3-4-2016

A Metallomics Study on Protein function assignment Using ProMOL

Venkata Aditya Kovuri
vsk5525@rit.edu

R·I·T

A Metallomics Study on Protein function assignment Using ProMOL

By

Venkata Aditya Kovuri

Submitted in partial fullfilment of the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

College of Science
Department of Biological Sciences
School of Life Sciences

Rochester Institute of Technology
Rochester, New York
March 04, 2016

Committee Approval:

Paul A. Craig, Ph.D.

(date)

Professor and Head / School of Chemistry and Materials Science
Thesis Project Advisor

Feng Cui, Ph.D.

(date)

Assistant Professor / School of Life Sciences
Committee Member

Rajendra K. Raj, Ph.D.

(date)

Professor / Computer Science Department
Committee Member

Abstract

Metal ions are an integral part in both the structural and functional stability of enzymes. The function of an enzyme is exhibited through its active site region, which carries out the reaction for that particular protein. When a metal ion is present in an active site, it not only acts as a structural support for the active site but is often plays a critical role in the functional behavior of the protein. An exhaustive screening for proposed metal ion active sites was performed through online resources including Metal MACiE, Metal PDB, the Catalytic site Atlas and Metal Mine. A total of 103 motifs were generated using ProMol, which is a plugin for the molecular visualization software PyMOL. ProMOL was able to identify these motifs in homologous structures. This new library of enzyme active site motifs was then used to suggest functions for three proteins of unknown function that are found in the Protein Data Bank. An additional functionality was added to ProMOL to enable PyMOL to visualize metal ions and prosthetic groups in enzyme active sites and to calculate the interatomic distances of the active site regions.

Table of Contents

Page #

List of Figures Page#

List of Tables Page#

List of Abbreviations

1. EC : Enzyme Commission
2. ATP : Adenosine triphosphate
3. PDB : Protein Data Bank
4. BRENDA : Braunschweig Enzyme Database
5. MACiE : Mechanism, Annotation and Classification in Enzymes
6. KEGG : Kyoto Encyclopedia of Genes and Genomes
7. SABIO-RK : System for the Analysis of Biochemical Pathways - Reaction Kinetics
8. IntEnz : Integrated relational Enzyme database
9. EMBL : European Molecular Biology Laboratory
10. CATH : Class, Architecture, Topology, Homologous superfamily.
11. MFS : Minimal Functional Sites
12. SCOP : Structural Classification of Proteins
13. PSI-BLAST : Position Specific Iterated Basic Local Alignment Search Tool
14. PKA : Protein Kinase A

1. Introduction

Pattern recognition is one of the more sought after fields in computer science these days from bio-scanners to streaming of text. This approach is being taken up by different bioinformatics programs that aim to achieve pattern recognition through local structural similarity in associating the function of a protein with known function and a query protein. ProMOL is one such program that aims to achieve characterization of proteins with unknown function through such pattern recognition principles [1].

ProMOL is a plugin for the PyMOL molecular graphics environment [2] that utilizes the 3-dimensional visualization and measurement capacities of PyMOL to align enzyme active sites with a library of active sites found in a motif template library. Through this method, ProMOL/PyMOL uses structural alignment to predict enzyme function for proteins of unknown function. The purpose for this initiative can be understood by looking at the number of structures in the Protein Data Bank that are labeled as having "unknown function"; to date, there are 3495 such structures in the Protein Data Bank.

An additional functionality for ProMOL to recognize metal ions in active sites can increase the motif template library considerably, as nearly 40% of all the PDB structures contain at least one metal ion [3]. This is biologically significant as metal ions such as Manganese (Mn), Zinc (Zn), Calcium (Ca), and Iron (Fe) play a vital role in both the structure and in the expression of function in many proteins [3]. The addition of metal ion recognition to ProMOL/PyMOL promises a better function assignment to this collection of protein structures. Using the current set of motif templates in ProMOL/PyMOL, nearly 12% of the 3495 structures of unknown function to date were assigned putative function [4]. Incorporation of metal ions in the search and alignment processes promises to increase this percentage significantly.

The decision to add the metal ion recognition to ProMOL was a natural extension of the project, propelled by a curiosity to understand the function of proteins that contain metal ions in their active

sites and the functionality that these metal ions bring. Many metal ions are an integral part of life [5]. Plants are no exception. Metal-binding proteins play a crucial role in the biological function of plants, such as nitrogen fixation [6] degradation of urea [7] and photosynthesis [8]. Plants are stationary so they cannot leave the environment that they are present in and not all geographical locations are favorable for a plant to survive. For example, they cannot leave the environments with toxic metal ions such as cadmium or arsenic, and cannot seek environments rich with the necessary metal ions. So, they develop several ways to sustain without the necessary metal ions. This inherently can even affect the evolutionary pathways and is an interesting avenue to pursue as it possesses a great potential due to its medical implication and its relevance in human systems.

One of the crucial roles of the metals in protein is their function as cofactors in enzymes. The functional and structural aspects of the study of metal-associated enzymes have garnered a certain curiosity from the field of bioinorganic chemistry in identifying the structural and functional aspects of a protein [3]. They also act as signal mediators with metal-sensing proteins, which have various applications [9, 10] along with the stabilization of protein structures [11]. Metal-binding proteins also act as metal chaperones in aiding the transport of metal ions [12]. In all structurally characterized proteins, around one-third structures contain metal ions [13]. The study of such metalloproteins can be constituted under metallomics. This indicates the interest and increase in the efforts to study and understand metalloproteins.

Enabling ProMOL to recognize metal ions as a part of the active site motifs allows users to create a wider set of motif templates and offers a different route from solely focusing on the active site motifs containing exclusively the 20 canonical amino acid residues. This new capability for ProMOL can help increase our understanding of metalloenzymes and potentially lead us to an increased rate in function assignment for metalloenzymes. There is no set of general rules that describes the behavior of a given

metal ion in an enzyme [14]. One of the major factors in the correct orientation of a substrate to a binding site is the electrostatic environment in the active site. Metal ions frequently assist in this process. Metal ions control the action of the enzyme by often binding groups in a stereo chemically rigid manner [15].

To carry out biological function, it is estimated that about 30-40% of proteins require a metal ion [16, 17]. Among them, magnesium is most frequently found in active sites when compared to other metal ions such as calcium, zinc, manganese, iron, copper, and molybdenum. These metal ions aid not only in the function of the protein but also in the catalysis of a reaction. Some metal ions provide their redox activity to the function of an enzyme, since some metal ions have several valence states [18].

The ability of a metal ion to compete effectively with a proton in a ligand binding is dictated in a large measure by the strength of the bond that both the ligand and the metal share. The metal-ligand bond is dependent on the detailed nature of the valence orbitals of ligands as well as the effective nuclear charge and coordination number and geometry of the metal ion [19]. Even small proteins have complex molecular architecture. Typically this architecture has evolved to promote function, which for metallobiomolecules is intimately coupled with the nature of the active site, which in turn is very critical to the function of some proteins [19].

“Metal ions generally bind in regions of high hydrophobicity contrast in proteins”[20]. This can be inferred from the electron distributions in metal ions that are highly symmetric, attracting electron-pair donors (Lewis bases) around the ion in a shell. Typically oxygen, nitrogen, and sulfur atoms are the electron pair donors in proteins. The same factors that cause these ligands to bind metal ions also cause them to be strongly solvated by water. Also there are other factors that contribute to such high hydrophobicity contrast in the metal binding sites. One may be that the hydrophobic sphere restricts flexibility of the site [21], reducing the decrease in entropy associated with binding. Such predisposition

for metal sites would be expected to favor binding [24]. While the soft metals such as Fe frequently bond to sulfur ligands from the side chains of cysteine and methionine residues, the bonding arrangements of harder metals like Na or Mg prefer oxygen and nitrogen as ligands. Despite these differences most metal sites are centered in a shell of hydrophilic ligands, surrounded by a shell of carbon-containing groups [20]. Zinc coordination site are predominantly present within the transmembrane domain and play a pivotal role in the Transcription factor proteins [21]. Ca^{2+} is often found in a structural role for example, Ca^{2+} in maintains the conformation of the N-terminal region of the EGF-like domain and demonstrates that Ca^{2+} can directly mediate protein-protein contacts structurally [22].

All metal ions exhibit a certain function both structurally and in terms of their catalytic function, for example magnesium exhibits such activity in ATPase and DNA polymerase. Calcium performs as a structural metal and also triggers intracellular messenger systems controlling processes such as muscle contraction, secretion, glycolysis, and ion transport [25]. A large majority of sites present in Metal-MACiE predominantly contain calcium or zinc [33]. The classic example of a structural zinc site is the tetrahedral binding site found in alcohol dehydrogenase [19]. By introducing the metal ion recognition capability to ProMOL, we can expand the target population of enzymes and also begin to codify motifs to allow structural alignment of non-catalytic metal ion motifs found in protein structures.

There are a number of databases that focus on enzymes that contain metal ions and their active sites. Two of the most well-known are Metal MACiE [26] and Metal PDB [3], which focus on set structures. One of my goals is to curate and accumulate meaningful data from these databases to provide candidate motif templates containing metal ions for use with ProMOL /PyMOL.

Our understanding of enzymes is very limited despite the fact that their diversity is the thing that makes the complexity of life possible. They help constitute organisms ranging from thermophiles that exist

survive in harsh conditions to complex higher organism like humans. We have an even less understanding of the complexity of their function. The Enzyme Commission (EC) first published their rules for enzyme nomenclature in 1964 along with a system to classify the overall reaction that an enzyme performs [27]. The first proteins with a fully defined sequence and assigned identifier from the curated portion of UniprotKB (Swiss-Prot) [28] were deposited in the 1980s, and the first crystal structures relating to an enzyme were deposited in the world wide PDB in the early 1970s [2]. Since then, the increase in the number of structures is exponential, but still there is still a lot of information that is yet to be understood.

Despite this lack of knowledge there is a lot of information that is still available on structures, gene sequences, mechanisms, metabolic pathways and kinetic data. However, these data are not available at one resource and throughout the literature they are spread between many different databases. Most web resources relating to enzymes [such as BRENDA [29], KEGG [30], SABIO-RK [31], the IUBMB Enzyme Nomenclature website [28] and IntEnz [32]] focus on the overall reaction, accompanied in some cases by a textual or graphical description of the mechanism. MACiE [33][35], which stands for Mechanism, Annotation and Classification in Enzymes, is a collaboration between the Thornton group (EMBL-EBI), Mitchell group (University of St Andrews, Scotland) and Bertini group (University of Florence, Italy) and was designed to provide a computational description of mechanism by including detailed stepwise mechanistic information for a wide coverage of both chemical space and protein structure.

The study of metal ions in biochemical process and in catalysis of reactions is an area of research that has intrigued many scientific minds and is certainly not a novelty. However, with the rise of data mining techniques and the insurgence of “Big Data”, a possibility of mining various available databases offers a new perspective to the study of metal ions in proteins in a high-level overview aided by the collection of large amounts of data in understanding of the basic tenets of metal-dependent catalysis through the information collected from various databases. This viewpoint stands in congruence with that of those

scientists who recognize the potential of the increasingly available 'omics' (e.g. genomics, proteomics) datasets, which help better understand cell biology, as it provides a detailed outlook of the molecular components of a cell that helps understand the complex mechanisms as the system-level properties emerge.

2. Materials and Methods

2.1 Databases and computational resources:

The first database that was created for metalloproteins is MDB (http://metallo.scripps.edu/) [35]. It was specifically geared towards metalloprotein design and consisted of many proteins with a description of the metal coordination environment. The MDB database has not been updated since 2003. MESPEUS [36] is a relatively recent database that has extensive information on the metal coordination environment. It has the geometric features of these metal sites, listed to provide more quantitative information for the sites. It was implemented in 2008. Mespeus also describes extensively geometric features of the metal coordination over any other current databases [36].

An extensive search for the metal ion based catalytic sites was performed. The Catalytic Site Atlas, which is an extensively used source for catalytic sites in proteins, predominantly focuses on the canonical amino acid residues [37]. Metal MACiE [35], Metal PDB [3] and Metal Mine [38] are three other databases with reliable metal ion based active site catalytic site information.

Apart from the databases, there are other computational resources used to analyze the proteins structurally, to identify protein function. One such key resource is Autodock Vina [39]. Autodock Vina analyses all the structures of unknown function that had good Levenshtein distances [40] and RMSD values are then screened out by comparing the Autodock Vina scores with the aid of their binding affinity scores. A binding score of greater than -5 is ignored as a poor binding even though there was a

binding involved between the ligand and the protein. The greater value suggests that the binding, though favorable wouldn't necessarily be easily recreated in a laboratory environment or in nature. Hence these results with a ($\Delta G > -5$ value have been rejected. Typically lower binding free energy values ($\Delta G < -6$ kcal/mol) were considered interesting. More details about Autodock Vina are described later in the document.

2.1.1 Metal MACiE:

MACiE (Mechanism, Annotation and Classification in Enzymes) is a database of enzyme reaction mechanisms [33]. There are 182 EC sub-classes in MACiE with approximately 90% of the structures containing a full set of 3D coordinates. MACiE also provides detailed chemical and structural information for a proposed active site.

MACiE lists the catalytic machinery of the reaction to describe the role of the different catalytic residues as well as the residues that bind to the metal ion cofactor ions. It also expands on cases where metal ions act as a core part of the mechanism. MACiE typically does not list the organic and the metal ion cofactors individually by themselves and does not list the associated canonical residues, since they are not always present in the representative crystal structure listed in the PDB entry. This may be the result of different methods being used for protein crystallization and structure determination. A comparison of the complement of the catalytic amino acid residues aids in understanding this machinery in a detailed perspective. However, the variations present in the annotations of the amino acid residues for catalytic sites leads to difficulty in ascertaining the role of each amino acid in the mechanism. This, in turn, can produce discrepancies in the results, which can sometimes be clarified by using a superimposition of the 3D coordinates of the homologous catalytic sites by IsoCleft [33].

Every individual Metal-MACiE entry has its associated sequence homologs listed, based on homolog lists found in the Catalytic Site Atlas [37]. Entries with EC number classification that are identical to the fourth EC level are listed with their CATH domains when they have at least one catalytic domain in common [37].

Only reactions with a Tanimoto [41] similarity score of higher than a set cut-off are given a higher priority by the Metal-MACiE [26]. Typically for an individual reaction fingerprint (when a single cofactor is involved), the cutoff is set for 0.75, but for the composite reaction fingerprint, when more than one cofactor is involved in the reaction, the cut-off is set at 0.65. These cut-off values are chosen in an arbitrary fashion and are listed to only to provide perspective on the similarity of the reactions catalyzed. A detailed listing of the mechanism of a proposed active site along with the amino acids and also the metal cofactors were defined in Metal-MACiE to provide an efficient supplement to the information provided in the metal listings for the database.

2.1.2 Metal PDB:

Metal PDB aims to convey the 3-dimensional information of the protein structures that contain metal ions. Metal PDB contains more than 39,000 structures that contain one or more of 56 metal ions [Metal PDB citation]. This is achieved by representing the metal binding sites in proteins and nucleic acids as Minimal Functional Sites (MFS). MFS describes the 3D template of the local environment around the metals by not taking the entire macromolecule into consideration. The primary focus would be on the local environment surrounding the metal ion. MFSs are grouped into equistructural (broadly defined as sites found in corresponding positions in similar structures) and equivalent sites (equistructural sites that contain the same metals), allowing users to easily analyze similarities and variations in metal–macromolecule interactions, and to link them to functional information [16].

It has been estimated that 30–40% of proteins require one or more metal ions to be able to carry out their biological function in cells [17, 18]. This value depends on the tissue or the organism in consideration as that effects the various metals required for the organism to function. Additionally, metal ions play a decisive role in stabilizing the structure of nucleic acids [42].

Metal ion coordination is a key feature in understanding the metal ions in macromolecular structures. The determination of this structure can help understanding the functional and biochemical aspects of the protein by understanding its interaction with the metal ions. A metal ion or cofactor along with the ligands in the metalloprotein constitutes a Minimal Functional Site (MFS) and the distance parameter is coordinated from about 5Å from the ligand. The local 3-dimensional environment is described by the MFS. The systematic structural comparison of MFSs of zinc proteins allowed a structure-based classification to be developed that is tightly connected to the functional properties of each site [43]. This example solidifies that the 3-dimensional structure can help predict the classification of a structure in the absence of proper biochemical data.

2.1.3 Metal-Mine:

Metal-Mine [38] is a database for metal-binding sites in metalloproteins [36]. The database is curated from the information of the metal-binding sites that were extracted from Protein Data Bank (PDB) structures, which are then classified based on the protein domains that contain the metal binding sites and then are manually curated by screening the data. Only the metal ions that are inherently present in the structure are considered and any tentative or artificial metal ion coordinates were excluded. They classify the metal binding sites using the Structural Classification of Proteins (SCOP) for defining the structural domains [44]. A protein can have multiple SCOP regions where the metals bind.

2.1.4 The Catalytic Site Atlas:

The Catalytic Site Atlas (CSA) [41] primarily provides curated annotations of the small number of highly conserved residues that are directly involved in undertaking the catalytic activity in enzymes, whose have a structure annotated and deposited in the Protein Data Bank (PDB) [42]. By PSIBlast method, through Homology these curated entries can be used in inferring catalytic residues in other enzyme structures. The original resource contained 177 hand-annotated entries and 2608 homologous entries, and covered 30% of all EC numbers found in PDB [41].

Primarily the significant data in the CSA consists of the catalytic sites in proteins. To be able to be designated as catalytic the residues should meet any of these mentioned criteria.

1) There must be a direct involvement in the catalytic mechanism.
2) An alteration in the pK_a of another residue or water molecule directly involved in the catalytic mechanism is performed by the residue.
3) A notable stabilization of a transition state or intermediate.
4) It plays an important role in the activation of a substrate.

The CSA does not include residues that are involved solely in ligand binding. This is a different approach than other databases like UniProtKB. Entries are made with respect to the deposited PDB structure, with the potential to have many catalytic sites within a single entry.

2.1.5 Selection and Creation of M-class Motifs in ProMOL :

There are two crucial steps in this process; the first step was to accumulate data from the above mentioned databases. The next step involved enabling ProMOL to recognize metal ions. ProMOL Version 4.0-r194 recognized only canonical amino acid residues. An additional functionality for ProMOL to

recognize these metal ions and their associated prosthetic groups was added to facilitate the generation of these motifs to determine function of proteins.

The recognition of the metal ions by ProMOL was performed by making a change in the code base of the current stable version of ProMOL. A major change in the way ProMOL analyses these atoms was done in a key method (MakeMotifCore) in the motif.py file where the definitions of each atom that ProMOL recognizes are registered. For the complex prosthetic groups that are associated with the metal ions, a different approach of individually defining each atom and then correlating those atoms to a string connotation, helps associate the ProMOL select function to the atoms in the PDB file and specified atoms in the PDB file are highlighted as the selection and then the 3-dimensional visualization of the prosthetic group occurs. Prior to visualizing the protein, two other methods (Populate and incnewcolor) were called from the Promolglobals.py file, which aids in selecting the residues and populating them on the 3-dimensional visualization template. The Metal ions are defined as "symbols" in the selection schema and the complex prosthetic groups are defined as "resn".

```
AminoLongList = ('alanine', 'arginine', 'asparagine', 'aspartate', 'cysteine',
    'glutamine', 'glutamate', 'glycine', 'histidine', 'isoleucine', 'leucine',
    'lysine', 'methionine', 'phenylalanine', 'proline', 'serine', 'threonine',
    'tryptophan', 'tyrosine', 'valine', 'calcium', 'molybdenum',
    'molybdenum4', 'magnesium', 'zinc', 'manganese', 'sodium',
    'hemes','b12','cub','fes','hea','mos','cua','fco','sf4','f3s','fe2','cfm',
    'clf','hec','cob','c2o','pcd','4mo', 'f43', '3co', 'cobalt', 'nickle', 'iron', 'copper')
AminoList = ('ala', 'arg', 'asn', 'asp', 'cys', 'gln', 'glu', 'gly', 'his',
    'ile', 'leu', 'lys', 'met', 'phe', 'pro', 'ser', 'thr', 'trp', 'tyr', 'val',
    'ca', 'mo', '4mo', 'mg', 'zn', 'mn', 'na', 'hem','b12','cub','fes','mos',
    'hea','cua','fco','sf4','f3s','fe2','cfm','clf','hec','cob','c2o','pcd','4mo','f43','3co',
    'co', 'ni', 'fe', 'cu')
AminoShortList = ('a', 'r', 'n', 'd', 'c', 'q', 'e', 'g', 'h', 'i', 'l', 'k',
    'm', 'f', 'p', 's', 't', 'w', 'y', 'v', 'ca', 'mo', '4mo',
    'mg', 'zn', 'mn', 'na', 'hem','b12','cub','fes','mos','hea','cua','fco',
    'sf4','f3s','fe2','cfm','clf','hec','cob','c2o','pcd','4mo','f43','3co', 'co', 'ni', 'fe', 'cu')
```

Figure 1: An assembly of different metal ions and prosthetic groups notations in the "incnewcolor" method in Promolglobals.py file, to help ProMOL identify the residues that need to be colored in order to visualize them in 3-dimensional renderings.

```
cmd.select('heme', 'resn hem')
cmd.select('b12', 'resn b12')
cmd.select('cub', 'resn cub')
cmd.select('fes', 'resn fes')
cmd.select('mos', 'resn mos')
cmd.select('hea', 'resn hea')
cmd.select('cua', 'resn cua')
cmd.select('fco', 'resn fco')
cmd.select('sf4', 'resn sf4')
cmd.select('f3s', 'resn f3s')
cmd.select('fe2', 'symbol fe')
cmd.select('cfm', 'resn cfm')
cmd.select('clf', 'resn clf')
cmd.select('hec', 'resn hec')
cmd.select('cob', 'resn cob')
cmd.select('c2o', 'resn c2o')
cmd.select('pcd', 'resn pcd')
cmd.select('f43', 'resn f43')
cmd.select('sodium', 'symbol na')
cmd.select('zinc', 'symbol zn')
cmd.select('3co', 'symbol co')
cmd.select('Cobalt', 'symbol co')
cmd.select('Nickle', 'symbol ni')
cmd.select('Iron', 'symbol fe')
cmd.select('Copper', 'symbol cu')
cmd.select('Manganese', 'symbol mn')
cmd.select('Magnesium', 'symbol mg')
cmd.select('4mo', 'symbol mo')
cmd.select('Molybdenum', 'symbol mo')
```

Figure 2: This figure shows the "resn" denotation for all the complex prosthetic groups containing metal ions and the simple metal ions as a "symbol" denotation in the "Populate" methods of Promolglobals.py file in aiding ProMOL understand the specific atoms to populate for the visualization to occur.

Through this work, ProMOL's dictionary has been updated to accommodate metal ions and other prosthetic groups that chelate metal ions and also act as key functional support. This functionality helps us identify the 3-dimensional positioning and the creation of these motifs. By facilitating this process we now not only can recognize the metal ions in ProMOL but also increase the probability to predict the function of metalloproteins of unknown function.

From the data curated from these different databases 103 metal ion motifs were generated. These are designated the M-Class of motif templates (M for metal).

To validate the function of these motifs, they have to pass three stages of screening to assume their performance matches our expectation in finding catalytic sites in other proteins.

1) They must demonstrate self-recognition at neutral D value (the interatomic distance between the residues).
2) They also must recognize and align with known homologs.
3) They must not recognize non-homologous structures, which are typically chosen randomly from the PDB.

Once the motifs have passed these tests, they are used to screen structures of unknown function for identical motifs.

2.2 Self- Recognition at Neutral d-value:

In the ProMOL work flow, the d value parameter in a motif file is the "distance parameter" in a motif. A motif file consists of a header and the pairwise distances of each atom in a residue to its adjacent specified atom. The header file has the information pertaining to that specific PDB ID, EC Number, residues associated in the active site with their residue numbers. The d value is associated with the pairwise distances for example:

```
cmd.select('asp101', 'n. CB&r. asp w. %s of n. MG&r. mg'%(d*10.00))
```

The "d" parameter in the above distance calculation for the protein 1bzy (Human HGPRTase with transition state inhibitor) that exhibits a functional property of a phosphoribosyl transferase is a multiplier of the distance between CB (the beta carbon) in Aspartate-101 and a nearest Magnesium in the active site, which is 10 Angstroms in this enzyme active site. The "d" parameter is a multiplier designed to enhance the ability of ProMOL to recognize homologous enzyme active sites. The default value is 2, which gives our standard search space for a structural alignment.

In the ProMOL work flow there are several stages for screening a motif. Initially, 185 metal ion motifs were generated from the data that is extracted from Metal MACiE, Catalytic Site Atlas, Metal PDB and Metal Mine. Only 103 of these motifs were successfully screened and self-identified the motifs in their native structure. This warranted further investigation to identify the potential problems and suggested solutions.

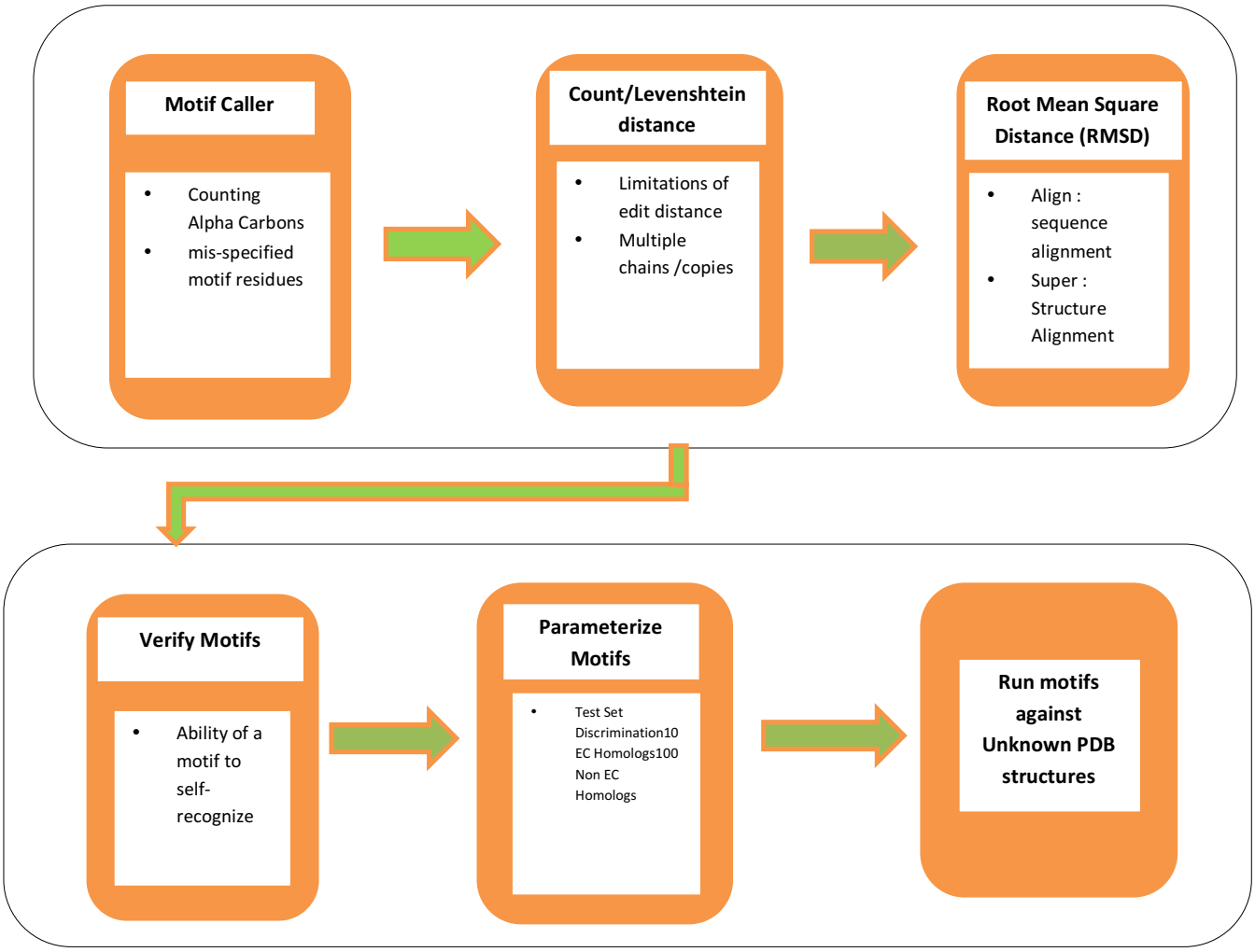


Figure 3: ProMOL workflow depicting the different phases before a motif is successfully generated. Once a motif has passed the Test Set Discrimination phase (10 EC homologs, 100 Non-EC homologs) it is then added to our library and used to test the structures of unknown function. If a motif fails such a test, it is not added in the library despite being listed in the metal active site databases.

2.2.1 Motif Caller:

The motif caller screening step validates the recognition of the residues. It calculates the number of residues present and then passes these motif parameters into the next screening tests. The primary challenge we faced, when it came to the motif caller is the inability of ProMOL to recognize metal ions and metal prosthetic groups. By increasing the dictionary capacity of ProMOL along with assigning specific atom lists to comprise the metal prosthetic groups, I was able to add this functionality to ProMOL. By changing the number of residue and the interactions that can possibly be found by Promolglobals, which is a global subroutine that validates the number of possible residues to be made, the metal ions and their associated prosthetic groups were successfully incorporated into the ProMOL residue library.

One of the potential problems that we encountered in the motif caller step was with mis-specified residues. The data for generating the metal ion motifs have been acquired from different databases. One such database is the Catalytic Site Atlas. If a residue is misrepresented as an active site in their records, ProMOL selects those motifs with the specified interatomic distances and the residue numbers. This will lead to a generation of inaccurate motifs that fail to recognize their functional homologs and sometimes even their native structure. One such documented case in my research is in 1HTO (Figure 4).

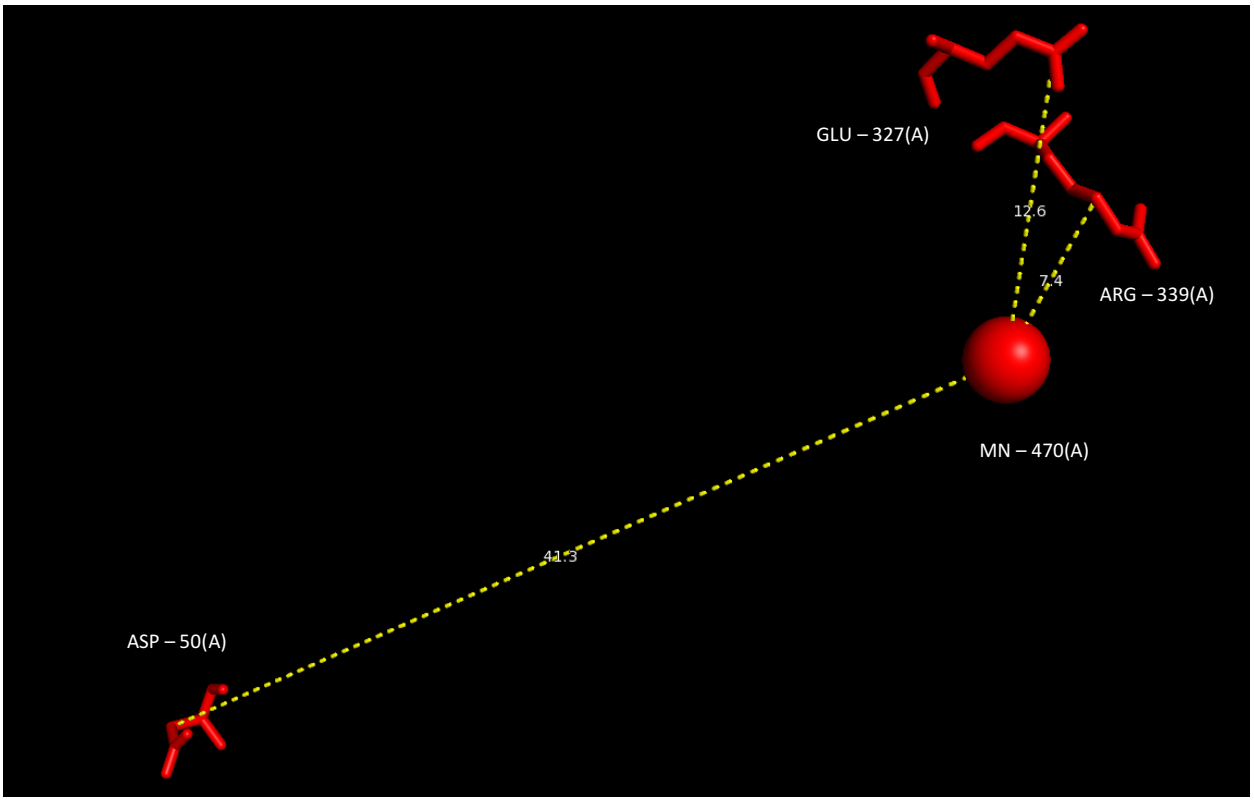


Figure 4A: A figure of 1HTO with residues specified on the wrong chain. This figure depicts the ASP- 50 on chain A, which is incorrectly annotated in the Catalytic Site Atlas. This inference is done by looking the 3-dimensional spatial alignment of the residue. The correct representation would be ASP-50 on chain B which is much closer to the density of the other active site residues.

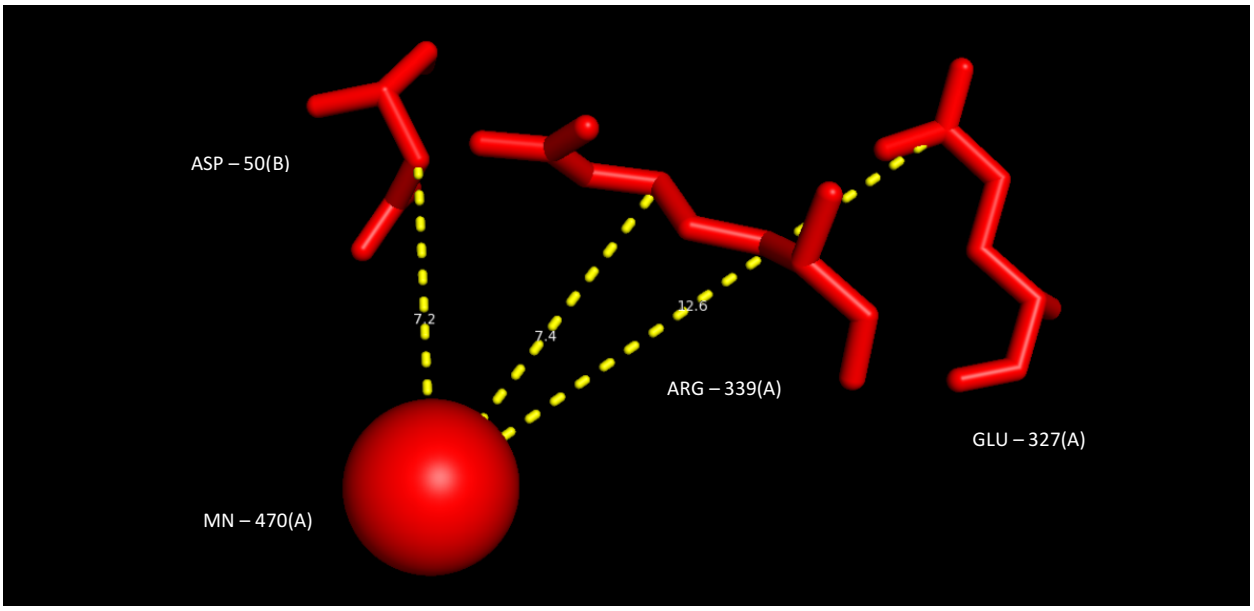


Figure 4B: This figure depicts the corrected chain positioning for the mis-specified residue. ASP- 50 on chain B is positioned at a relatively closer interatomic distance with no hydrophobic regions in the domain to prevent its interaction with other specified active sites rather than the proposed active site of ASP- 50 on chain A.

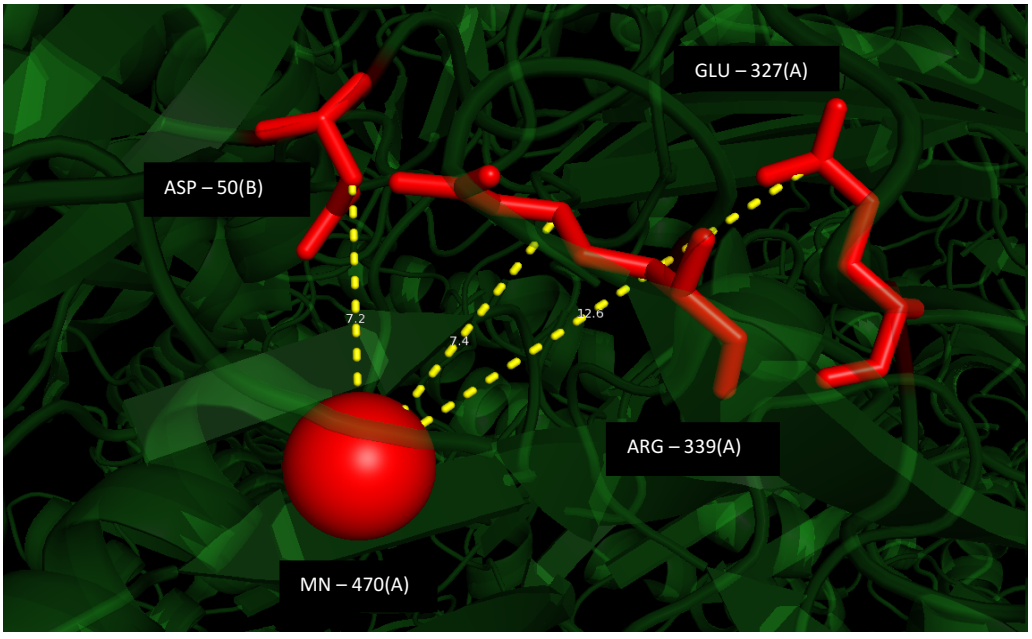


Figure 5A: A 3-dimensional placement of the corrected residues in the native structure of 1HTO.

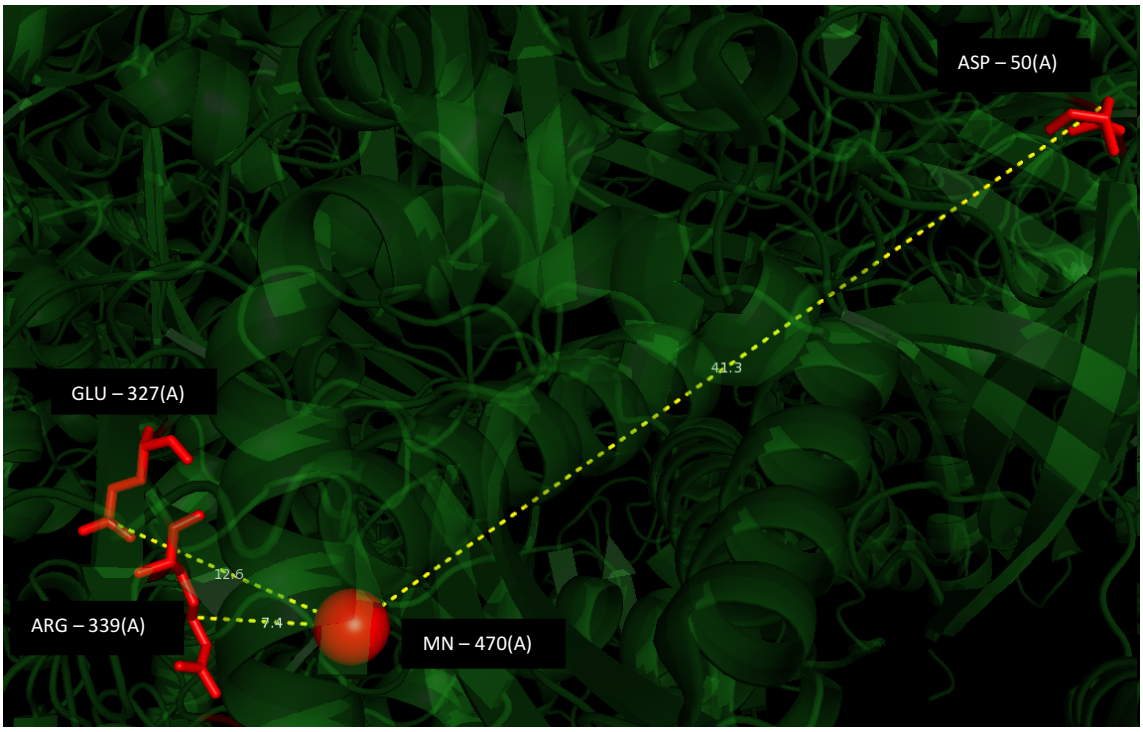


Figure 5B: A 3-dimensional visualization of the placement of the proposed active site in its native structure.

2.2.2 Count / Levenshtein distance:

Once the motif has passed through the motif caller stage, it enters a calculator for the count/Levenshtein distance [45], which is a string metric that measures the difference between two sequences. If two sequences or sets of residues are identical, the Levenshtein distance is 0. If there is one substitution (e.g., DHS ➔ EHS), then the Levenshtein distance is 1. It is also known as edit distance. When the motif passes through this phase, it counts the number of residues that pass through the list. This is helpful in the later steps of RMSD (Root Mean Square Distance) calculations. If the Levenshtein distance is greater than an acceptable range of residues of 5-7 at this point, ProMOL aborts that alignment and moves on the next structure or template. Such a large Levenshtein distance gap could occur if the template found in the query is present as multiple replicates on different chains, which

would then cause ProMOL to abort the alignment. This could lead to a false negative during the screening step. Consider the case of a template containing 4 residues. Suppose the query structure contained 8 copies of the active site for this enzyme, but each copy had a single residue difference. The Levenshtein distance would be 8 and the alignment would be aborted. To solve this issue of false negatives a higher threshold of 5-7 residues is proposed for Levenshtein distances. The new algorithm was tested with true negatives and false negatives. . The increase in the PyMOL threshold did not impact the false negative rate and true negatives stayed constant despite increasing the threshold values in the motif caller and the count subroutines.

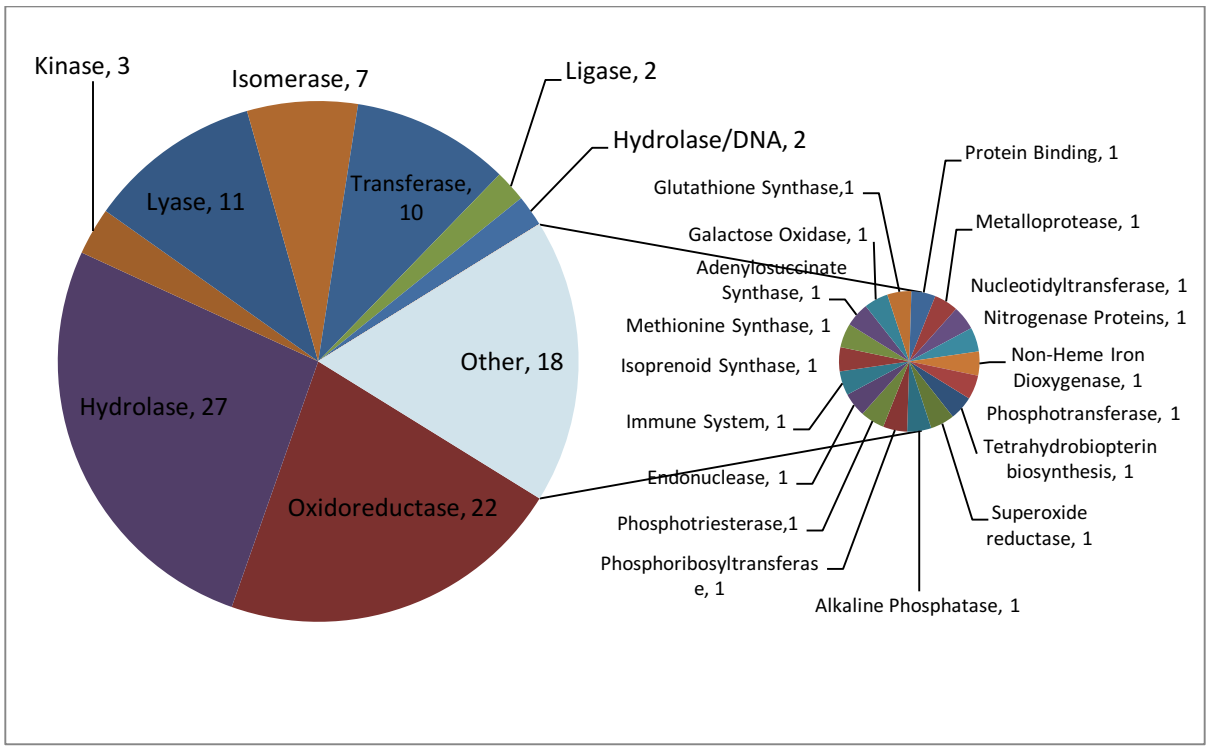


Figure 6: The distribution of EC functional classifications in the M-set.

2.3 Testing with Homologs:

To improve confidence in the 103 self-recognized motifs, they were run against their known functional homologs that were recognized as having the same functional activity. A known homolog to the generated motif is structurally similar and would be expected to have the same or a similar function. A test against 10 known true positives and 10 true negatives with 100 random PDB IDS was performed on each motif to register their performance before searching them against proteins with unknown function.

2.4 Docking Studies:

Autodock Vina:

Autodock Vina[42] is one of the computational tools used in predicting the binding affinity between macromolecule – macromolecule interactions or ligand binding to a macromolecule. Molecular docking is a computational procedure that aims to understand the binding affinity between non-covalent binding partners. The structures are primarily obtained from MD simulations or homology modeling. The binding prediction of macromolecules and small ligands plays an important role in drug screenings and further drug development. Our approach at predicting the protein function is also based on such screening methods implied in drug screenings.

Autodock Vina is used as a secondary screening of the ProMOL good hits to predict the possibility of the ligand binding to an uncharacterized protein to confer the possibility of a function assignment. This is achieved by selecting a ligand that binds near the proposed active site of the protein with the known function (motif) and then trying to dock that ligand onto the region of the protein with the unknown structure (query protein). If there is a possible binding then, Autodock vina reports a negative value for binding energy and a positive value if binding is unlikely. The more negative the binding energy, the, stronger the binding is assumed.

3. Results and Discussion

3.1 Metal ions and Prosthetic groups

Recognition of metal ions and their prosthetic groups has enabled us to expand the motif template library for ProMOL to include 103 metal ion motifs. Table 5 contains a list of the metal ions and prosthetic groups that can now be used as residues in ProMOL to create new motifs. These motifs act as resourceful tools in assigning function for metalloenzymes. Currently the metal ion motifs are rather large as they contain more residues than individual motifs in the A-set or P-set (A and P are single letter assignments for motifs that contain the 20 standard canonical residues. The A set contains motifs that have been generated automatically by ProMOL [46]. The P set (Pab, Pfa, P) consists of motifs that were generated manually with the Motif Maker in ProMOL. Pab means that the motif was created based on the alpha and beta carbon of the amino acid in the PDB template. Pfa means the motif was created by alpha carbons, beta carbons and one side chain atom found in the PDB template. A "P" designation means that the motif was created based on all the amino acid side chain atoms in the PDB template. Due to their rather large sizes, it is computationally intensive to calculate metal ion motifs in a local fashion. The memory problems with large iterative combinatorics in PyMOL are clearly prevalent when run through the M set motifs, as it takes ProMOL more than twice as long to compare query structures to M set motifs when compared to motifs from the A or P set.

	SYMBOL	NAME
1	**ZN**	**Zinc**
2	**MN**	**Manganese**
3	**NA**	**Sodium**
4	**MG**	**Magnesium**
5	**FE**	**Iron**
6	**CU**	**Copper**
7	**CO**	**Cobalt**
8	**CA**	**Calcium**
9	**MO**	**Molybdenum**
10	**NI**	**Nickel**
11	**4MO**	**Molybdenum(IV) Ion**
12	**HEM**	**Protoporphyrin IX Containing Iron**
13	**B12**	**Cobalamin**
14	**CUB**	**Cu(I)-S-Mo(IV)(=O)O-NBIC Cluster**
15	**FES**	**FE2/S2 Cluster**
16	**MOS**	**Dioxothiomolybdenum(VI) Ion**
17	**HEA**	**HEME-A**
18	**CUA**	**Dinuclear Copper Ion**
19	**FCO**	**Carbonmonoxide-(Dicyano) Iron**
20	**SF4**	**Iron/Sulphur Cluster**
21	**F3S**	**FE3-S4 Cluster**
22	**Fe2**	**FE (II) Ion**
23	**CU1**	**Copper(I) Ion**
24	**CFM**	**FE-MO-S Cluster**
25	**CLF**	**FE(8)-S(7) Cluster**
26	**HEC**	**HEME C**
27	**COB**	**Co-Methylcobalamin**
28	**C2O**	**CU-O-CU Linkage**
29	**PCD**	**(Molybdopterin-Cytosine Dinucleotide-S,S)-Dioxo-Aqua-Molybdenum(V)**
30	**F43**	**Factor 430**
31	**3Co**	**Cobalt (III) Ion**

Table 1: A list of all the additional metal ions and prosthetic groups that are recognized by PyMOL/ProMOL besides the 20 standard canonical residues.

3.2 M Set Motif Testing

Once the M set motifs were created, each motif was tested against 10 homologs (same EC#) and 10 random structures to look for true positives (homologs that were identified as homologs in ProMOL/PyMOL), false positives (random nonhomologous structures that were identified as homologs by ProMOL/PyMOL), true negatives (random nonhomologous structures that homologs that were not identified as homologs by ProMOL/PyMOL), and false negatives (homologs that were not identified as homologs by ProMOL/PyMOL). Figures 7 and 8 contain two examples of alignments between motif templates and structures of known function. As a whole, the M-set motif templates gave lower RMSD scores for homologs than for random structures (Figure 9). For homologs, the RMSD values were found mainly in the ≤1Å (78%) and 1-5Å (8%) while the unrelated structures clustered at higher RMSD values.

The 103 Metal ion motifs were run against their Structural homologs, with similar EC class designation.

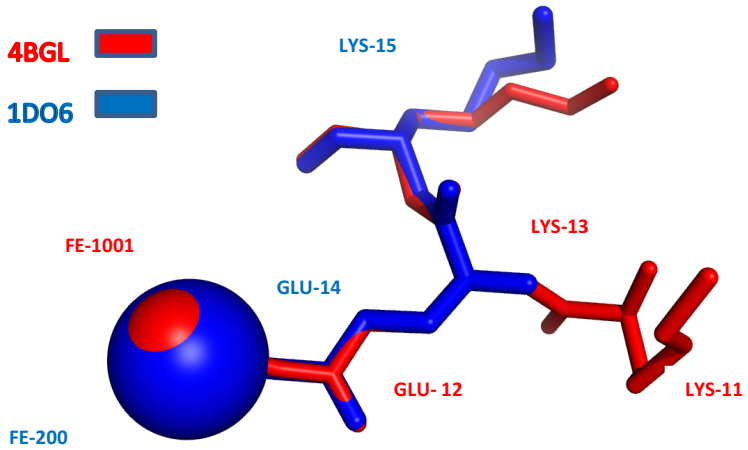


Figure 7: An overlap of the query structure 4BGL (Superoxide reductase (Neelaredoxin) from *Archaeoglobus fulgidus*) and motif template 1DO6 (Crystal Structure of Superoxide Reductase in the Oxidized state at 2.0 Angstrom Resolution). Both of the structures show the functional property of an oxidoreductase and ProMOL garnered it as a good hit. This example shows the confidence in motifs to find its functional relatives.

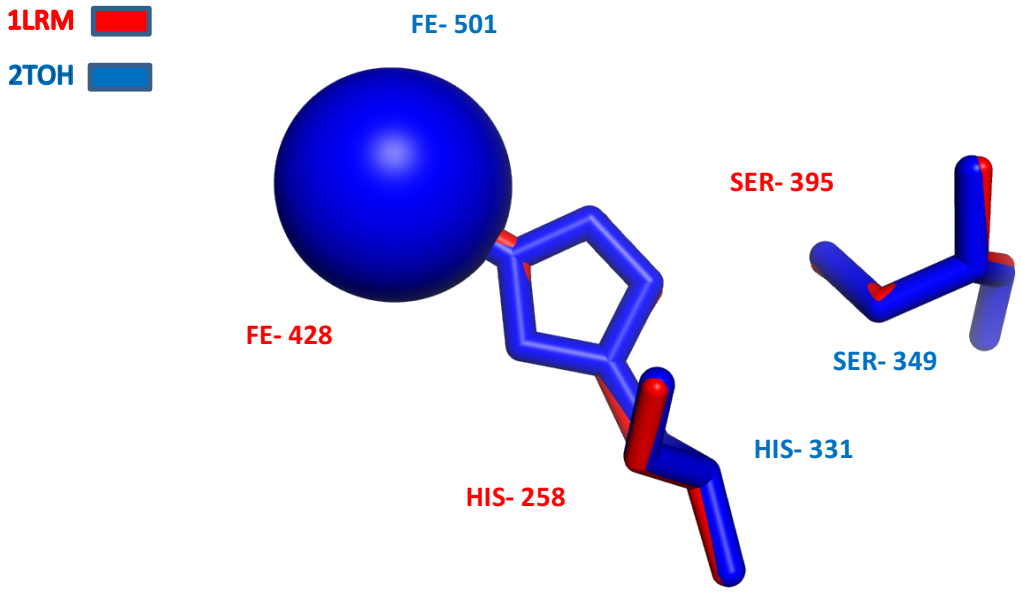


Figure 8: An overlap of the query structure 1LRM (crystal structure of binary complex of the catalytic domain of human phenylalanine hydroxylase with dihydrobiopterin (BH2)) and motif template 2TOH (tyrosine hydroxylase catalytic and tetramerization domains from rat). 1LRM is an oxidoreductase and 2TOH is functionally categorized as a hydrolase.

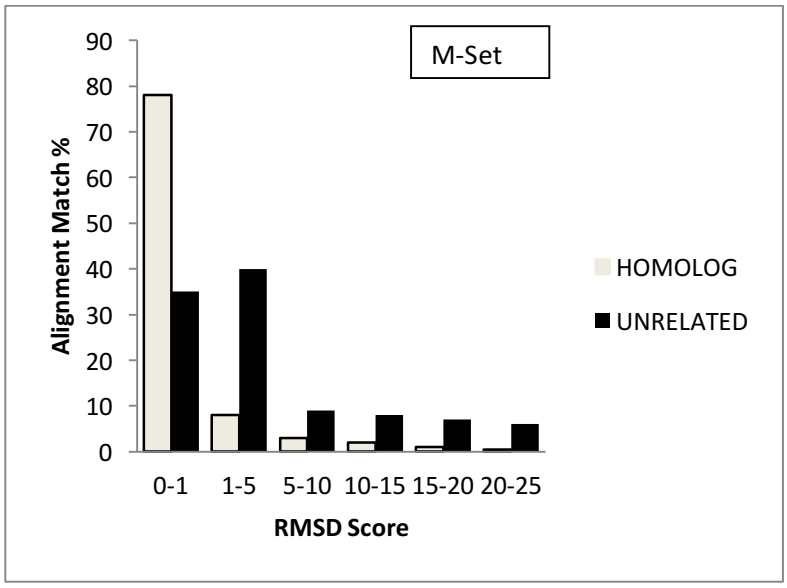


Figure 9: This bar graph depicts the performance of the RMSD values for the M-set motifs with homologs and their unrelated structures.

3.3 Screening Proteins of Unknown Function with the M Set

An exhaustive run was performed on all the metal ion structures listed in PDB, who belong to the unknown function list. Of the 3495 structures of unknown function in the PDB, there are 298 that contain metal ions. All 298 structures (Figure 10) were then queried against the 103 existing metal ion motifs to possibly assign a function to these structures. Among those hits the best hits (Table 1) were chosen based on both the visual fidelity of an overlap between the motif and the query structures along with the Levenshtein distance scores. Table 2 lists the residues found in these high quality alignments. The PDB IDs with the best alignment scores were analyzed using Autodock Vina [40] to confirm their binding affinities. In such cases, a "pdbqt" file was generated denoting the binding affinities of each ligand to the proposed macromolecule. A pdbqt file is format of a plain text readable file that Autodock uses as a proprietary output format after the calculations are performed. The Autodock Vina results are reported in Table 3, with the best hits from these studies reported in Table 4.

Unknown function structures with metal ions
1HTW,1I6N,1IN0,1IUJ,1J3W,1J9J,1JN1,1K26,1K2E,1K77,1LV3,1M65,1M68,1MOG,1MWQ,
1NC7,1NF2,1NMO,1NMP,1NNH,1NNQ,1NNW,1NO5,1NR9,1NX4,1NX8,1NZA,1NZJ,1P1M,1PB0,
1PBJ,1Q4R,1Q7H,1Q8C,1Q9U,1R4V,1R5X,1R9P,1RKQ,1RLH,1RVK,1S4C,1SAW,1SED,1SS4,
1SU0,1SU1,1T0B,1T0T,1T2B, 1T57,1T5J,1T8H,1TQX,1TT4,1TU9,1TWY,1TXL,1TXZ,1TY8,
1TZZ,1U05,1UE8,1V5N,1V70,1V8D,1VCT,1VHE,1VIZ,1VK1,1VK9,1VMF,1VMH,1VPY,1WDT,
1WFK,1WIG,1WII,1WIL,1WUF,1X4J,1XAF,1XBV,1XBX,1XBY,1XBZ,1XFI,1XJS,1XK8,1XKF,
1XM5,1XMX,1XQA,1XTL,1XTM,1XTO,1XV2,1XX7,1Y0Z,1Y63,1Y7P,1Y8A,1Y9I,1YDF,1YEY,
1YLK,1YLO,1YN4,1YWS,1YX1,1Z1S,1Z67,1Z6N,1Z84,1ZD0,1ZEL,1ZKE,1ZKP,1ZL0,1ZNO,
1ZSW,1ZUU,1ZWJ,1ZX8,1ZZM,2A33,2A5Z,2A9F,2AKL,2AMH,2ASF,2AZ4,2AZH,2B06,2B0C,
2B0V,2CQZ,2CS7,2CSY,2CUQ,2D16,2DB6,2DCT,2DDT,2DEG,2DEH,2DEV,2DJW,2DPW,2DSY,
2EZ1,2EP6,2EQ7,2EA2,2EFF,2EJQ,2EQ0,2EQ1,2EQ2,2EQ3,2EQW,2F4I,2F6S,2FBL,2FDR,
2FE1,2FIY,2FKB,2FSU,2G2C,2G2N,2G2P,2G38,2G6B,2G7Z,2G84,2GFG,2GFQ,2GGE,2GJU,
2GKP,2GL5,2GPY,2GSL,2GTA,2GU3,2GWG,2GWN,2GX8,2GYQ,2GZX,2H28,2H5N,2H6L,2HEK,
2HHG,2HIY,2HMC,2HNE,2HSI,2HSJ,2HUH,2HV6,2HXT,2HXU,2I0M,2I2O,2I3D,2I5R,2I5U,
2I71,2IAF,2IBD,2IDA,2IDL,2IEC,2IIZ,2IL5,2IMR,2JRP,2JZ8,2KGO,2KII,2KIL,2KPI,
2KW4,2M7A,2M7B,2NLY,2NN5,2NQL,2NYD,2O14,2O1E,2O1O,2O34,2O35,2O56,2O5A,2OBB,
2OD0,2OKQ,2OOI,2OPL,2OT9,2OTM,2OX6,2OY9,2OYN,2OYY,2P06,2P0N,2P0O,2P17,2P2L,
2P3P,2P4P,2P6Y,2P9X,2PJS,2PLI,2PLM,2PLS,2POD,2PS2,2PW6,2Q02,2Q3L,2Q3P,2Q40,
2Q4O,2QEN,2QGS,2QGY,2QH1,2QM2,2QMW,2QSV,2QZ7,2QZI,2R6O,2R6S,2R84,2R85,2R86,
2RA9,2RAR,2RAV,2RB5,2RBK,2RDX,2RG4,2RJB,2VH3,2W0M,2WNY,2WZ7

Figure 10: This figure lists all the Unknown Function PDB IDs with metal ions as a part of their structure.

Motif	Query	Precision Factor	Levenshtein Distance	RMSD
M_1ah7_3_1_4_3	1i6n	1	1	0.9594
M_1b66_4_6_1_10	1iuj	1	1	0.765166
M_1b66_4_6_1_10	1i6n	1	2	0.967468
M_1bg0_2_7_3_3	1j3w	1	1	0.555387
M_1rdd_3_1_26_4	1htw	0.8	1	1.250948
M_1rdd_3_1_26_4	1j3w	0.8	1	0.947101
M_1dqs_4_6_1_3	1i6n	1	1	1.150567
M_1dqs_4_6_1_3	1iuj	1	1	0.68116
M_1e7l_3_1_22_4	1i6n	1	3	3.905346
M_1eb6_3_4_24_39	1i6n	1.1	1	0.051607
M_1eb6_3_4_24_39	1iuj	1.1	1	1.480631
M_1ez1_2_1_2_-	1htw	1	1	1.281219
M_1ez2_3_1_8_1	1i6n	1	1	1.012578
M_1f48_3_6_2_16	1htw	1	2	1.871943
M_1fr2_3_1_21_1	1i6n	1	1	1.446791
M_1fr2_3_1_21_1	1iuj	1.1	1	1.117894
M_1fsg_2_4_2_8	1htw	1	1	0
M_1fsg_2_4_2_8	1j3w	1	1	0
M_1fua_4_1_2_17	1iuj	0.8	1	1.074152
M_1fua_4_1_2_17	1i6n	1	1	1.206053
M_1g4p_2_5_1_3	1htw	1	1	0.555653
M_1g4p_2_5_1_3	1j3w	1	1	0.564647
M_1ge7_3_4_24_20	1i6n	1.1	2	2.669183
M_1goj_3_6_4_4	1htw	1	1	0.1353
M_1goj_3_6_4_4	1j3w	1	1	0.585678
M_1gsa_6_3_2_3	1j3w	1.1	1	1.154524
M_1gt7_4_1_2_19	1i6n	1	2	8.948014
M_1gt7_4_1_2_19	1iuj	1	2	2.044327
M_1itq_3_4_13_19	1i6n	1	1	1.108363
M_1j09_6_1_1_17	1htw	1	1	0.937957
M_1jms_2_7_7_31	1htw	1	1	0.97168
M_1kcz_4_3_1_2	1htw	0.8	1	0.947252
M_1l0o_2_7_11_1	1j3w	1	2	1.989227
M_1l0o_2_7_11_1	1j3w	1	2	1.989227
M_1pvd_4_1_1_1	1htw	1	3	7.774868
M_1qh5_3_1_2_6	1i6n	1	1	1.307151
M_1qum_3_1_21_2	1i6n	1	1	0.800894
M_1qum_3_1_21_2	1iuj	1	1	1.393676
M_1r1j_3_4_24_11	1i6n	1	2	2.771363
M_1r44_3_4_13_22	1iuj	1	2	5.513898
M_1r44_3_4_13_22	1i6n	1	2	3.99714
M_1rdd_3_1_26_4	1htw	1	1	1.250948
M_1rdd_3_1_26_4	1j3w	1	1	0.947101
M_1xa8_4_4_1_22	1i6n	1	1	0.090155
M_1sml_3_5_2_6	1i6n	1.5	1	0.027548
M_1w0h_3_1_-_-	1j3w	1.1	2	2.825222
M_1ck7_3_4_24_24	1iuj	0.8	1	0.037948
M_1ck7_3_4_24_24	1i6n	1	1	0.039435

Table 2: Motif hits with the unknown function proteins (query)

Motif	Query	Residues
M_1ah7_3_1_4_3	1i6n	[['A', 'ASP', '174'], ['A', 'ZN', '401']]
M_1b66_4_6_1_10	1iuj	[['B', 'GLU', '60'], ['B', 'GLU', '67']]
M_1b66_4_6_1_10	1i6n	[['A', 'ASP', '45'], ['A', 'HIS', '46'], ['A', 'GLU', '142'], ['A', 'ASP', '174'], ['A', 'GLU', '246'], ['A', 'ZN', '401']]
M_1bg0_2_7_3_3	1j3w	[['D', 'ARG', '124']]
M_1rdd_3_1_26_4	1htw	[['A', 'HIS', '83'], ['A', 'MG', '561'], ['B', 'HIS', '83'], ['B', 'MG', '661'], ['C', 'HIS', '83'], ['C', 'MG', '761']]
M_1rdd_3_1_26_4	1j3w	[['D', 'HIS', '121'], ['D', 'MG', '1305']]
M_1dqs_4_6_1_3	1i6n	[['A', 'HIS', '177'], ['A', 'HIS', '200'], ['A', 'ZN', '401']]
M_1dqs_4_6_1_3	1iuj	[['A', 'HIS', '75'], ['A', 'ZN', '2007'], ['B', 'HIS', '75'], ['B', 'ZN', '2006']]
M_1e7l_3_1_22_4	1i6n	[['A', 'GLU', '32'], ['A', 'GLU', '142'], ['A', 'HIS', '177'], ['A', 'HIS', '200'], ['A', 'GLU', '246'], ['A', 'ZN', '401']]
M_1eb6_3_4_24_39	1i6n	[['A', 'TYR', '252']]
M_1eb6_3_4_24_39	1iuj	[['A', 'TYR', '50']]
M_1ez1_2_1_2_-	1htw	[['B', 'THR', '67'], ['B', 'ASP', '85'], ['B', 'ARG', '88'], ['C', 'ASP', '85']]
M_1ez2_3_1_8_1	1i6n	[['A', 'ASP', '174']]
M_1f48_3_6_2_16	1htw	[['A', 'GLY', '40'], ['A', 'GLY', '43'], ['A', 'GLY', '45'], ['A', 'LYS', '46'], ['A', 'MG', '561']]
M_1fr2_3_1_21_1	1i6n	[['A', 'HIS', '200']]
M_1fr2_3_1_21_1	1iuj	[['A', 'HIS', '75']]
M_1fsg_2_4_2_8	1htw	[['A', 'MG', '561'], ['A', 'MG', '562'], ['B', 'MG', '661'], ['C', 'MG', '761']]
M_1fsg_2_4_2_8	1j3w	[['D', 'MG', '1305']]
M_1fua_4_1_2_17	1iuj	[['B', 'GLU', '60'], ['B', 'GLU', '67'], ['B', 'ZN', '2001'], ['B', 'ZN', '2003']]
M_1fua_4_1_2_17	1i6n	[['A', 'GLU', '142'], ['A', 'GLU', '246'], ['A', 'ZN', '401']]
M_1g4p_2_5_1_3	1htw	[['A', 'ALA', '44'], ['A', 'MG', '561'], ['A', 'MG', '562'], ['B', 'ALA', '44'], ['B', 'MG', '661'], ['C', 'ALA', '44'], ['C', 'MG', '761']]
M_1g4p_2_5_1_3	1j3w	[['D', 'ALA', '125'], ['D', 'MG', '1305']]
M_1ge7_3_4_24_20	1i6n	[['A', 'GLU', '32'], ['A', 'GLU', '142'], ['A', 'TYR', '199'], ['A', 'ZN', '401']]
M_1goj_3_6_4_4	1htw	[['A', 'GLY', '40'], ['A', 'GLY', '43'], ['A', 'GLY', '45'], ['A', 'MG', '561'], ['A', 'MG', '562']]
M_1goj_3_6_4_4	1j3w	[['D', 'GLY', '122'], ['D', 'MG', '1305']]
M_1gsa_6_3_2_3	1j3w	[['D', 'ARG', '124']]
M_1gt7_4_1_2_19	1i6n	[['A', 'GLU', '7'], ['A', 'GLU', '12'], ['A', 'GLU', '24'], ['A', 'GLU', '32'], ['A', 'GLU', '41'], ['A', 'GLU', '81'], ['A', 'GLU', '85'] ['A', 'ZN', '401']]
M_1gt7_4_1_2_19	1iuj	[['B', 'GLU', '57'], ['B', 'GLU', '59'], ['B', 'GLU', '60'], ['B', 'GLU', '67'], ['B', 'GLU', '73'], ['B', 'ZN', '2001'], ['B', 'ZN', '2002']]
M_1itq_3_4_13_19	1i6n	[['A', 'ASP', '174'], ['A', 'HIS', '177'], ['A', 'HIS', '200'], ['A', 'ZN', '401']]
M_1j09_6_1_1_17	1htw	[['A', 'LYS', '46'], ['A', 'MG', '561'], ['A', 'MG', '562'], ['B', 'LYS', '46'], ['B', 'MG', '661'], ['C', 'LYS', '46'], ['C', 'MG', '761']]
M_1jms_2_7_7_31	1htw	[['B', 'ASP', '85'], ['B', 'MG', '661']]
M_1kcz_4_3_1_2	1htw	[['A', 'LYS', '46'], ['B', 'LYS', '46'], ['C', 'LYS', '46']]
M_1l0o_2_7_11_1	1j3w	[['D', 'GLU', '26'], ['D', 'ARG', '124'], ['D', 'MG', '1305']]
M_1l0o_2_7_11_1	1j3w	[['D', 'GLU', '26'], ['D', 'ARG', '124'], ['D', 'MG', '1305']]
M_1pvd_4_1_1_1	1htw	[['B', 'HIS', '29'], ['B', 'ASP', '107'], ['B', 'GLU', '113'], ['B', 'MG', '661'], ['C', 'GLU', '113'], ['C', 'MG', '761']]
M_1qh5_3_1_2_6	1i6n	[['A', 'ASP', '174'], ['A', 'ZN', '401']]
M_1qum_3_1_21_2	1i6n	[['A', 'GLU', '142'], ['A', 'GLU', '246'], ['A', 'ZN', '401']]
M_1qum_3_1_21_2	1iuj	[['B', 'GLU', '60'], ['B', 'GLU', '67'], ['B', 'ZN', '2001'], ['B', 'ZN', '2002'], ['B', 'ZN', '2003'], ['B', 'ZN', '2004'], ['B', 'ZN', '2005']]
M_1r1j_3_4_24_11	1i6n	[['A', 'GLU', '142'], ['A', 'GLU', '246']]
M_1r44_3_4_13_22	1iuj	[['B', 'GLU', '59'], ['B', 'GLU', '60'], ['B', 'ARG', '63'], ['B', 'ZN', '2003'], ['B', 'ZN', '2004']]
M_1r44_3_4_13_22	1i6n	[['A', 'GLU', '142'], ['A', 'ARG', '217'], ['A', 'GLU', '246'], ['A', 'ZN', '401']]
M_1rdd_3_1_26_4	1htw	[['A', 'HIS', '83'], ['A', 'MG', '561'], ['B', 'HIS', '83'], ['B', 'MG', '661'], ['C', 'HIS', '83'], ['C', 'MG', '761']]
M_1rdd_3_1_26_4	1j3w	[['D', 'HIS', '121'], ['D', 'MG', '1305']]
M_1xa8_4_4_1_22	1i6n	[['A', 'CYS', '4'], ['A', 'CYS', '23'], ['A', 'CYS', '93'], ['A', 'CYS', '149'], ['A', 'ZN', '401']]
M_1sml_3_5_2_6	1i6n	[['A', 'TYR', '199'], ['A', 'ZN', '401']]
M_1w0h_3_1_-_-	1j3w	[['D', 'GLU', '26'], ['D', 'HIS', '121'], ['D', 'MG', '1305']]
M_1ck7_3_4_24_24	1iuj	[['B', 'ALA', '64'], ['B', 'ZN', '2002']]
M_1ck7_3_4_24_24	1i6n	[['A', 'ALA', '8'], ['A', 'ZN', '401']]

Table 3: A list of all residue matches between associated motif and the query.

Motif	Query	Metal	Probable Function
1qlh	1iuj	ZN2+	oxidoreductase
1qlh	1i6n	ZN2+	oxidoreductase
1b66	1i6n	ZN2+	Tetrahydrobiopterin Biosynthesis
1dqs	1i6n	ZN2+	Lyase
1e7l	1i6n	ZN2+	endonuclease
1f48	1htw	MG2+	Hydrolase
1fua	1i6n	ZN2+	Lyase(aldehyde)
1ge7	1i6n	ZN2+	Hydrolase
1itq	1i6n	ZN2+	Hydrolase
1j09	1htw	MG2+	Ligase
1xa8	1i6n	ZN2+	Lyase
1kcz	1j3w	MG2+	Lyase
1ksj	1htw	MG2+	signaling protein/hydrolase
1v25	1htw	MG2+	ligase
1w0h	1htw	MG2+	hydrolase
1rdd	1htw	MG2+	hydrolase
1rdd	1j3w	MG2+	hydrolase(endoribonuclease)
1dqs	1iuj	ZN2+	Lyase
1fsg	1htw	MG2+	Transferase
1fua	1iuj	ZN2+	Lyase
1fua	1i6n	ZN2+	Lyase
1g4p	1htw	MG2+	Transferase
1g4p	1j3w	MG2+	Transferase
1goj	1j3w	MG2+	Motor Protein
1jms	1htw	MG2+	Transferase
1l0o	1j3w	MG2+	protein Binding
1pvd	1htw	MG2+	Lyase
1r44	1i6n	ZN2+	Hydrolase
1rdd	1htw	MG2+	hydrolase(endoribonuclease)
1w0h	1j3w	MG2+	hydrolase
1ck7	1iuj	ZN2+	hydrolase
1ck7	1i6n	ZN2+	hydrolase

Table 4: A list of all the structures with good ProMOL hits and a positive binding confirmation of the ligand to the unknown structure.

Assigned Function	Motif	Ligand	PF	RMSD	Query	Residues
Hydrolase	M_1w0h	ADP	1.4	1.022523	query:1htw	[['A', 'HIS', '83'], ['A', 'GLU', '113'], ['A', 'MG', '561']]
Lyase	M_1dqs	NAD	1	0.68116	query:1iuj	[['A', 'HIS', '75'], ['A', 'ZN', '2007']]
Hydrolase	M_1f48	ADP	1	1.871943	query:1htw	[['A', 'GLY', '45'], ['A', 'LYS', '46']] 'MG', '561']]
Transferase	M_1g4p	FQP	1	0.555653	query:1htw	[['A', 'ALA', '44'], ['A', 'MG', '561'], ['A', 'MG', '562']]
Transferase	M_1g4p	FQP	1	0.564647	query:1j3w	[['D', 'ALA', '125'], ['D', 'MG', '1305']]
Hydrolase	M_1w0h	AMP	1	1.989227	query:1j3w	[['D', 'GLU', '26'], ['D', 'ARG', '124'], ['D', 'MG', '1305']]

Table 5: A list of PDB IDs with good hits of proteins with unknown function that went through the high throughput screening techniques such as Autodock Vina for a more in-depth analysis.

Selected structures from the above tables were explored in more depth.

1HTW:

1HTW is a crystal structure of the YjeE protein from *Haemophilus influenzae*: a putative ATPase involved in cell wall synthesis. It contains sodium and zinc in its structure. A ProMOL run against the known metal ion motifs yielded positive hits with the structures 1FSG, 1F48, 1W0H, and 1G4P (Table 5). The functional significance of 1FSG, 1F48, 1W0H and 1G4P are transferase, hydrolase, hydrolase and transferase respectively. This is not unexpected as most metal ion motifs that contain sodium or zinc as a part of their structure tend to be either a transferase or a hydrolase. To understand the results in more detail a docking study was performed to understand the binding affinity of 1HTW with the ligands of the good hit protein structures. The results are shown below.

1HTW		
PDB ID	Enzyme Class	Function
1FSG	2.4.2.8	Transferase
1F48	3.6.1.-	Hydrolase
1W0H	3.1.-.-	Hydrolase
1G4P	2.5.1.3	Transferase

Table 6: Hits for 1HTW query protein that were acquired through structural screening.

1HTW and 1FSG:

1FSG is a transferase that belongs to the EC class 2.4.2.8 and the ligand that binds to 1FSG is 9DG (9-Deazaguanine) near its proposed active site. A pdbqt file of 9DG was generated and was docked to the query structure of 1HTW to estimate the binding affinity between the ligand and the macromolecule. If there is a significant binding between these two molecules, then we can infer from both the ProMOL data and the Autodock data that that these two proteins may be homologs with similar ancestors and similar functions.

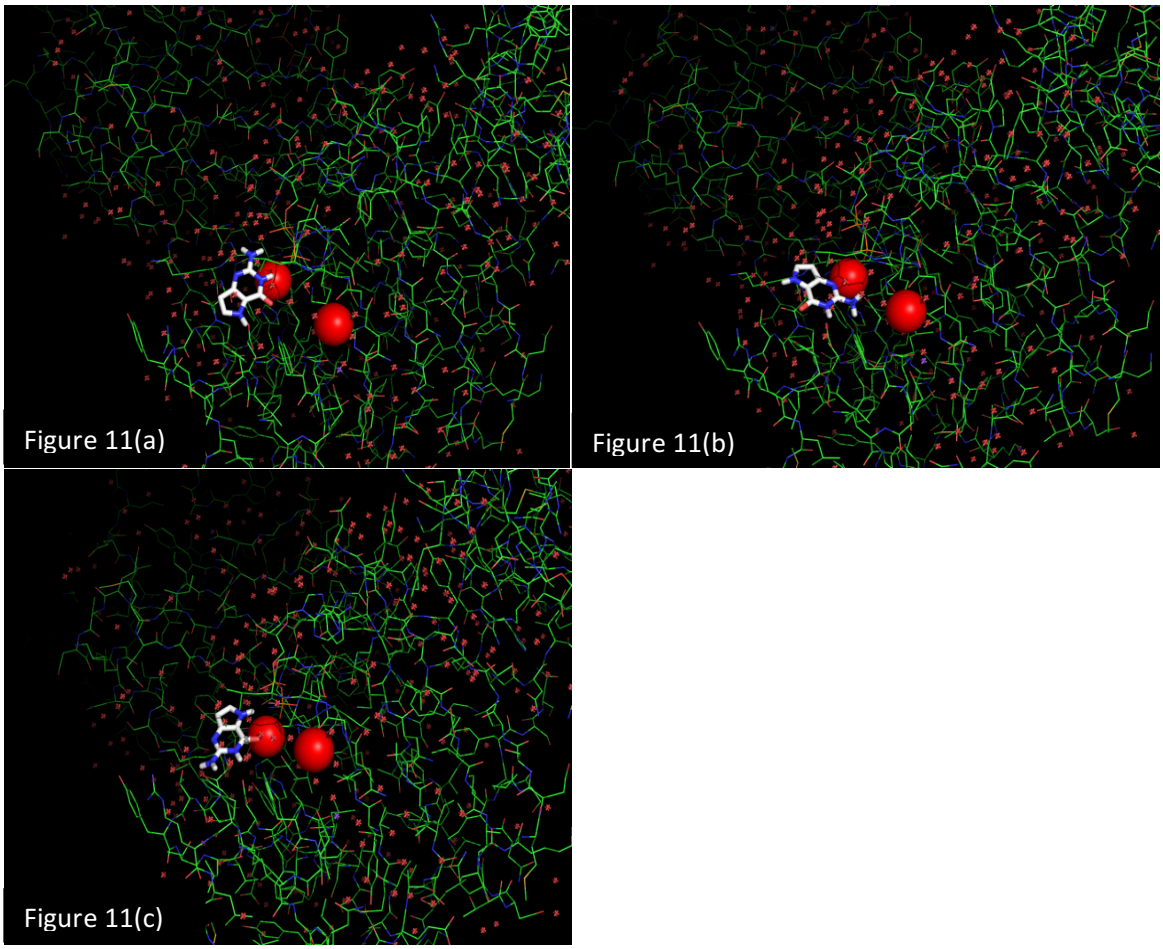


Figure 11 (a)(b)(c): This figure depicts the active site containing MG-301 and MG-302 with 9DG and the different confirmations of the ligand 9DG binding to the active site of 1HTW.

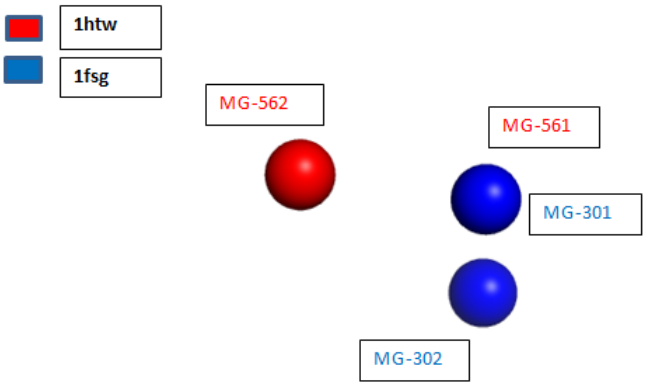


Figure 12: A ProMOL Alignment for 1HTW and 1FSG where the residue MG-561 completely overlaps with MG-301 of the motif.

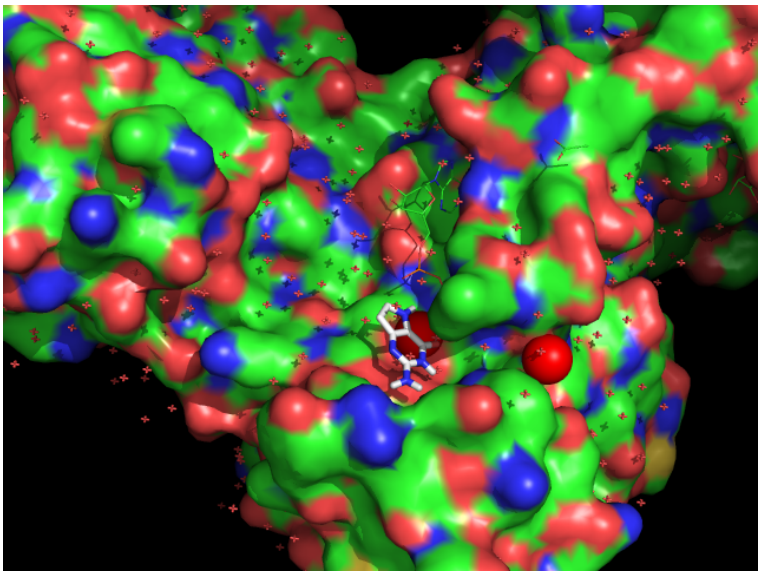


Figure 13: This figure shows confirmation 1 of ligand 9DG interacting with the active site in the binding pocket of the protein 1HTW.

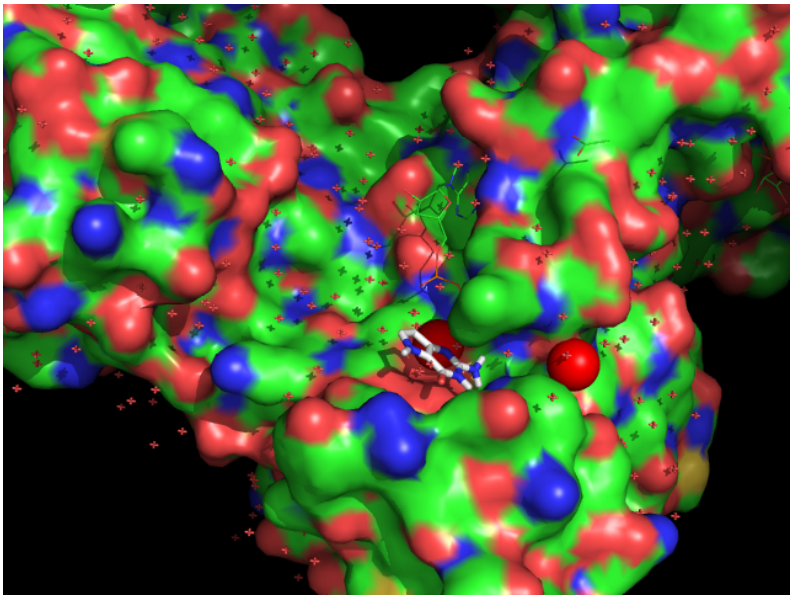


Figure 14: This figure shows confirmation 2 of ligand 9DG interacting with the active site in the binding pocket of the protein 1HTW.

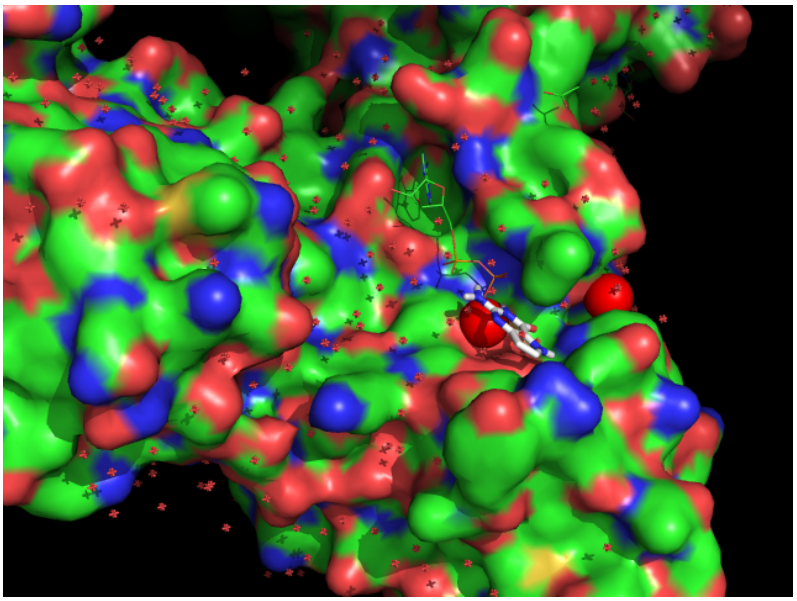


Figure 15: This figure shows confirmation 3 of ligand 9DG interacting with the active site in the binding pocket of the protein 1HTW.

Apart from pictographically conforming the binding of the protein, Autodock vina provides a log file that contains the different possible confirmations of the protein to interact with assigning a binding affinity score. For the three alignments shown above, the free energies of binding were -4.7kcal/mol , -4.8 kcal/mol and -5.1 kcal/mol repectively.

1HTW and 1F48:

1F48 is a Hydrolase that belongs to the EC class 3.6.1.- and the ligand that binds to 1F48 is ADP (ADENOSINE-5'-DIPHOSPHATE) near its proposed active site. A pdbqt file of ADP was generated and docked to the query structure of 1HTW to understand the binding affinity between the ligand and the macromolecule. If there is a significant binding between these two molecules, then we can infer from both the ProMOL data and the Autodock data that a function assignment is possible between 1HTW and 1F48 proteins.

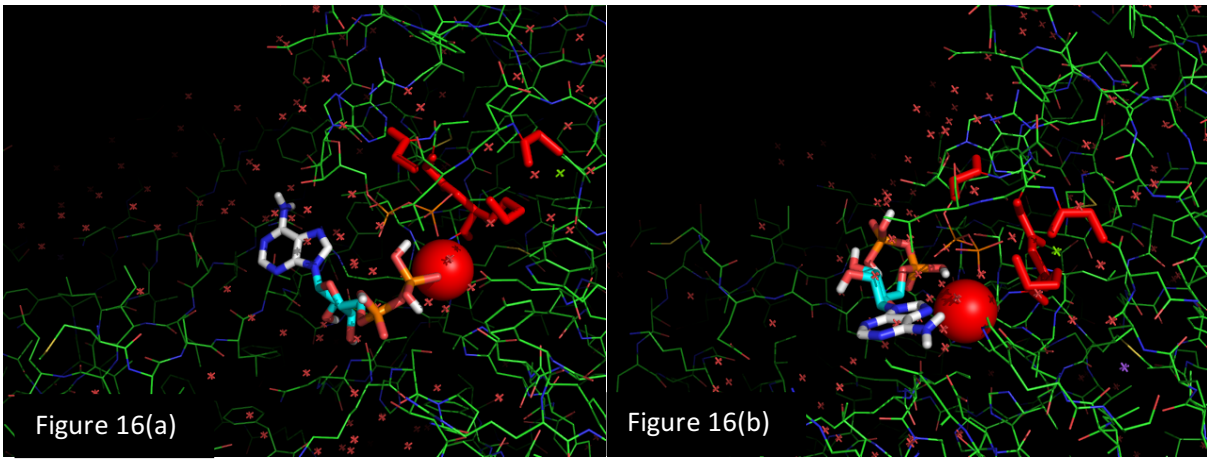


Figure 16 (a)(b): This figure depicts the different confirmations of the ligand ADP binding to the active site of 1HTW.

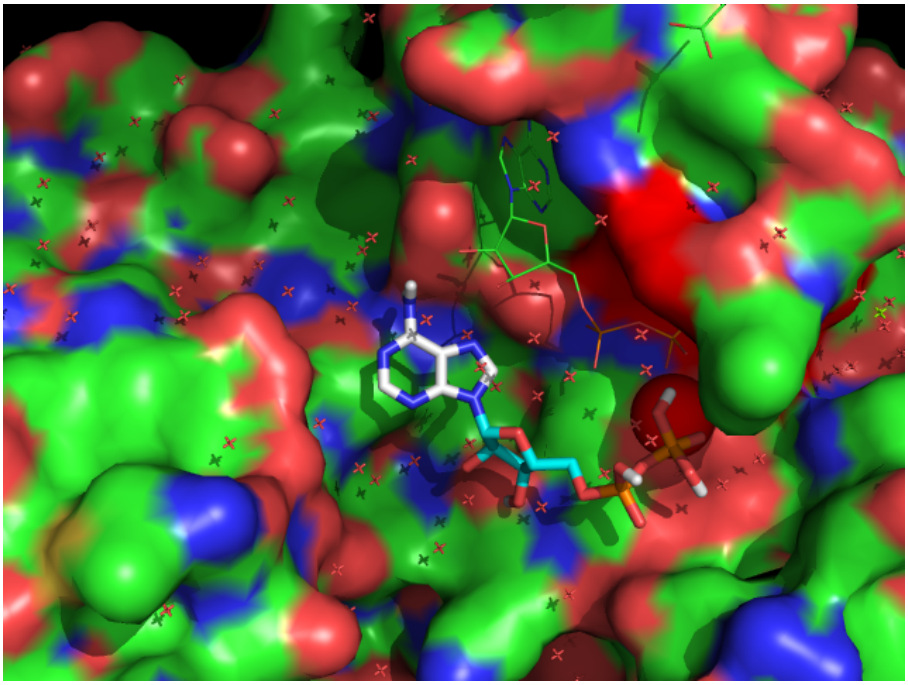


Figure 17: This figure shows a confirmation-1 of ligand ADP interacting with the active site in the binding pocket of the protein 1HTW.

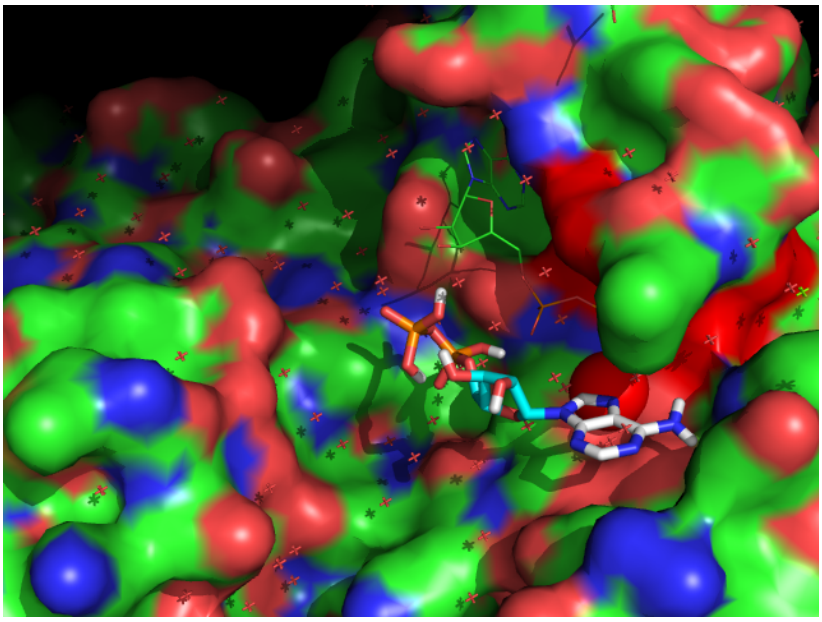


Figure 18: This figure shows a confirmation-2 of ligand ADP interacting with the active site in the binding pocket of the protein 1HTW.

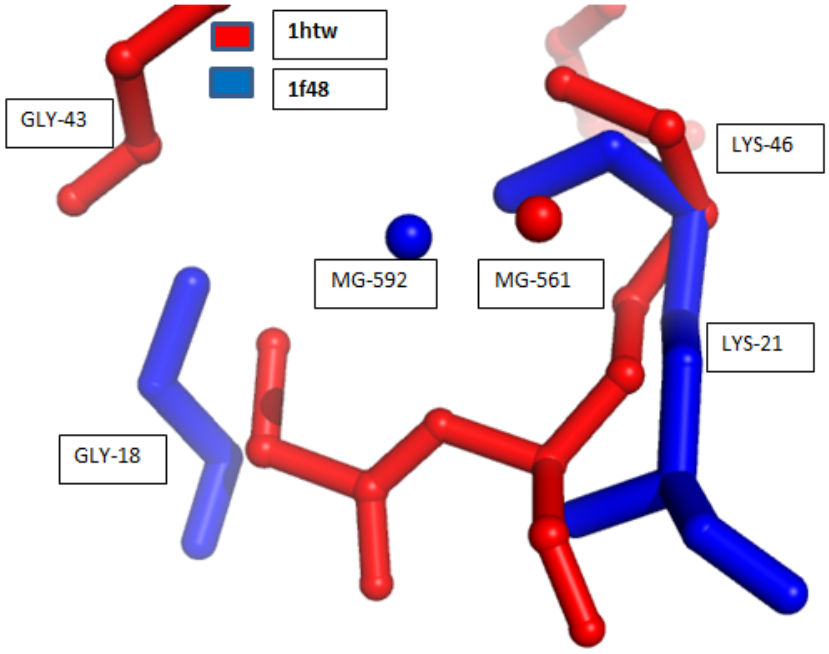


Figure 19: ProMOL Alignment of the proposed active sites of proteins 1HTW and 1F48.

The results from Autodock vina and the ProMOL alignment suggest that the query protein could be a hydrolase as the protein we are comparing against i.e. 1F48, is a hydrolase. The binding of the ligand to the proposed active region is favorable (ΔG = -6.8kcal/mol) which suggest that the active site region of the protein, potentially could exhibit the function of a hydrolase protein.

1HTW and 1W0H:

1W0H is a hydrolase that belongs to the EC class 3.1.-.- and the ligand that binds to 1W0H is AMP (Adenosine Monophosphate) near its proposed active site. A pdbqt file of AMP was generated and docked to the query structure 1HTW to understand the binding affinity between the ligand and the macromolecule. If there is a significant binding between these two molecules, then we can infer from both the ProMOL data and the Autodock data that a function assignment is possible between 1HTW and 1W0H proteins.

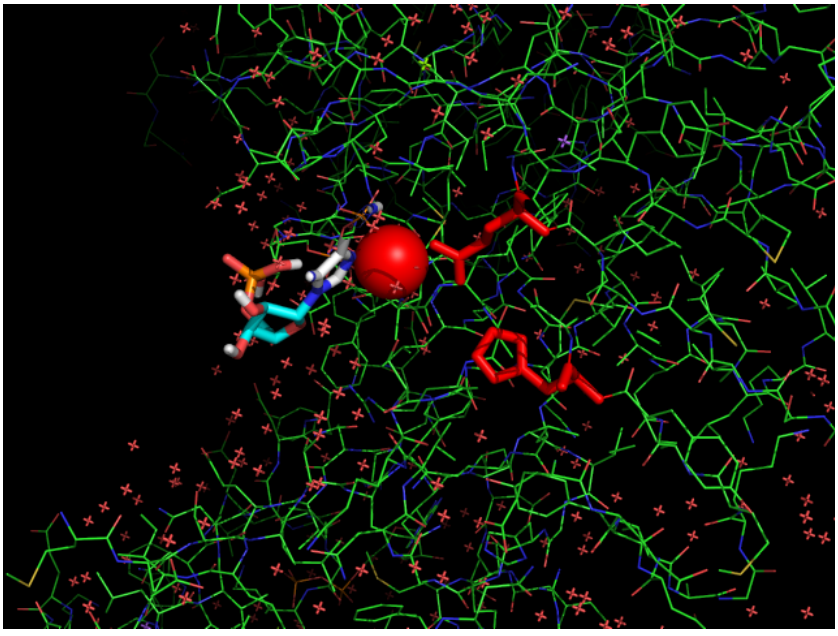


Figure 20: This figure depicts the confirmation of the ligand AMP binding to the active site of 1HTW.

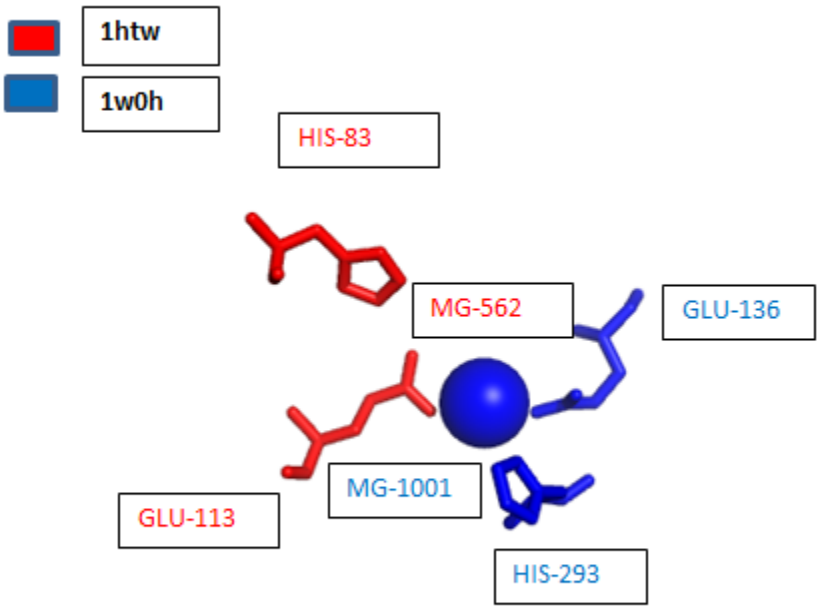


Figure 21: A ProMOL alignment of 1htw and 1w0h suggests a near perfect alignment with the metal ions MG-562 and MG-1001 with a considerable distance in the other canonical residues that are part of the active site.

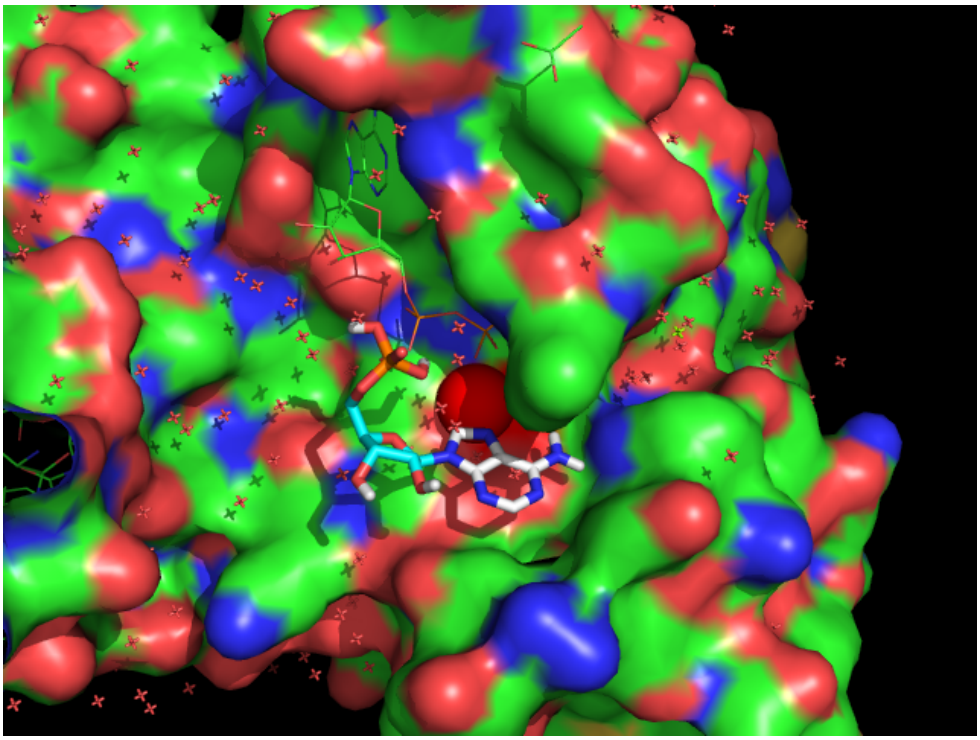


Figure 22: This figure shows a confirmation of ligand AMP interacting with the active site in the binding pocket of the protein 1HTW.

The results from Autodock vina and the ProMOL alignment suggest that the query protein could be a hydrolase. The binding of the ligand to the proposed active region is favorable (ΔG = -5.8kcal/mol) which suggest that the active site region of the protein, potentially could exhibit the function of a hydrolase protein.

1HTW and 1G4P:

1G4P is a transferase that belongs to the EC class 2.5.1.3 and the ligand that binds near the proposed active site of 1G4P is FQP (4-Amino-2-Trifluoromethyl-5-Hydroxymethylpyrimidine Pyrophosphate); it should be noted that FQP is not similar to a common metabolite. A pdbqt file of FQP was generated and docked to the query structure of 1HTW to understand the binding affinity between the ligand and the macromolecule. If there is a significant binding

between these two molecules, then we can infer from both the ProMOL data and the Autodock data that a function assignment is possible between 1HTW and 1G4P proteins.

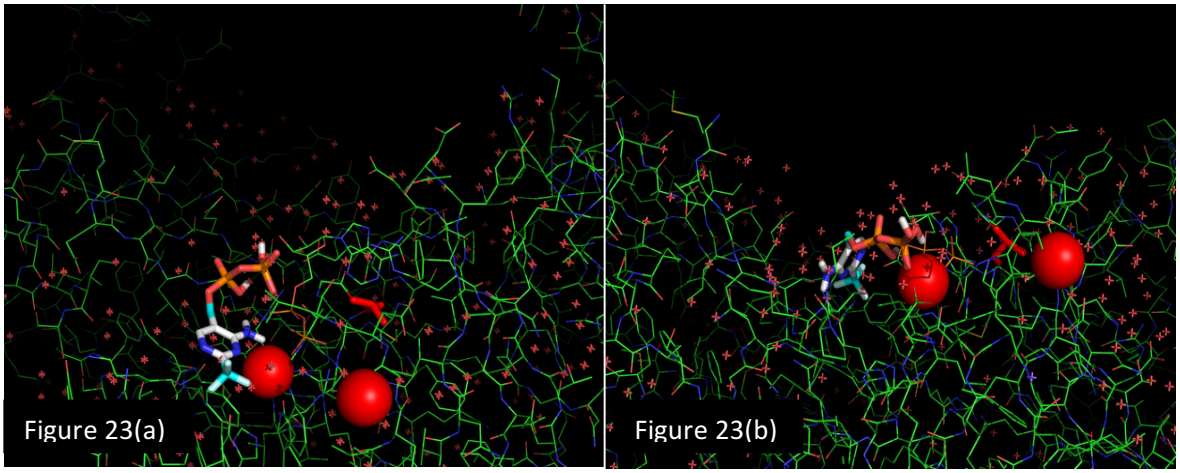


Figure 23(a)(b): This figure depicts the different confirmations of the ligand FQP binding to the active site of 1HTW.

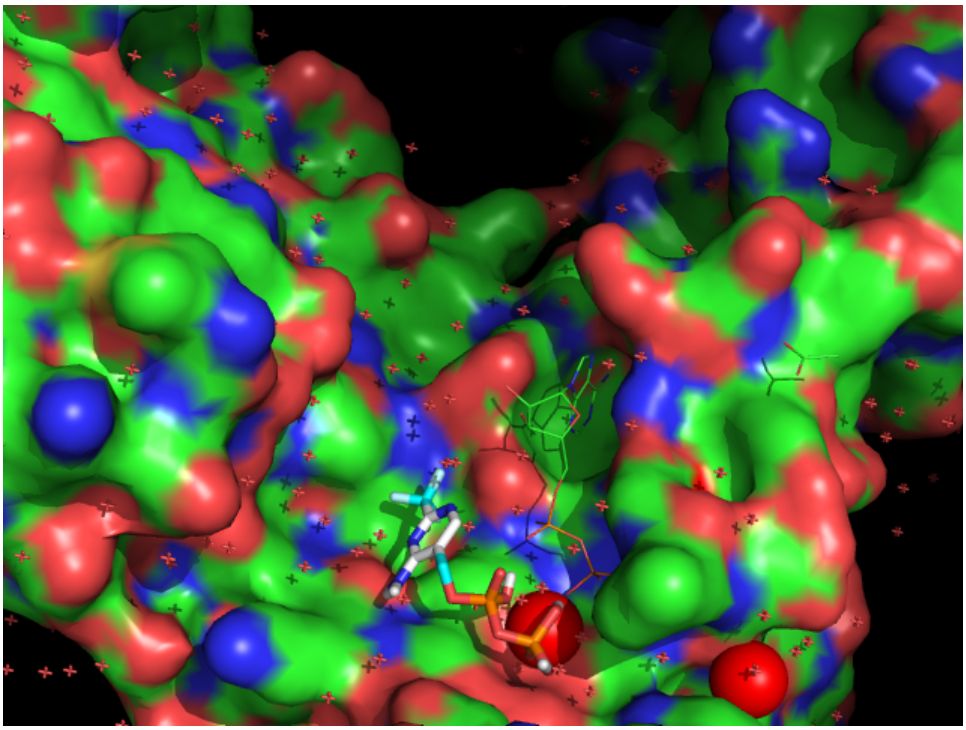


Figure 24: This figure shows confirmation-1 of ligand FQP interacting with the active site in the binding pocket of the protein 1HTW.

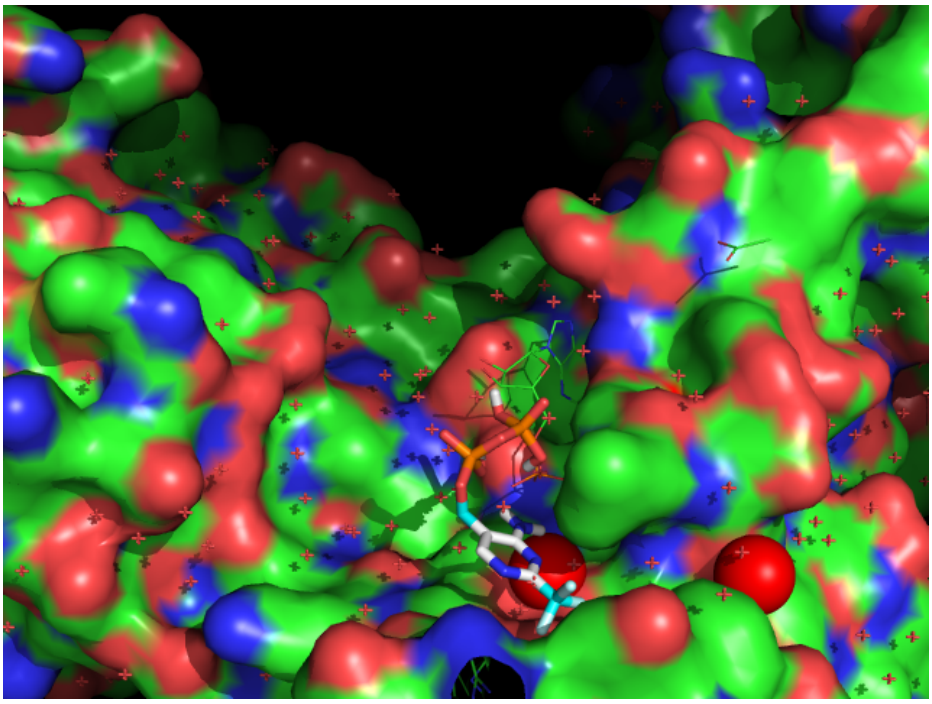


Figure 25: This figure shows confirmation-2 of ligand FQP interacting with the active site in the binding pocket of the protein 1HTW.

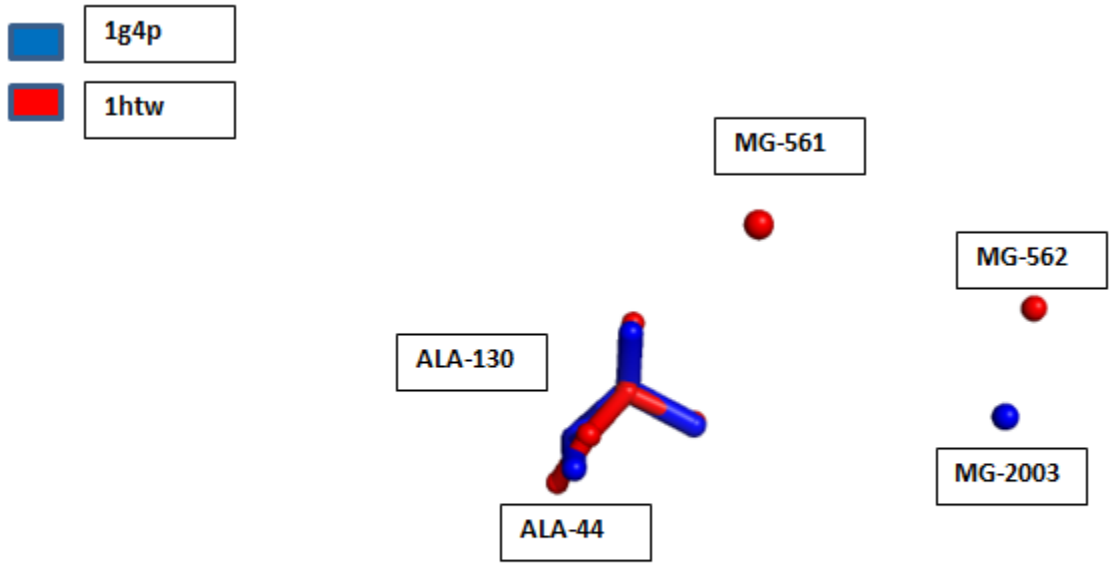


Figure 26: This Figure shows the ProMOL Alignment of the proposed active sites of proteins 1HTW and 1G4P

The Results from Autodock vina and the ProMOL alignment suggest that the query protein could be a transferase. The binding of the ligand to the proposed active region is favorable (ΔG = -5.7kcal/mol) which suggest that the active site region of the protein, potentially could exhibit the function of a transferase.

1HTW and 1FSG:

```
mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1        -5.1      0.000      0.000
   2        -5.1      8.929     10.231
   3        -4.8      8.325      9.783
   4        -4.8      2.755      3.868
   5        -4.7     48.795     49.471
   6        -4.7     15.893     16.208
   7        -4.6     45.911     46.380
   8        -4.6     23.492     24.428
   9        -4.6     17.520     17.781
Writing output ... done.
```

Figure 27: The scoring values for the binding affinity of 9DG (ligand) with 1HTW in its 9 possible confirmation states of binding.

1HTW and 1F48:

```
mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1        -6.2      0.000      0.000
   2        -5.9     23.918     26.536
   3        -5.7      3.899      5.953
   4        -5.5     24.236     26.402
   5        -5.4     19.027     21.271
   6        -5.3     20.679     22.916
   7        -5.3     13.005     15.007
   8        -5.3     35.367     36.846
   9        -5.2     28.910     30.253
Writing output ... done.
```

Figure 28: The scoring values for the binding affinity of ADP (ligand) with 1HTW in its 9 possible confirmation states of binding

1HTW and 1W0H:

mode	affinity (kcal/mol)	dist from best mode rmsd l.b.	rmsd u.b.
1	-5.8	0.000	0.000
2	-5.6	35.602	36.668
3	-5.5	4.244	7.052
4	-5.4	4.824	6.958
5	-5.3	29.542	30.316
6	-5.3	2.182	2.475
7	-5.1	22.985	25.745
8	-5.0	6.837	8.991
9	-5.0	5.955	7.555

Writing output ... done.

Figure 29: The scoring values for the binding affinity of AMP (ligand) with 1HTW in its 9 possible confirmation states of binding.

1HTW and 1G4P:

mode	affinity (kcal/mol)	dist from best mode rmsd l.b.	rmsd u.b.
1	-5.7	0.000	0.000
2	-5.6	16.698	17.431
3	-5.6	14.005	15.050
4	-5.5	14.729	15.799
5	-5.4	13.346	14.297
6	-5.4	20.562	21.524
7	-5.2	10.237	11.304
8	-5.1	10.072	10.989
9	-5.1	20.575	21.442

Writing output ... done.

Figure 30: The scoring values for the binding affinity of FQP (ligand) with 1HTW in its 9 possible confirmation states of binding.

1IUJ:

1IUJ is a crystal structure of the conserved hypothetical protein TT1380 from *Thermus*

thermophilus HB8, which contains zinc. A ProMOL run against the known metal ion motifs yielded a positive hit with the structure 1DQS, a lyase. 1DQS is a structure of dehydroquinate synthase (EC 4.2.3.4) that contains an active site capable of multistep catalysis in *Aspergillus nidulans*. To understand the results in more detail a docking study was performed to understand the binding affinity of 1IUJ with the ligands of the good hit protein structures. The results are shown below.

1IUJ and 1DQS:

1DQS is a lyase that belongs to the EC class 4.6.1.3 and the ligand that binds to 1DQS is NAD (Nicotinamide-Adenine-Dinucleotide) near its proposed active site. A pdbqt file of NAD was generated and docked to the query structure of 1IUJ to understand the binding affinity between the ligand and the macromolecule. If there is a significant binding between these two molecules, then we can infer from both the ProMOL data and the Autodock data that a function assignment is possible between 1IUJ and 1DQS proteins.

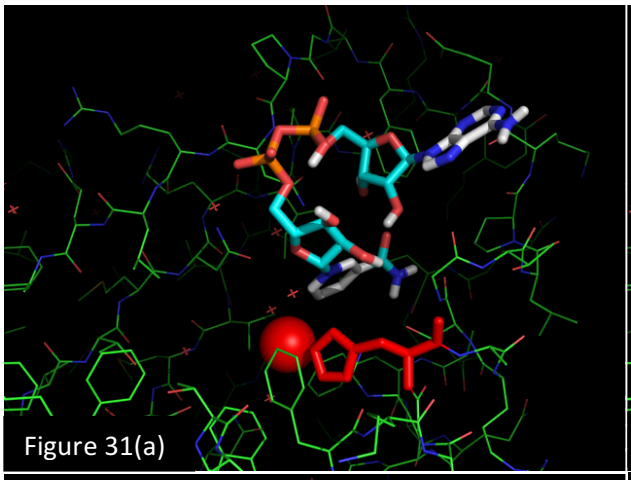


Figure 31(a)

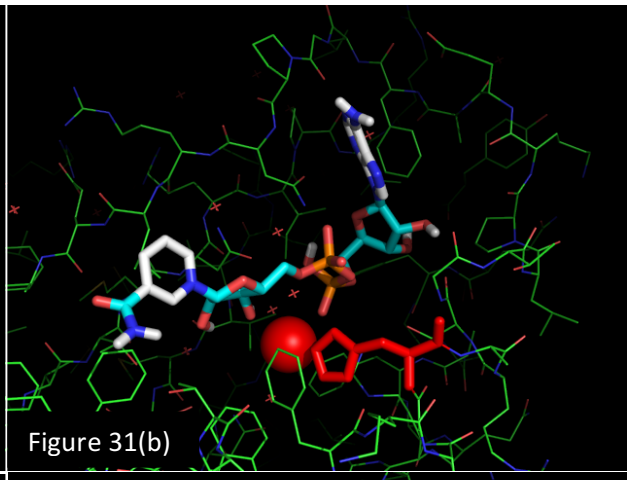


Figure 31(b)

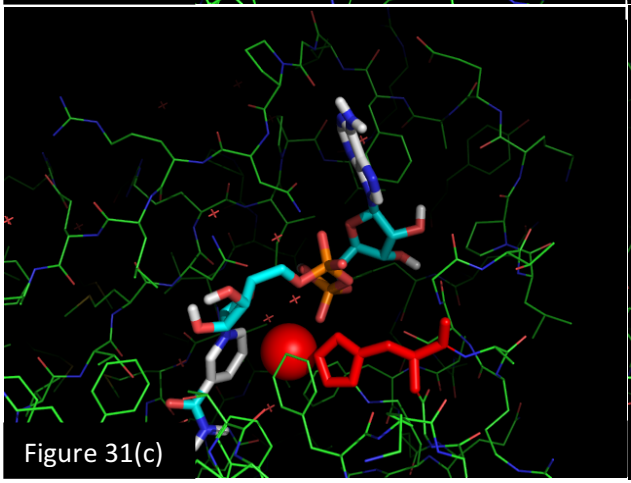


Figure 31(c)

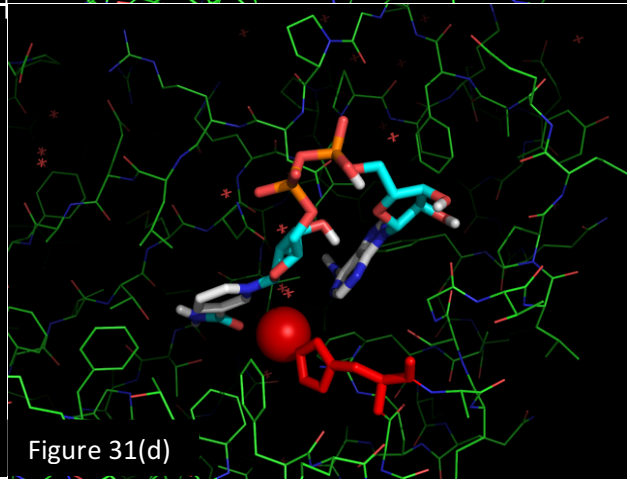


Figure 31(d)

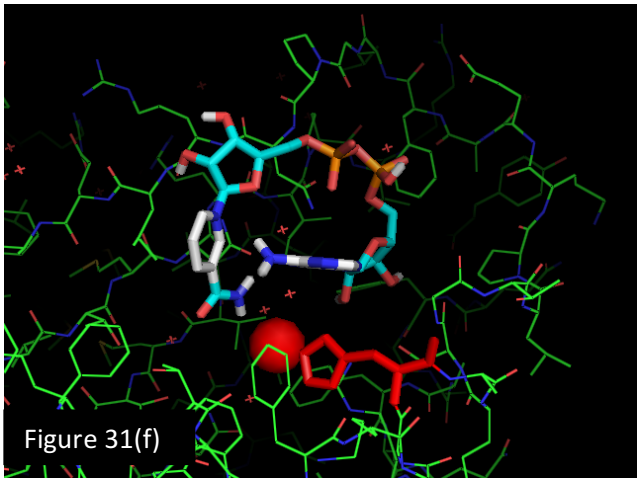
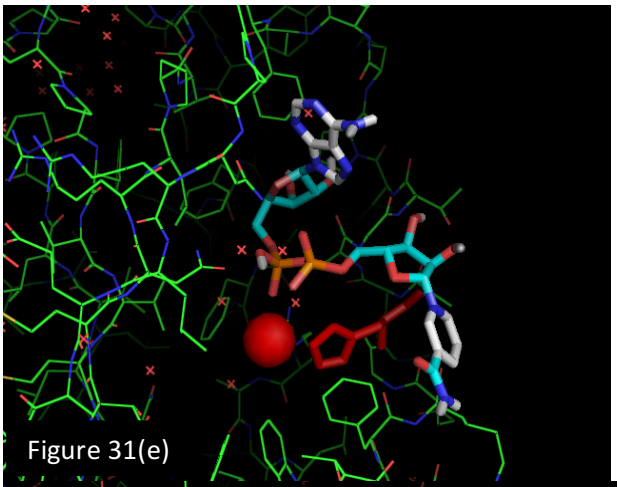
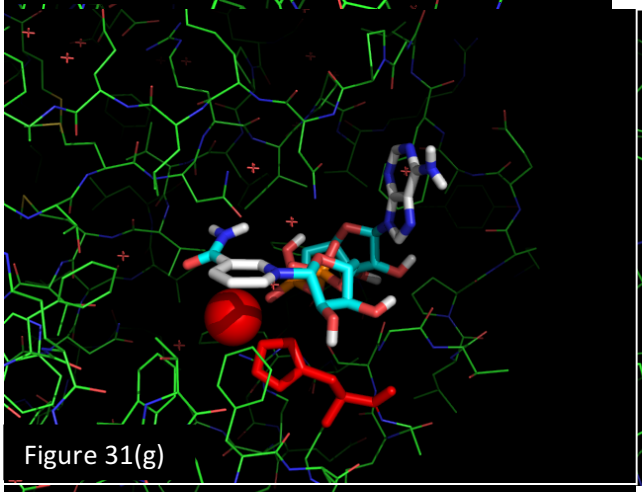
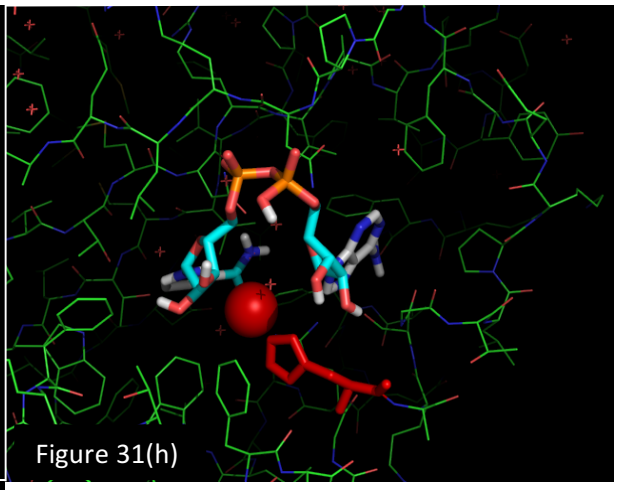


Figure 31(f)







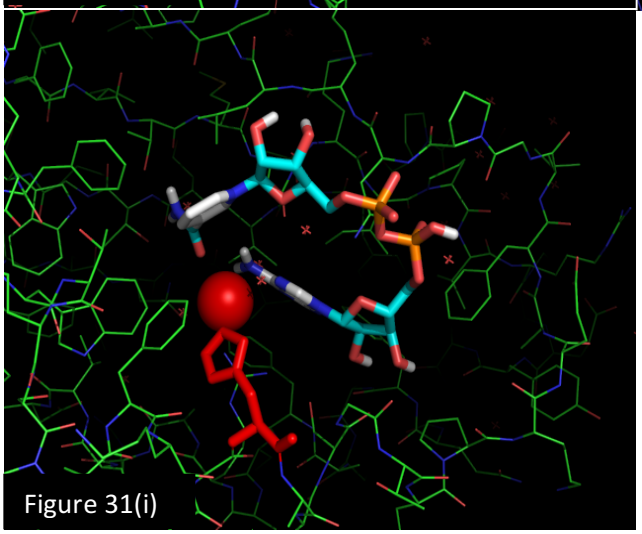


Figure 31 (a - i): This figure depicts the different confirmations of the ligand NAD binding to the active site of 1IUJ.

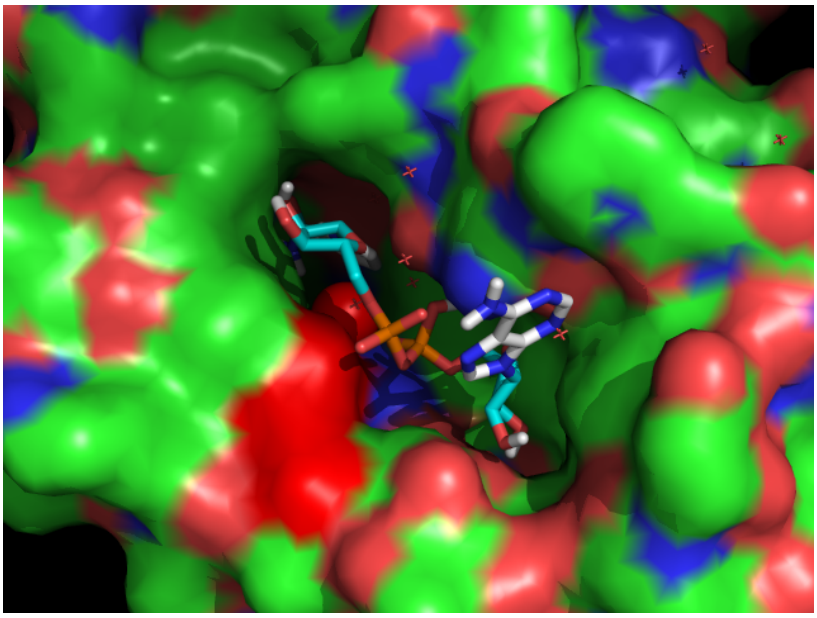


Figure 32: This figure shows a confirmation of ligand NAD interacting with the active site in the binding pocket of the protein 1IUJ.

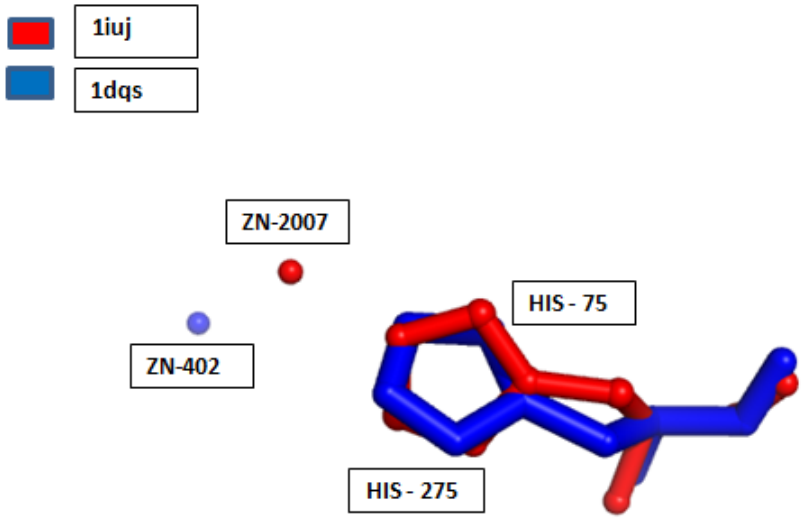


Figure 33: This Figure shows the ProMOL Alignment of the proposed active sites of proteins 1IUJ and 1DQS

1IUJ – 1DQS:

mode	affinity (kcal/mol)	dist from best mode rmsd l.b.	rmsd u.b.
1	-8.3	0.000	0.000
2	-8.2	3.382	5.615
3	-8.0	3.728	5.962
4	-8.0	4.016	7.347
5	-7.9	2.039	3.053
6	-7.9	3.784	5.794
7	-7.8	2.479	4.346
8	-7.7	2.174	3.478
9	-7.7	5.538	7.581

Writing output ... done.

Figure 34: The scoring values for the binding affinity of NAD (ligand) with 1HTW in its 9 possible confirmation states of binding.

The results from Autodock vina and the ProMOL alignment suggest that the query protein could be a lyase. Typically the function of lyases is to catalyze the breaking of various chemical bonds by means other than hydrolysis and oxidation. The binding of the ligand to the proposed active region is favorable (ΔG = -8.3kcal/mol) which suggest that the active site region of the protein, potentially be a lyase.

1J3W:

1J3W is a structure of gliding protein-mglB from *Thermus Thermophilus* HB8. It contains magnesium as a part of its structure. A ProMOL run against the known metal ion motifs yielded a positive hit with the motif templates, 1W0H and 1G4P. The functional significance of 1W0H is a hydrolase and 1G4P is a transferase. To understand the results in more detail a docking study was performed to understand the binding affinity of 1J3W with the ligands of the good hit protein structures. The results are shown below.

1J3W		
PDB ID	Enzyme Class	Function
1G4P	2.5.1.3	Transferase
1W0H	3.1.-.-	Hydrolase

Table 7: Hits for 1J3W query protein had been acquired through structural screening.

1J3W and 1G4P

1G4P is a transferase that belongs to the EC class 2.5.1.3 and the ligand that binds to 1G4P is FQP (4-Amino-2-Trifluoromethyl-5-Hydroxymethylpyrimidine Pyrophosphate) near its proposed active site; FQP is not similar to a common metabolite. A pdbqt file of FQP was generated and docked to the query structure of 1J3W to understand the binding affinity between the ligand and the macromolecule. If there is a significant binding between these two molecules, then we can infer from both the ProMOL data and the Autodock data that a function assignment is possible between 1J3W and 1G4P proteins.

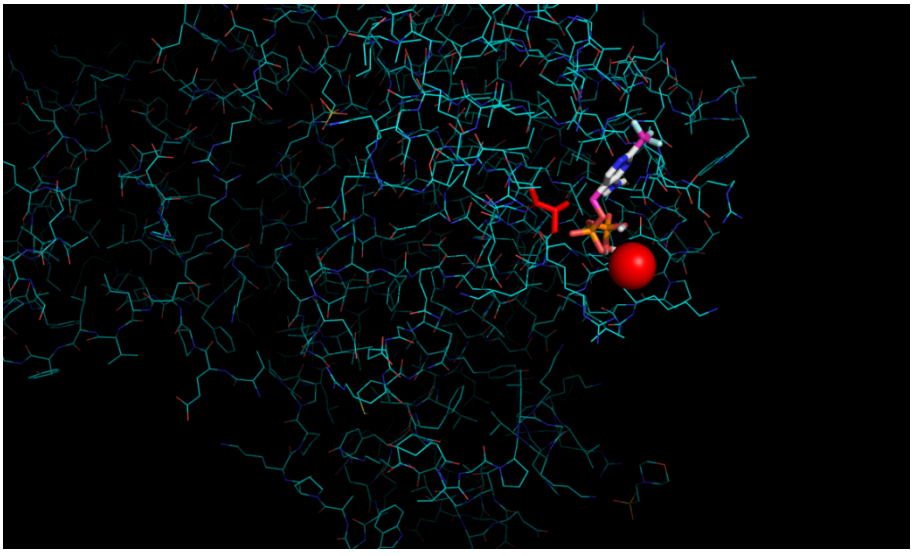


Figure 35: This figure depicts the confirmation of the ligand FQP binding to the active site of 1J3W.

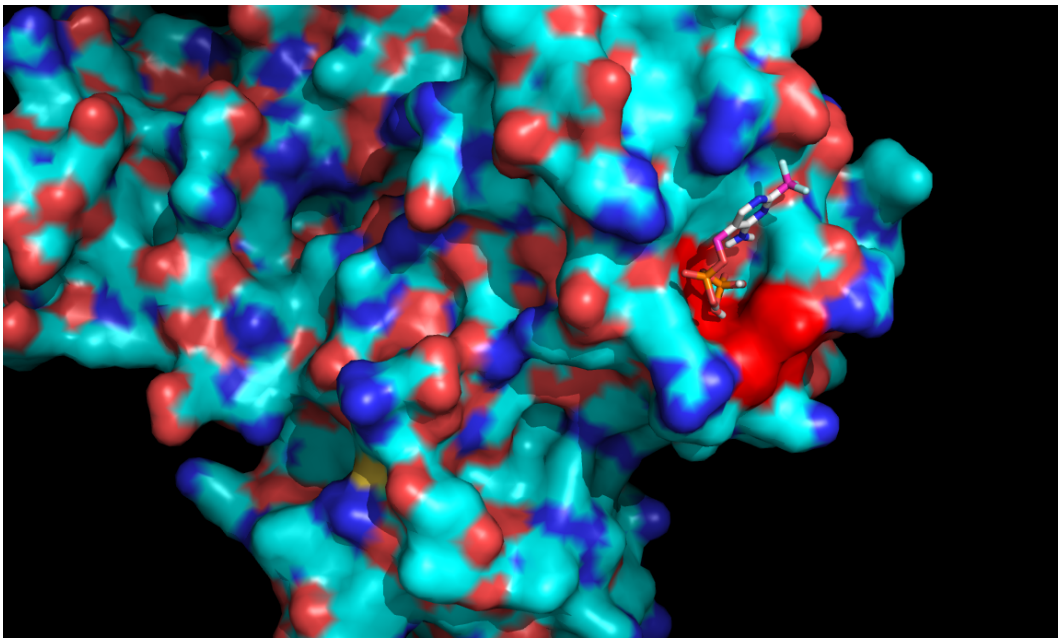


Figure 36: This figure shows a confirmation of ligand FQP interacting with the active site in the binding pocket of the protein 1J3W.

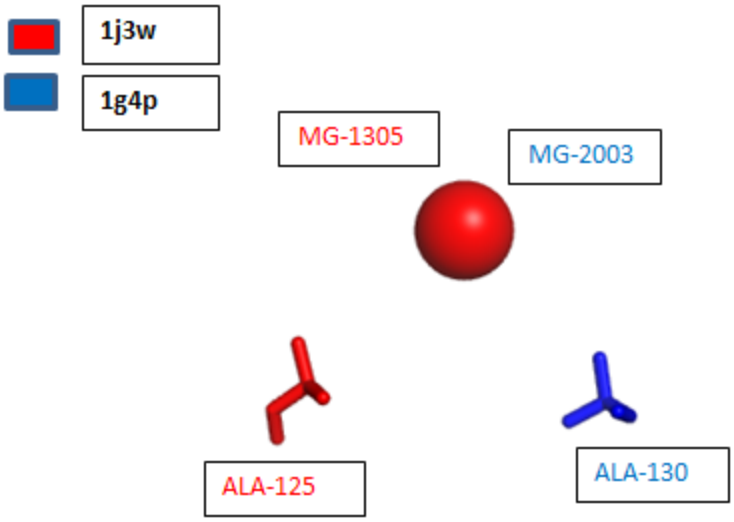


Figure 37: A ProMOL alignment of 1j3w with the active site residues MG-1305, ALA-125 and 1g4p with the active site MG- 2003, ALA -130.

The results from Autodock vina and the ProMOL alignment suggest that the query protein could be a transferase. The binding of the ligand to the proposed active region is favorable (ΔG = -6.9kcal/mol)which suggest that the active site region of the protein, potentially could exhibit the function of a transferase.

1J3W and 1W0H:

1W0H is a Hydrolase that belongs to the EC class 3.1.-.- and the ligand that binds to 1W0H is AMP (Adenosine Monophosphate) near its proposed active site. A pdbqt file of AMP was generated and docked to the query structure of 1J3W to understand the binding affinity between the ligand and the macromolecule. If there is a significant binding between these two molecules, then we can infer from both the ProMOL data and the Autodock data that a function assignment is possible between 1J3W and 1W0H proteins.

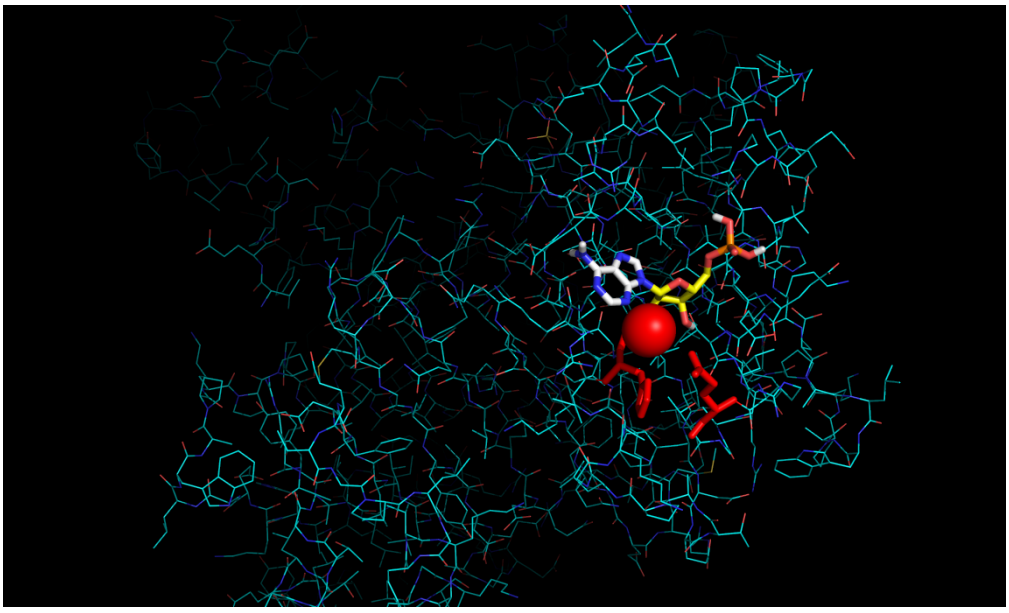


Figure 38: This figure depicts the confirmation of the ligand AMP binding to the active site of 1J3W.

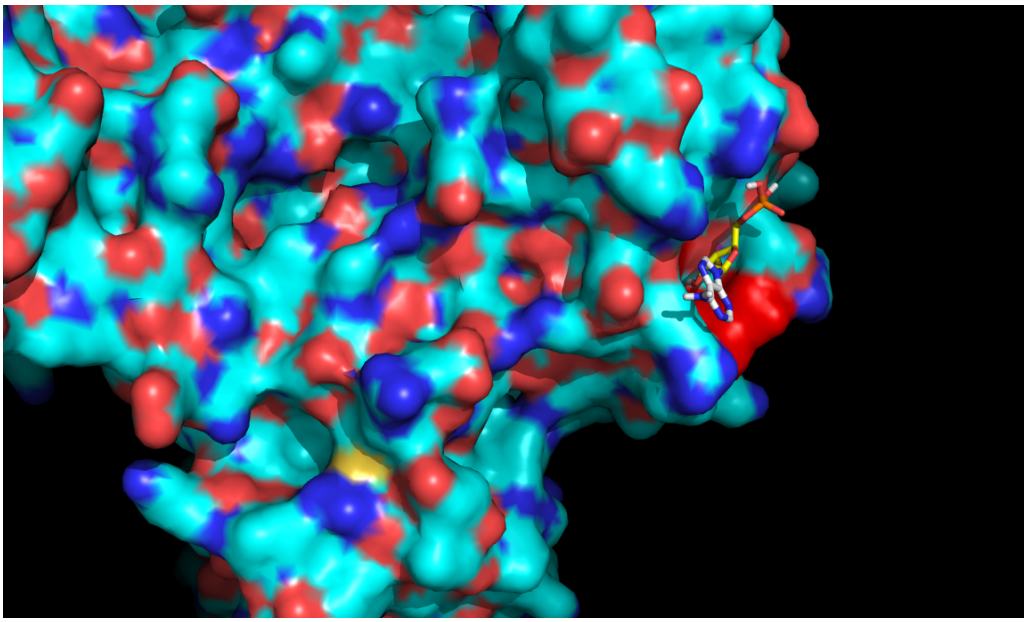


Figure 39: This figure shows a confirmation of ligand AMP interacting with the active site in the binding pocket of the protein 1J3W.

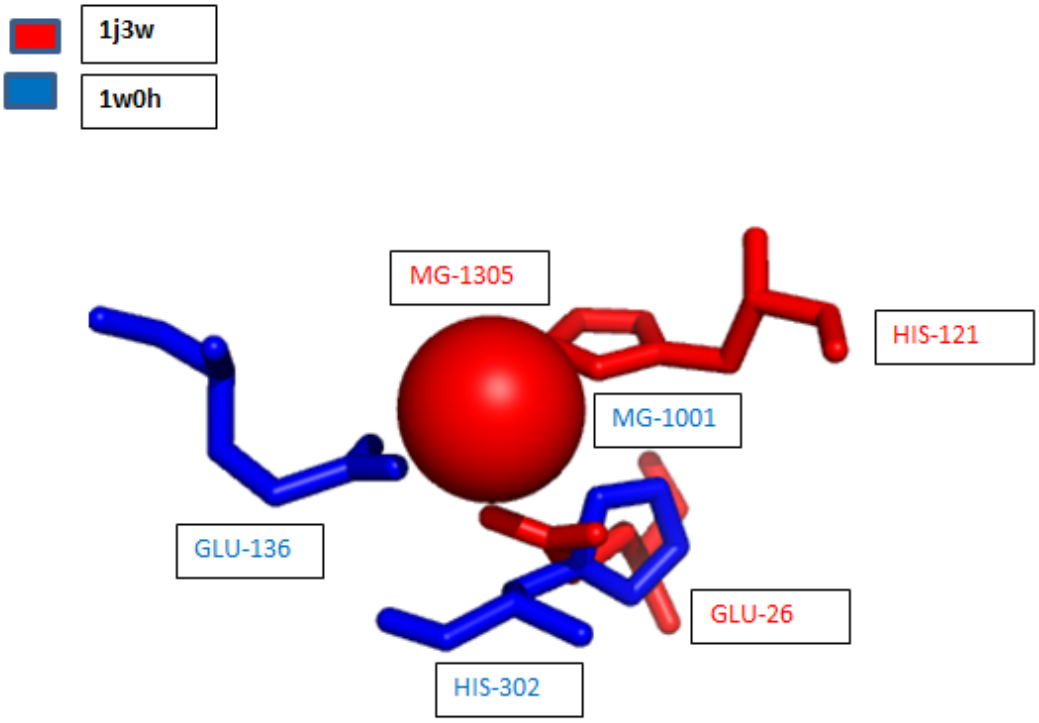


Figure 40: A ProMOL alignment of 1j3w with the active site residues MG-1305, HIS-121, GLU-26 and 1w0h with active site residues GLU-136, MG-1001, HIS-302. where the metal active site MG- 1305 overlaps with MG-1001 of the motif.

The results from Autodock vina and the ProMOL alignment suggest that the query protein could be a hydrolase. The binding of the ligand to the proposed active region is favorable which suggest that the active site region of the protein, potentially could exhibit the function of a hydrolase.

mode	affinity (kcal/mol)	dist from best mode rmsd l.b.	rmsd u.b.
1	-6.9	0.000	0.000
2	-6.8	2.173	3.217
3	-6.8	20.184	21.488
4	-6.6	1.979	2.854
5	-5.9	23.912	24.741
6	-5.9	24.597	25.727
7	-5.8	14.985	15.952
8	-5.8	20.139	21.565
9	-5.8	23.377	24.619

Writing output ... done.

Figure 41: The scoring values for the binding affinity of FQP (ligand) with 1J3W in its 9 possible confirmation states of binding.

1J3W- 1W0H:

mode	affinity (kcal/mol)	dist from best mode rmsd l.b.	rmsd u.b.
1	-6.6	0.000	0.000
2	-6.3	28.180	29.542
3	-6.3	2.296	3.669
4	-6.1	28.515	30.089
5	-6.0	25.449	27.079
6	-6.0	28.503	30.256
7	-5.9	2.264	3.398
8	-5.9	4.305	7.204
9	-5.8	42.434	44.073

Writing output ... done.

Figure 42: The scoring values for the binding affinity of AMP (ligand) with 1J3W in its 9 possible confirmation states of binding.

3.4 Analyzing Results:

ProMOL predicted functions for three PDB structures with unknown function:

- 1HTW(YjeE protein from *Haemophilus influenzae*: a putative Atpase involved in cell wall synthesis)
- 1IUJ (conserved hypothetical protein TT1380 from *Thermus thermophilus* HB8)

- 1J3W (gliding protein-mglB from *Thermus Thermophilus* HB8).

1HTW has multiple good hits with proteins 1W0H, 1F48, 1FSG and 1G4P. This suggests that the given protein could be a hydrolase or a transferase. The ligand binding affinity for all the proteins is nearly identical. The active site for 1F48 contains a GLY, LYS and MG residues while for 1G4P has ALA and two MG ions in the active site. This suggests that there are two active site regions in the protein where it is favorable for a ligand to bind and are present at a favorable binding pocket of the protein. This suggests that this protein could be a moonlighting protein. Extensive wet lab work is the only way this protein can be assigned a definitive function of either a hydrolase or a transferase. As for 1IUJ, there is a favorable hit with the protein 1DQS, which exhibits the functional property of a lyase. Both the ProMOL alignment and the Autodock results suggests that 1IUJ can likely function as a lyase with the active site residues being HIS and Zn on chain A with a favorable binding because of their positional spacing in a binding pocket region of the protein. 1J3W has a favorable alignment with 1G4P and 1W0H, which are a transferase and a hydrolase respectively, with a similar binding affinity for their respective ligands. The regions of the proposed active sites are present in a favorable binding pocket at different locations on the protein; this suggests that the protein can behave differently if interacted at different parts of both the active sites. An in-depth wet lab analysis is required to prove the functional significance of the protein. Below are the affinity values for the query and their respective hits listed from the Autodock Vina results.

3.5 Moonlighting Function:

In the process of validating the M-set motifs against random structures (which are not likely to be homologs), one structure gave an interesting alignment. 1ADN is a DNA methylphosphotrieaster repair domain protein from *E. coli*, which shows potential moonlighting activity. A moonlighting protein is a protein that exhibits a special character of performing more than one function. Through evolution from their ancestral proteins this phenomenon could be possible. The most common function of a moonlighting protein is enzymatic catalysis [45].

```
WARNING: The search space volume > 27000 Angstrom^3 (See FAQ)
Detected 4 CPUs
Reading input ... done.
Setting up the scoring function ... done.
Analyzing the binding site ... done.
Using random seed: 2141294416
Performing search ...
0%   10   20   30   40   50   60   70   80   90   100%
|----|----|----|----|----|----|----|----|----|----|
**************************************************
done.
Refining results ... done.

mode |   affinity | dist from best mode
     | (kcal/mol) | rmsd l.b.| rmsd u.b.
-----+------------+----------+----------
   1        -10.5      0.000      0.000
   2        -10.5      5.192      8.467
   3        -10.1      4.690      9.702
   4        -10.0      7.045     11.605
   5         -9.8      4.649      7.472
   6         -9.7      4.546      8.907
   7         -9.6      4.027     10.485
   8         -9.6      4.258      7.696
   9         -9.5      4.526      7.983
Writing output ... done.
```

Figure 43: Binding affinity of NAD with the macromolecule 1adn.

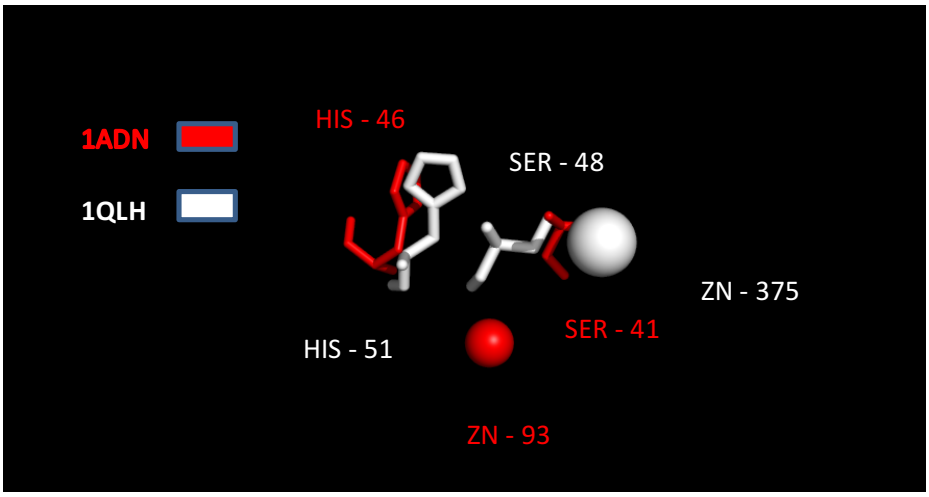


Figure 44: ProMOL alignment of the active sites from macromolecules 1ADN, 1QLH. A zinc is a common metal cofactor for these two structures.

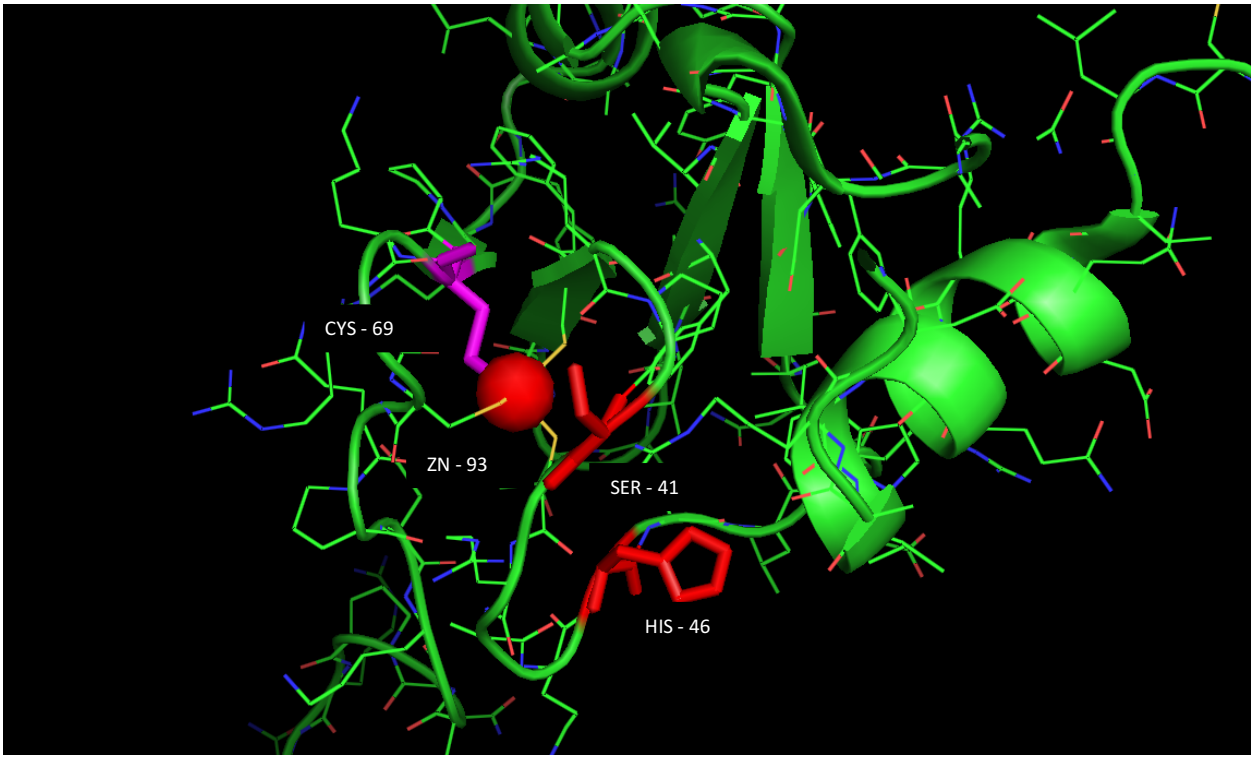


Figure 45: This figure shows the active site proposed by the catalytic site atlas in (pink) CYS -69 for the macromolecule 1adn. The residues in (red) are ZN-93 and HIS-46, these residues are identified by ProMOL as a putative hit for the protein as an active site, with zinc as a common chelator in both the reactions. The moonlighting function of 1ADN will need to be tested in the wet lab.

3.6 Novel Motif generation:

By recognizing the patterns in an existing metal ion motif we can generate a novel metal ion motif that can essentially contain the same function. For example, in PDB a protein 1GQG has a listed histidine and a copper as a registered active site on multiple active site locations on the chains A, B, C, D. The homolog of this structure 4ERO has no associated active site in the catalytic site atlas but has cobalt in the place of a copper. With relation to the atomic replacements between copper and cobalt there is a high probability that there can be a

substitution between 4ERO and 1GQG as both of them are oxidoreductases. According to their characterizations there is no listed catalytic active site for 4ERO in the catalytic active site database, but as 1GQG and 4ERO are both homologs and their residue positioning matches that of the other because of their structural, functional and sequential similarities. We can predict/ generalize that the active sites can be more coordinated than others. Taking this example into consideration a synthetic protein can be designed by building the protein around a proposed active site. This can be essentially helpful for proteins that have an ideal function but are either copyright infringed or very expensive to stabilize.

4. Conclusion and Future Plans

Metal ions are an integral part of biological systems. They provide structural and functional support to macromolecules and aid in the catalysis of their reactions. They play an active role in the mechanism of various biological pathways by assuming the role of either an inhibitor or a catalyst. Understanding the function of metals in a protein can help predict the characteristic that the protein exhibits. Around one-third of the known structures in the PDB contain one or more metal ions as a part of their structure. ProMOL utilizes PyMOL's molecular visualization and distance calculation modules to assign a 3-dimensional coordinate spatial positioning to find patterns of the same active site resurfacing in other proteins, this helps narrow down the possibilities for a potential function assignment. By adding the ability for PyMOL to recognize metal ions via ProMOL, I was able to generate metal ion catalytic active sites along with prosthetic group containing proteins as a part of the active site such as heme and cobalamin. A possible future step for the M set motifs could be the minimal set representation of the motifs

that can essentially reduce the number of pairwise distance calculations and can reduce the combinatoric calculations that occur in real time to reduce the memory issues with PyMOL.

Typically generating a motif with the canonical residues would take approximately 5-8 seconds. The A-set and the P-set motifs consists only the canonical residues. On the other hand the M-set contains metal ions and the metal based prosthetic residues which have typically a higher atom count, to make the prosthetic group. The higher the number of atoms, the greater the number of combinatorics involved in calculating the distances to generate a motif. For these reasons it is a more time-consuming task to manually query structures of unknown function against these motifs with unknown structures and hence a distributed implementation of this software is the next logical step to iteratively search through most of the structures in PDB.

Currently the M-set contains 103 Metal ion motifs curated from various databases, primarily from Metal MACiE and the Catalytic Site Atlas with reference to Metal PDB and Metal Mine. These motifs were able to self-identify, identify homologs and discriminate between homologous and non-homologous structures with known functions. By increasing the number of motifs as these databases grow, we can potentially improve function prediction of existing and future entries into the unknown function category. Increasing the selectivity and the sensitivity of the motifs, by allowing metal replacements in the current motifs we can generate novel motif active sites that can potentially find additional good hits in assigning protein function. Increasing the number of motifs that belong to the same EC class probabilistically increases the chances for an accurate protein function assignment. By adding the metal ion recognitions and metal ion motifs, we achieve better visual overlap fidelity for both the query

and the template motifs and understand the underlying biological function and the presence of metal ions in a protein. Present work has sizably increased ProMOL's motif library.

5. Bibliography

[1] Hanson, B., Westin, C., Rosa, M., Grier, A., Osipovitch, M., MacDonald, M. L., ... & Craig, P. A. (2014). Estimation of protein function using template-based alignment of enzyme active sites. BMC bioinformatics, 15(1), 87.

[2] DeLano, W. L. (2002). The PyMOL Molecular Graphics System; DeLano Scientific: Palo Alto, CA, 2002.

[3] Andreini, C., Cavallaro, G., Lorenzini, S., & Rosato, A. (2013). MetalPDB: a database of metal sites in biological macromolecular structures. Nucleic acids research, 41(D1), D312-D319.

[4] McKay T., Hart, K., Horn, A., Kessler, H., Dodge, G., Bardhi, K., Bardhi, K., Mills, J.L., Bernstein, H.J., Craig, P.A. Annotation of Proteins of Unknown Function: Initial Enzyme Results, Journal of Structural and Functional Genomics, 16:43-54, (2015) DOI: 10.1007/s10969-015-9194-5

[5] Frausto da Silva JJR, Williams RJP (2001) The Biological Chemistry of the Elements: The norganic Chemistry of Life., Second edition. Oxford University Press, Oxford

[6] Marschner H (1986) Mineral Nutrition of Higher Plants. Academic Press, San Diego

[7] Polacco JC, Holland MA (1993) Roles of urease in plant cells. In: Jeon KW, Jarvik J (eds) International Review of Cytology. Vol 145, Academic Press, San Diego, pp 65–102

[8] Rhee K-H, Morris EP, Barber J, Kuhlbrandt W (1998) Threedimensional structure of the plant photosystem II reaction center at 8 Å resolution. Nature 396: 283–286

[9] O'Halloran TV (1993) Transition metals in control of gene expression. Science 261: 715–725

[10] Ukaegbu UE, Henery S, Rosenzweig AC (2006) Biochemical characterization of MmoS, a sensor protein involved in copperdependent regulation of soluble methane monooxygenase. Biochem 45: 10191–10198

[11] Christianson DW (1991) Structural biology of zinc. Adv Protein Chem 42: 281–355

[12] Rae TD, Schmidt PJ, Pufahl RA, Culotta VC, O'Halloran TV (1999) Undetectable intraellular free copper: the requirement of a copper chaperone for superoxide dismutase. Science 284: 805–808

[13] Jernigan R, Raghunathan G, Bahor I (1994) Characterization of interactions and metal binding sites in proteins. Curr Opin Struct Biol 4: 256–263

[14] J. R Glusker. Structural aspects of metal liganding to functional groups in proteins. Adv. Protein Chem., 1991, 42, 1-73.

[15] Sigel, A., & Sigel, H. (Eds.). (1998). Metal ions in biological systems. CRC Press.

[16] Andreini,C., Bertini,I. and Rosato,A. (2009) Metalloproteomes: a bioinformatic approach. Acc. Chem. Res., 42, 1471–1479.

[17] Andreini,C., Bertini,I., Cavallaro,G., Holliday,G.L. and Thornton,J.M. (2008) Metal ions in biological catalysis: from enzyme databases to general principles. J. Biol. Inorg. Chem., 13, 1205–1218.

[18] Ludwig, M. L.; Metzger, A. L.; Pattridge, K. A.; Stallings, W. C. Manganese superoxide dismutase from Thermus thermophilus: A structural model refined at 1.8 Å resolution. J. Mol. Biol. 1991, 219, 335.

[19] Holm, R. H., Kennepohl, P., & Solomon, E. I. (1996). Structural and functional aspects of metal sites in biology. Chemical Reviews, 96(7), 2239-2314.

[20] Yamashita, M. M., Wesson, L., Eisenman, G., & Eisenberg, D. (1990). Where metal ions bind in proteins. Proceedings of the National Academy of Sciences,87(15), 5648-5652.

[21] Stojanovic, A., Stitham, J., & Hwa, J. (2004). Critical role of transmembrane segment zinc binding in the structure and function of rhodopsin. Journal of Biological Chemistry, 279(34), 35932-35941.

[22] Rao, Z., Handford, P., Mayhew, M., Knott, V., Brownlee, G. G., & StuartZ, D. (1995). The structure of a Ca 2+-binding epidermal growth factor-like domain: its role in protein-protein interactions. Cell, 82(1), 131-141.

[23] Serpersu, E. H., Shortle, D. & Mildvan, A. S. (1986) Biochemistry 25, 68-77.

[24] Cram, D. (1986) Angew. Chem. Int. Ed. Engl. 25, 1039-1057.

[25] Kretsinger, R. H., & Nelson, D. J. (1976). Calcium in biological systems.Coordination Chemistry Reviews, 18(1), 29-124.

[26] Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L., & Thornton, J. M. (2009). Metal-MACiE: a database of metals involved in biological catalysis.Bioinformatics, 25(16), 2088-2089.

[27] McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. Nucleic Acids Res., 37, D593–D597

[28] The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res., 39, D214–D219.

[29] Scheer,M., Grote,A., Chang,A., Schomburg,I., Munaretto,C.,Rother,M., So¨hngen,C., Stelzer,M., Thiele,J. and Schomburg,D.(2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res., 39, D670–D676

[30] Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res., 38, D355–D360

[31] Wittig,U., Golebiewski,M., Kania,R., Krebs,O., Mir,S., Weidemann,A., Anstein,S., Saric,J. and Rojas,I. (2006) SABIO-RK: Integration and Curation of Reaction Kinetics Data. Lect. Notes Bioinformatics, 4075, 94–103.

[32] Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. Nucleic Acids Res., 32, D434–D437

[33] Holliday,G.L., Almonacid,D.E., Bartlett,G.J., O'Boyle,N.M., Torrance,J.W., Murray-Rust,P., Mitchell,J.B.O. and Thornton,J.M. (2007) MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. Nucleic Acids Res., 35, D515–D520

[34] Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L., & Thornton, J. M. (2009). Metal-MACiE: a database of metals involved in biological catalysis.Bioinformatics, 25(16), 2088-2089.

[35] Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A., & Pique, M. E. (2002). MDB: the metalloprotein database and browser at the Scripps Research Institute. Nucleic acids research, 30(1), 379-382.

[36] Hsin,K., Sheng,Y., Harding,M.M., Taylor,P. and Walkinshaw,M.D. (2008) MESPEUS: a database of the geometry of metal sites in proteins. J. Appl. Cryst., 41, 963–968.

[37] Furnham, N., Holliday, G. L., de Beer, T. A., Jacobsen, J. O., Pearson, W. R., & Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic acids research, 42(D1), D485-D489.

[38] Nakamura, K., Hirai, A., Altaf-Ul-Amin, M., & Takahashi, H. (2009). MetalMine: a database of functional metal-binding sites in proteins. Plant biotechnology, 26(5), 517-521.

[39] Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of computational chemistry, 31(2), 455-461.

[40] Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(6), 1091-1095.

[41] Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical similarity searching. Journal of chemical information and computer sciences, 38(6), 983-996.

[42] Pechlaner,M. and Sigel,R.K.O. (2012) Characterization of metal ion-nucleic acid interactions in solution. Met. Ions Life Sci., 10, 1–42

[43] Andreini,C., Bertini,I. and Cavallaro,G. (2011) Minimal functional sites allow a classification of zinc sites in proteins. PLoS One, 10, e26325.

[44] Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540

[45] Yujian, L., & Bo, L. (2007). A normalized Levenshtein distance metric.Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(6), 1091-1095.

[46] Osipovitch, M., Lambrecht, M., Baker, C., Madha, S., Mills, J. L., Craig, P. A., & Bernstein, H. J. (2015). Automated protein motif generation in the structure-based protein function prediction tool ProMOL. Journal of structural and functional genomics, 1-11.

[47] Jeffery, C. J. (1999). Moonlighting proteins. Trends in biochemical sciences,24(1), 8-11.

6. APPENDIX

PDBID	Protein Name
1afr	Stearoyl-Acyl Carrier Protein Desaturase From Castor Seeds
1alk	Reaction Mechanism Of Alkaline Phosphatase Based On Crystal Structures.
1aq2	Phosphoenolpyruvate Carboxykinase
1b57	Class Ii Fructose-1,6-Bisphosphate Aldolase In Complex With Phosphoglycolohydroxamate
1b66	6-Pyruvoyl Tetrahydropterin Synthase
1bg0	Transition State Structure Of Arginine Kinase
1bzy	Human Hgprtase With Transition State Inhibitor
1c9u	Crystal Structure Of The Soluble Quinoprotein Glucose Dehydrogenase In Complex With Pqq
1ctt	Transition-State Selectivity For A Single Oh Group During Catalysis By Cytidine Deaminase
1d8c	Malate Synthase G Complexed With Magnesium And Glyoxylate
1do6	Crystal Structure Of Superoxide Reductase In The Oxidized State At 2.0 Angstrom Resolution
1e1a	Crystal Structure Of Dfpase From Loligo Vulgaris
1e7l	Endonuclease Vii (Endovii) N62d Mutant From Phage T4
1ez2	Three-Dimensional Structure Of The Zinc-Containing Phosphotriesterase With Bound Substrate Analog Diisopropylmethyl Phosphonate.
1fr2	Crystal Structure Of The E9 Dnase Domain With A Mutant Immunity Protein Im9(E41a)
1fua	L-Fuculose 1-Phosphate Aldolase Crystal Form T
1fui	L-Fucose Isomerase From Escherichia Coli
1g4p	Thiamin Phosphate Synthase
1g72	Catalytic Mechanism Of Quinoprotein Methanol Dehydrogenase
5eat	5-Epi-Aristolochene Synthase From Nicotiana Tabacum With Substrate Analog Farnesyl Hydroxyphosphonate
13pk	Ternary Complex Of Phosphoglycerate Kinase From Trypanosoma Brucei
1bmt	X-Ray Structure Of The B12-Binding Domains Of Methionine Synthase
1ge7	Zinc Peptidase From Grifola Frondosa
1gim	Crystal Structure Of Adenylosuccinate Synthetase From Escherichia Coli
1gog	Novel Thioether Bond Revealed By A 1.7 Angstroms Crystal Structure Of Galactose Oxidase
1gp5	Anthocyanidin Synthase From Arabidopsis Thaliana Complexed With Trans-Dihydroquercetin
1gqg	Quercetin 2,3-Dioxygenase In Complex With The Inhibitor Diethyldithiocarbamate
1gsa	Structure Of Glutathione Synthetase Complexed With Adp And Glutathione
1h19	Structure Of [E271q] Leukotriene A4 Hydrolase
1hdh	Arylsulfatase From Pseudomonas Aeruginosa
1i1i	Neurolysin (Endopeptidase 24.16) Crystal Structure
1i6p	Crystal Structure Of E. Coli Beta Carbonic Anhydrase (Ecca)
1ir3	Phosphorylated Insulin Receptor Tyrosine Kinase In Complex With Peptide Substrate And Atp Analog
1it4	Solution Structure Of The Prokaryotic Phospholipase A2 From Streptomyces Violaceoruber
1j53	Structure Of The N-Terminal Exonuclease Domain Of The Epsilon Subunit Of E.Coli Dna Polymerase Iii At Ph 8.5

1jms	Crystal Structure Of The Catalytic Core Of Murine Terminal Deoxynucleotidyl Transferase
1kcz	Crystal Structure Of Beta-Methylaspartase From Clostridium Tetanomorphum
1l0o	Crystal Structure Of The Bacillus Stearothermophilus Anti-Sigma Factor Spoiiab With The Sporulation Sigma Factor Sigmaf
1l7n	Transition State Analogue Of Phosphoserine Phosphatase
1lws	Crystal Structure Of The Intein Homing Endonuclease Pi-Scei Bound To Its Recognition Sequence
1mqw	Structure Of The Mt-Adprase In Complex With Three Mn2+ Ions And Ampcpr, A Nudix Enzyme
1muc	Structure Of Muconate Lactonizing Enzyme At 1.85 Angstroms Resolution
1n29	Crystal Structure Of The N1a Mutant Of Human Group Iia Phospholipase A2
1nml	Di-Haemic Cytochrome C Peroxidase From Pseudomonas Nautica 617, Form In (Ph 4.0)
1o8a	Crystal Structure Of Human Angiotensin Converting Enzyme (Native).
1os7	Crystal Structure Of Taud With Iron, Alpha-Ketoglutarate And Taurine Bound At Ph 7.5
1pym	Phosphoenolpyruvate Mutase From Mollusk In With Bound Mg2-Oxalate
1qum	Crystal Structure Of Escherichia Coli Endonuclease Iv In Complex With Damaged Dna
1r1j	Structural Analysis Of Neprilysin With Various Specific And Potent Inhibitors
1r44	Crystal Structure Of Vanx
1req	Methylmalonyl-Coa Mutase
1v25	Crystal Structure Of Tt0168 From Thermus Thermophilus Hb8
1vzx	Roles Of Active Site Tryptophans In Substrate Binding And Catalysis By Alpha-1,3 Galactosyltransferase
1w0h	Crystallographic Structure Of The Nuclease Domain Of 3'hexo, A Deddh Family Member, Bound To Ramp
2phk	The Crystal Structure Of A Phosphorylase Kinase Peptide Substrate Complex: Kinase Substrate Recognition
2toh	Tyrosine Hydroxylase Catalytic And Tetramerization Domains From Rat
7atj	Recombinant Horseradish Peroxidase C1a Complex With Cyanide And Ferulic Acid
1do8	Crystal Structure Of A Closed Form Of Human Mitochondrial Nad(P)+-Dependent Malic Enzyme
1ah7	Phospholipase C From Bacillus Cereus
1ca2	Human Carbonic Anhydrase Ii
1ck7	Gelatinase A
1dii	Crystal Structure Of P-Cresol Methylhydroxylase
1dl2	Crystal Structure Of Class I Alpha-1,2-Mannosidase From Saccharomyces Cerevisiae
1dqs	Crystal Structure Of Dehydroquinate Synthase (Dhqs) Complexed With Carbaphosphonate, Nad+ And Zn2+
1eb6	Deuterolysin From Aspergillus Oryzae
1ehk	Crystal Structure Of The Aberrant Ba3-Cytochrome-C Oxidase From Thermus Thermophilus
1ez2	Three-Dimensional Structure Of The Zinc-Containing Phosphotriesterase With Bound Substrate Analog Diisopropylmethyl Phosphonate
1f7l	Holo-(Acyl Carrier Protein) Synthase In Complex With Coenzyme A
1fa0	Structure Of Yeast Poly(A) Polymerase Bound To Manganate And 3'-Datp
1fcb	Molecular Structure Of Flavocytochrome B2
1fft	The Structure Of Ubiquinol Oxidase From Escherichia Coli

1foa	Crystal Structure Of N-Acetylglucosaminyltransferase I
1fsg	Toxoplasma Gondii Hypoxanthine-Guanine Phosphoribosyltransferase Complexed With 9-Deazaguanine, Alpha-D-5-Phosphoribosyl-1-Pyrophosphate (Prpp) And Two Mg2+ Ions
1fwj	Klebsiella Aerogenes Urease, Native
1g8k	Crystal Structure Analysis Of Arsenite Oxidase From Alcaligenes Faecalis
1gt7	L-Rhamnulose-1-Phosphate Aldolase From Escherichia Coli
1hxq	The Structure Of Nucleotidylated Galactose-1-Phosphate Uridylyltransferase From Escherichia Coli
1itq	Human Renal Dipeptidase
1j09	Crystal Structure Of Thermus Thermophilus Glutamyl-Trna Synthetase Complexed With Atp And Glu
1mpy	Structure Of Catechol 2,3-Dioxygenase (Metapyrocatechase) From Pseudomonas Putida Mt-2
1mqw	Structure Of The Mt-Adprase In Complex With Three Mn2+ Ions And Ampcpr, A Nudix Enzyme
1muc	Structure Of Muconate Lactonizing Enzyme
1n2c	Nitrogenase Complex From Azotobacter Vinelandii Stabilized By Adp-Tetrafluoroaluminate
1n20	(+)-Bornyl Diphosphate Synthase: Complex With Mg And 3-Aza-2,3-Dihydrogeranyl Diphosphate
1n62	Crystal Structure Of The Mo,Cu-Co Dehydrogenase (Codh), N-Butylisocyanide-Bound State
1ndo	Naphthalene 1,2-Dioxygenase
1nfs	Structure And Mechanism Of Action Of Isopentenylpyrophosphate-Dimethylallylpyrophosphate Isomerase: Complex With Nipp
1nia	The Structure Of Cu-Nitrite Reductase From Achromobacter Cycloclastes At Five Ph Values, With Nitrite Bound And With Type Ii Cu Depleted
1o98	Crystal Structure Of Phosphoglycerate Mutase From Bacillus Stearothermophilus Complexed With 2-Phosphoglycerate
1ogy	Crystal Structure Of The Heterodimeric Nitrate Reductase From Rhodobacter Sphaeroides
1pow	The Refined Structures Of A Stabilized Mutant And Of Wild-Type Pyruvate Oxidase From Lactobacillus Plantarum
1pvd	Crystal Structure Of The Thiamin Diphosphate Dependent Enzyme Pyruvate Decarboxylase From The Yeast Saccharomyces Cerevisiae At 2.3 Angstroms Resolution
1qlh	Horse Liver Alcohol Dehydrogenase Complexed To Nad Double Mutant Of Gly 293 Ala And Pro 295 Thr
1ra0	Bacterial Cytosine Deaminase D314g Mutant Bound To 5-Fluoro-4-(S)-Hydroxy-3,4-Dihydropyrimidine
1rdd	Crystal Structure Of Escherichia Coli Rnase Hi In Complex With Mg2+ At 2.8 Angstroms Resolution: Proof For A Single Mg2+ Site
1ru4	Crystal Structure Of Pectate Lyase Pel9a
1sml	Metallo Beta Lactamase L1 From Stenotrophomonas Maltophilia
1sox	Sulfite Oxidase From Chicken Liver
1ti6	Crystal Structure Of Pyrogallol-Phloroglucinol Transhydroxylase From Pelobacter Acidigallici Complexed With Inhibitor 1,2,4,5-Tetrahydroxy-Benzene
1uaq	The Crystal Structure Of Yeast Cytosine Deaminase
1uw8	Crystal Structure Of Oxalate Decarboxylase
1v04	Serum Paraoxonase By Directed Evolution

Table 1: List of all the M-set motifs with their protein names

PDBID	Function	Superfamily	EC class
1afr	Oxidoreductase	Ferritin	1.14.19.2
1alk	Alkaline Phosphatase	Alakaline phophotase	3.1.3.1
1aq2	Kinase	PEP carboxykinase	4.1.1.49
1b57	Lyase	Aldolase	4.1.2.13
1b66	Tetrahydrobiopterin Biosynthesis	Tetrahydrobiopterin biosynthesis enzymes	4.2.3.12
1bg0	Kinase	Guanido kinase N-terminal domain	2.7.3.3
1bzy	Phosphoribosyltransferase	PRTase	2.4.2.8
1c9u	Oxidoreductase	Soluble quinoprotein glucose dehydrogenase	1.1.5.2
1ctt	Hydrolase	Cytidine deaminase	3.5.4.5
1d8c	Lyase	Malate synthase G	2.3.3.9
1do6	superoxide reductase	Superoxide reductase	1.15.1.2
1e1a	Phosphotriesterase	Calcium-dependent phosphotriesterase	3.1.8.2
1e7l	Endonuclease	Recombination endonuclease VII, C-terminal and dimerization domains	3.1.22.4
1ez2	Hydrolase	Metallo-dependent hydrolases	3.1.8.1
1fr2	Immune system	Colicin E immunity proteins	3.1.21.1
1fua	Lyase	AraD/HMP-PK domain	4.1.2.17
1fui	Isomerase	FucI/AraA C-terminal domain	5.3.1.3
1g4p	Transferase	Thiamin phosphate synthase	2.5.1.3
1g72	Oxidoreductase	Methanol dehydrogenase subunit	1.1.2.7
5eat	Isoprenoid synthase	Terpenoid cyclases/Protein prenyltransferases	4.2.3.61
13pk	Kinase	Phosphoglycerate kinase	2.7.2.3
1bmt	methionine synthase	Methionine synthase domain	2.1.1.13
1ge7	Hydrolase	Metalloproteases ("zincins"), catalytic domain	3.4.24.20
1gim	ADENYLOSUCCINATE SYNTHETASE	P-loop containing nucleoside triphosphate hydrolases	6.3.4.4
1gog	GALACTOSE OXIDASE	Galactose-binding domain	1.1.3.9
1gp5	Oxidoreductase	Clavaminate synthase	1.14.11.19
1gqg	Oxidoreductase	RmlC-like cupins	1.13.11.24
1gsa	GLUTATHIONE SYNTHETASE	PreATP-grasp domain	6.3.2.3
1h19	Hydrolase	Leukotriene A4 hydrolase N-terminal domain	3.3.2.6
1hdh	Hydrolase	Alkaline phosphatase	3.1.6.1
1i1i	Hydrolase	Metalloproteases ("zincins"), catalytic domain	3.4.24.16
1i6p	Lyase	beta-carbonic anhydrase, cab	4.2.1.1
1ir3	Transferase	Protein kinase	2.7.10.1

1it4	Hydrolase	Phospholipase A2, PLA2	3.1.1.4
1j53	Transferase	Ribonuclease H	2.7.7.7
1jms	Transferase	DNA polymerase beta, N-terminal domain	2.7.7.31
1kcz	Lyase	Enolase C-terminal domain	4.3.1.2
1L0O	Protein Binding	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	2.7.11.1
1l7n	Hydrolase	HAD	3.1.3.3
1LWS	Hydrolase/dna	Hedgehog/intein (Hint) domain	3.6.3.14
1mqw	Hydrolase	Nudix	3.6.1.13
1muc	Isomerase	Enolase C-terminal domain	5.5.1.1
1n29	Hydrolase	Phospholipase A2, PLA2	3.1.1.4
1nml	oxidoreductase	Cytochrome c	1.11.1.5
1o8a	Metalloprotease	Metalloproteases ("zincins"), catalytic domain	3.4.15.1
1os7	Oxidoreductase	Clavaminate synthase	1.14.11.17
1pym	Phosphotransferase	Phosphoenolpyruvate/pyruvate domain	5.4.2.9
1qum	Hydrolase/dna	Xylose isomerase	3.1.21.2
1r1j	Hydrolase	Metalloproteases ("zincins"), catalytic domain	3.4.24.11
1r44	Hydrolase	Hedgehog/DD-peptidase	3.4.13.22
1req	Isomerase	Cobalamin (vitamin B12)-dependent enzymes	5.4.99.2
1v25	Ligase	Acetyl-CoA synthetase	6.2.1.3
1vzx	Transferase	Nucleotide-diphospho-sugar transferases	2.4.1.87
1w0h	Hydrolase	Ribonuclease H	3.1.-.-
2phk	Transferase	Protein kinase-like	2.7.11.1
2toh	Hydrolase	Aromatic aminoacid monoxygenases, catalytic and oligomerization domains	1.14.16.2
7atj	Oxidoreductase	Heme-dependent peroxidases	1.11.1.7
1do8	Oxidoreductase	NAD(P)-binding Rossmann-fold domains	1.1.1.38
1ah7	Hydrolase	Phospholipase C/P1 nuclease	3.1.4.3
1ca2	Lyase	Carbonic anhydrase	4.2.1.1
1ck7	Hydrolase	Metalloproteases ("zincins"), catalytic domain	3.4.24.24
1dii	Oxidoreductase	FAD-binding/transporter-associated domain	1.17.99.1
1dl2	Hydrolase	Seven-hairpin glycosidases	3.2.1.113
1dqs	Lyase	Dehydroquinate synthase	4.2.3.4
1eb6	Hydrolase	Metalloproteases ("zincins"), catalytic domain	3.4.24.39
1ehk	Oxidoreductase	Cupredoxins	1.9.3.1
1ez2	Hydrolase	Metallo-dependent hydrolases	3.1.8.1

1f7l	Transferase	4'-phosphopantetheinyl transferase	2.7.8.7
1fa0	Transferase	PAP/Archaeal CCA-adding enzyme, C-terminal domain	2.7.7.19
1fcb	Oxidoreductase	FMN-linked oxidoreductases	1.1.2.3
1fft	Oxidoreductase	Cupredoxins	1.10.3.10
1foa	Transferase	Nucleotide-diphospho-sugar transferases	2.4.1.101
1fsg	Transferase	PRTase	2.4.2.8
1fwj	Hydrolase	Metallo-dependent hydrolases	3.5.1.5
1g8k	Oxidoreductase	Formate dehydrogenase/DMSO reductase, domains 1- 3	1.20.9.1
1gt7	Lyase	AraD/HMP-PK domain	4.1.2.19
1hxq	Nucleotidyltransferase	HIT	2.7.7.12
1itq	Hydrolase	Metallo-dependent hydrolases	3.4.13.19
1j09	Ligase	Nucleotidylyl transferase	6.1.1.17
1mpy	Oxidoreductase	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	1.13.11.2
1mqw	Hydrolase	Nudix	3.6.1.13
1muc	Isomerase	Enolase C-terminal domain	5.5.1.1
1n2c	Complex of Nitrogenase Proteins	P-loop containing nucleoside triphosphate hydrolases	1.18.6.1
1n20	Isomerase	Terpenoid synthases	5.5.1.8
1n62	Oxidoreductase	CO dehydrogenase molybdoprotein N-domain	1.2.99.2
1ndo	Non Heme Iron Dioxygenase	ISP domain	1.14.12.12
1nfs	Isomerase	Nudix	5.3.3.2
1nia	Oxidoreductase	Cupredoxins	1.7.2.1
1o98	Isomerase	Alkaline phosphatase	5.4.2.12
1ogy	Oxidoreductase	Multiheme cytochromes	1.7.99.4
1pow	Oxidoreductase	DHS-like NAD/FAD-binding domain	1.2.3.3
1pvd	Lyase	DHS-like NAD/FAD-binding domain	4.1.1.1
1qlh	Oxidoreductase	NAD(P)-binding Rossmann-fold domains	1.1.1.1
1ra0	Hydrolase	Metallo-dependent hydrolases	3.5.4.1
1rdd	Hydrolase	Ribonuclease H	3.1.26.4
1ru4	Lyase	Pectin lyase	4.2.2.2
1sml	Hydrolase	Metallo-hydrolase/oxidoreductase	3.5.2.6
1sox	Oxidoreductase	Cytochrome b5-like heme/steroid binding domain	1.8.3.1
1ti6	Oxidoreductase	Cna protein B-type domain	1.97.1.2
1uaq	Hydrolase	Cytidine deaminase	3.5.4.1
1uw8	Lyase	RmlC-like cupins	4.1.1.2
1v04	Hydrolase	Calcium-dependent phosphotriesterase	3.1.1.2

Table 2: List of all the M-set Motifs with the proposed function and the superfamily the proteins along with their EC Number.