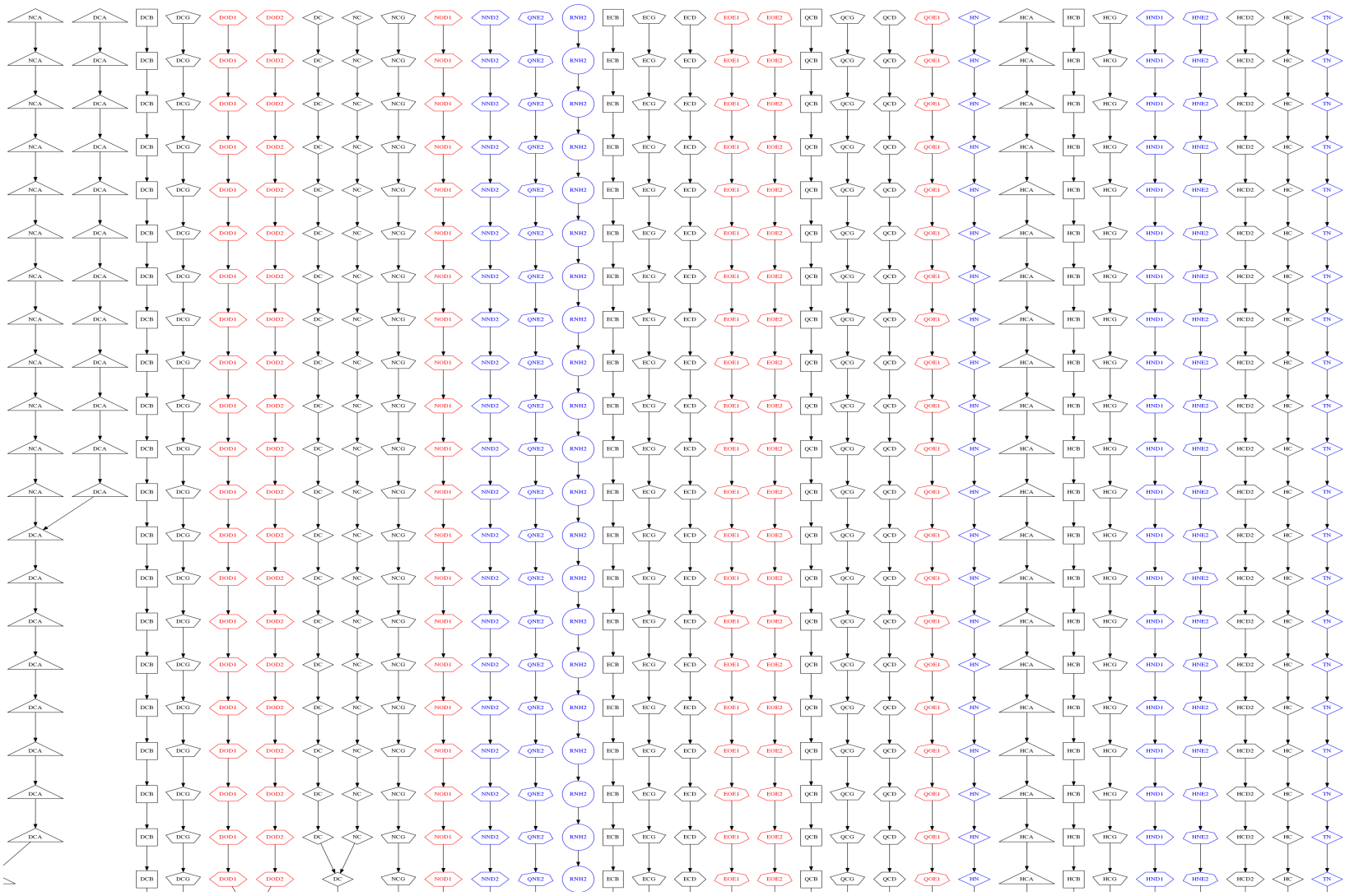

Figure A.5 Atom Type Merge Graph: KB_Top500, Pane 5

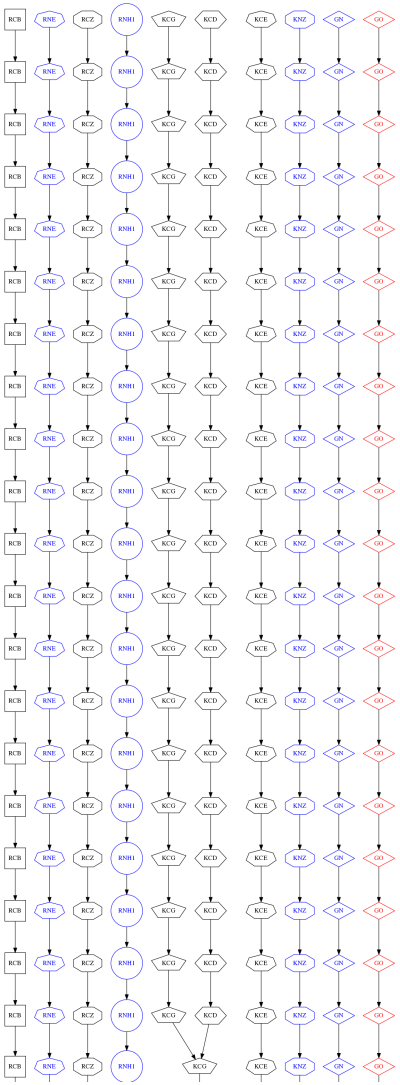

Figure A.6 Atom Type Merge Graph: KB_Top500, Pane 6

A.2.2 Atom Type Merge Graph for KB_Top500_1.00vdw

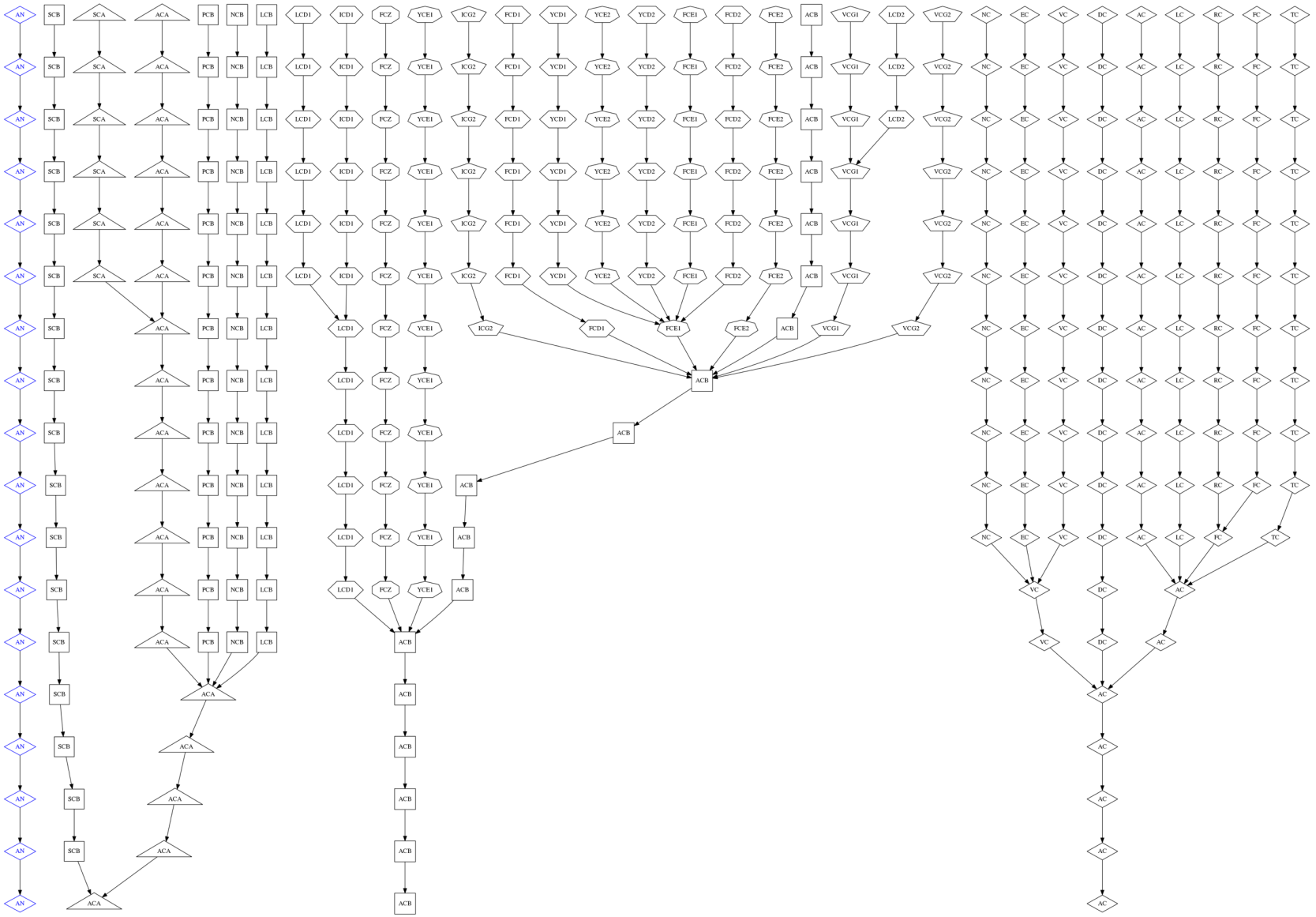

Figure A.7 Atom Type Merge Graph: KB_Top500_1.00vdw, Pane 1

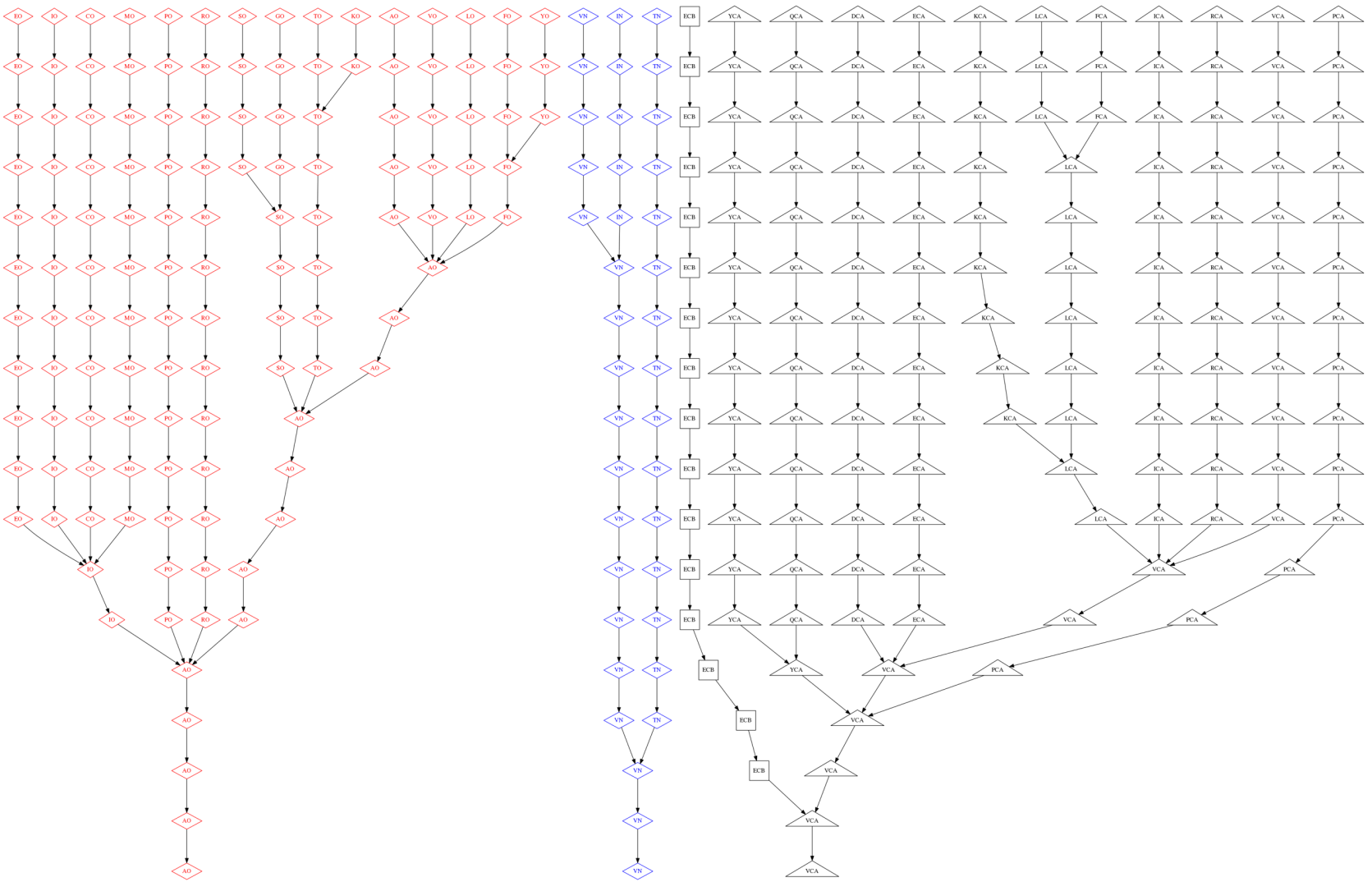

Figure A.8 Atom Type Merge Graph: KB_Top500_1.00vdw, Pane 2

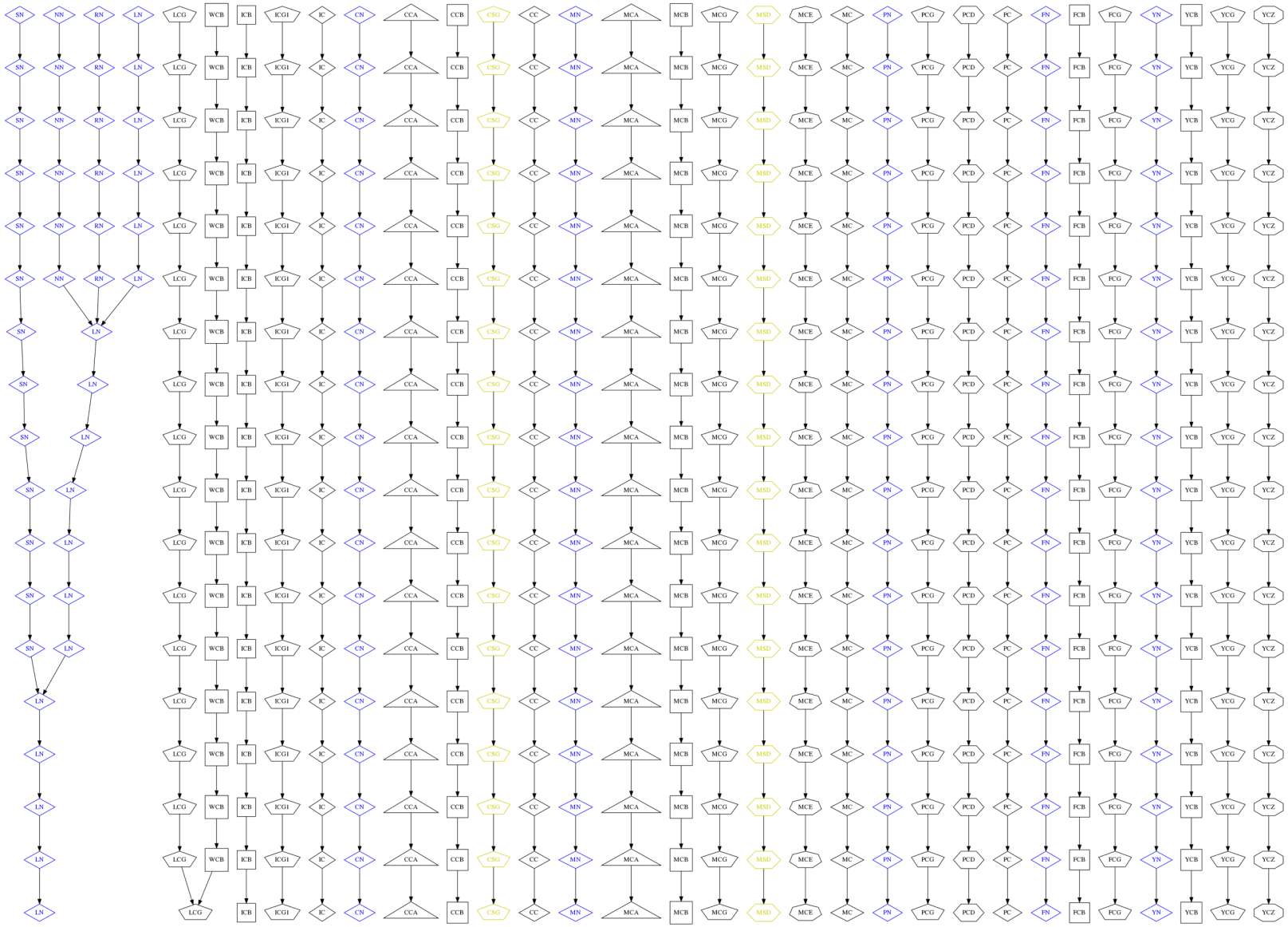

Figure A.9 Atom Type Merge Graph: KB_Top500_1.00vdw, Pane 3

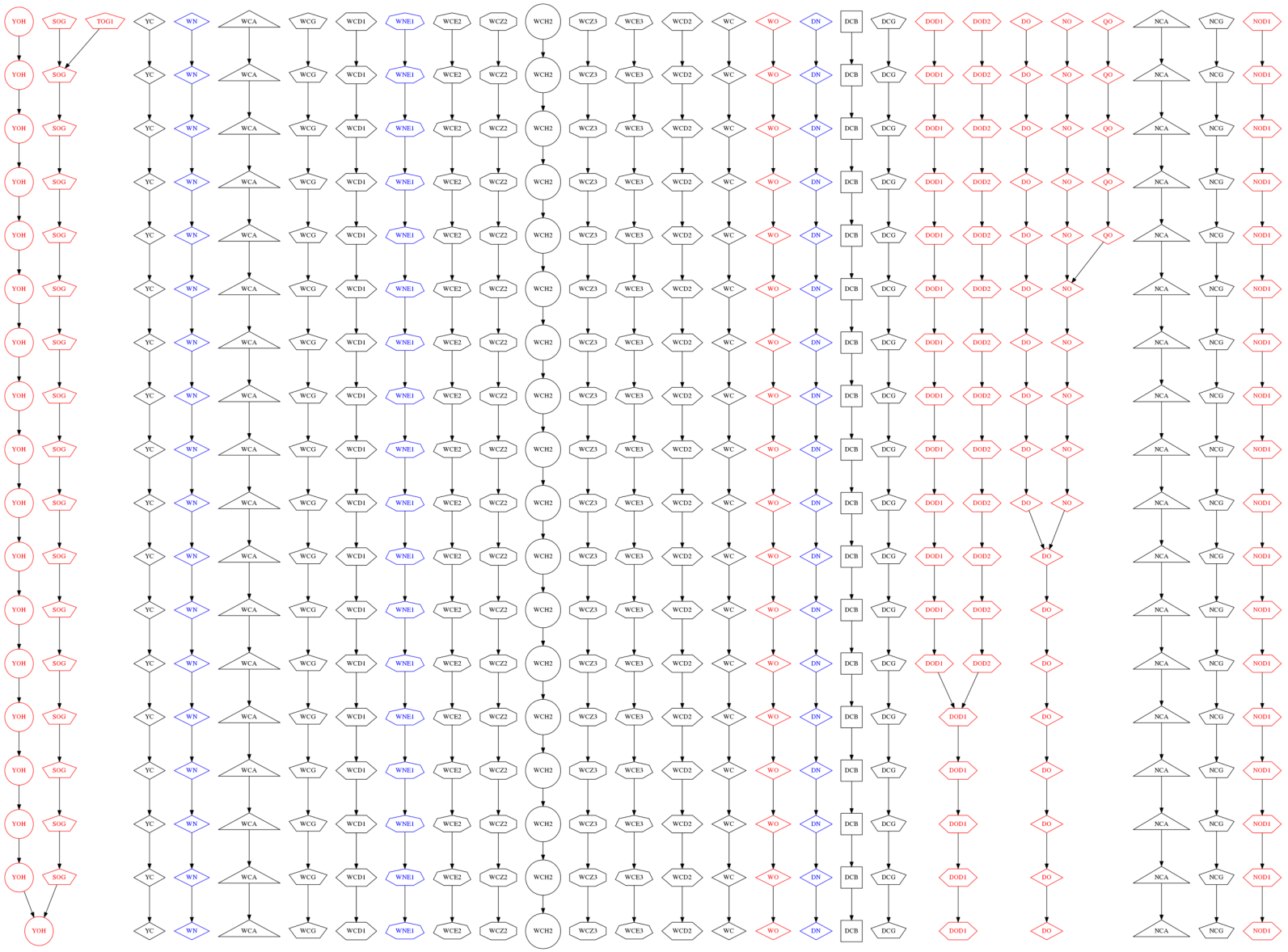

Figure A.10 Atom Type Merge Graph: KB_Top500_1.00vdw, Pane 4

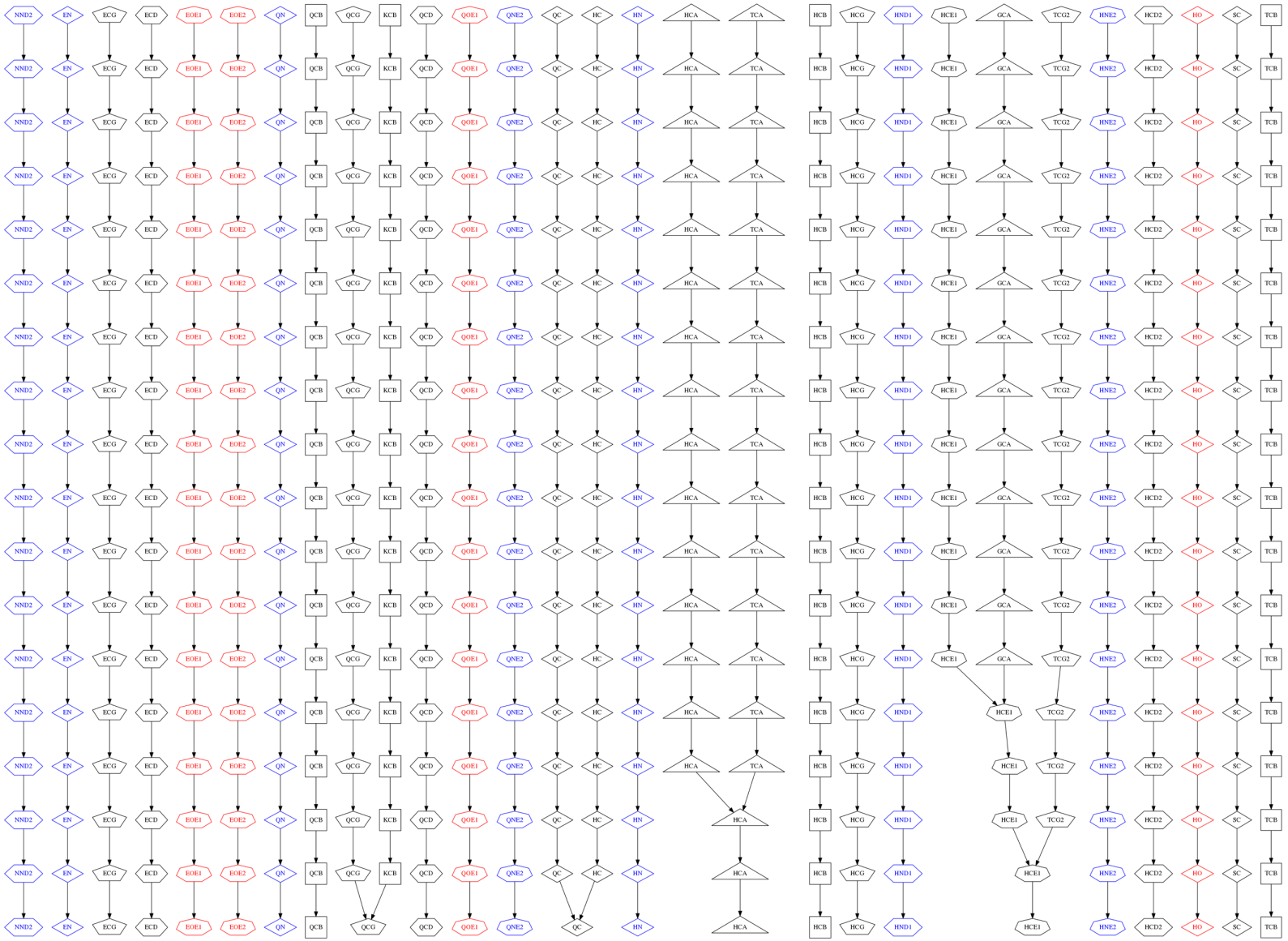

Figure A.11 Atom Type Merge Graph: KB_Top500_1.00vdw, Pane 5

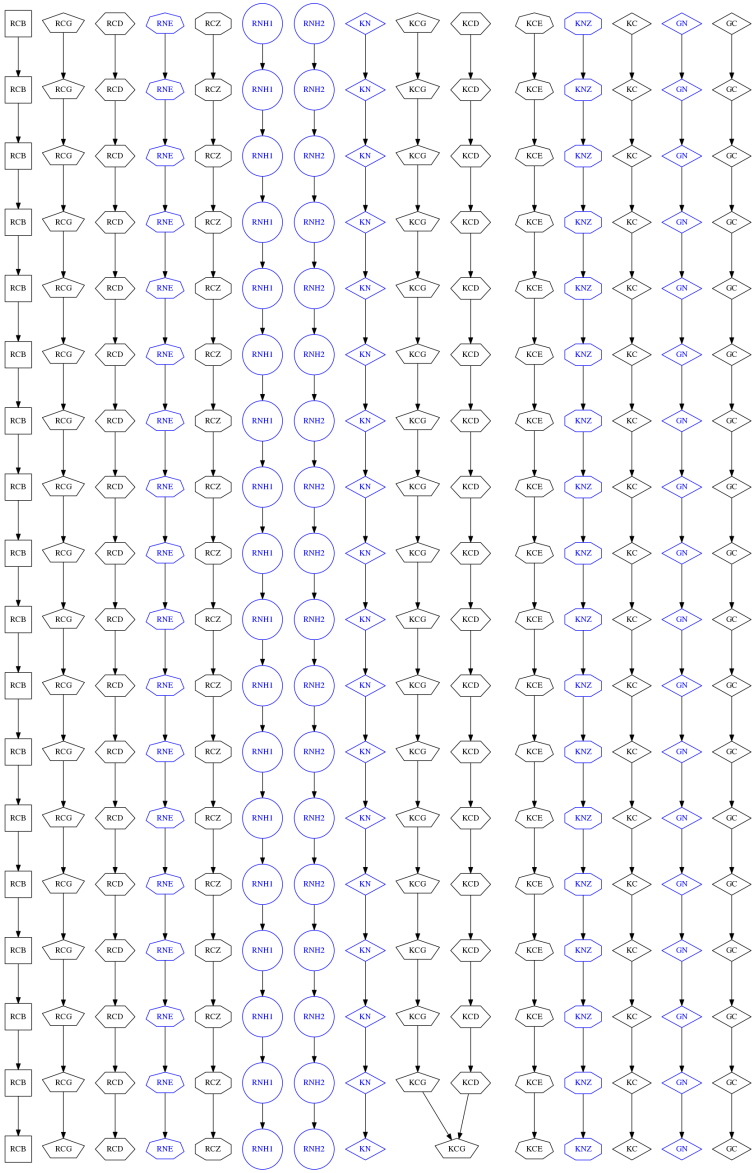

Figure A.12 Atom Type Merge Graph: KB_Top500_1.00vdw, Pane 6

Vita

The author was born in Slidell, Louisiana. He obtained his Bachelor's Degrees in Computer Science and Mathematics from the University of New Orleans (UNO) in 2011. In 2012 he joined the UNO computer science graduate program to pursue a PhD in Engineering and Applied Science (concentration: computer science) and became a member of the Summa Laboratory of the UNO computer science department. In the Summa Laboratory, he carried out his dissertation work as a graduate research assistant under the supervision of Dr. Christopher M. Summa.