COMPUTATIONAL SCREENING OF NATURAL PROTEINS FOR MORITA-BAYLIS-HILLMAN ACTIVITY

by

Belkıs Akbulut

Submitted to Graduate School of Natural and Applied Sciences
in Partial Fulfillment of the Requirements
for the Degree of Master of Science in
Chemical Engineering

Yeditepe University

2020

COMPUTATIONAL SCREENING OF NATURAL PROTEINS FOR MORITA-BAYLIS-HILLMAN ACTIVITY

APPROVED BY:

Assoc. Prof. Dr. Nihan Çelebi Ölçüm .....
(Thesis Supervisor)
(Yeditepe University)

Assoc. Prof. Dr. Hilal Demir Kıvrak .....
(Van Yüzüncü Yıl University)

Assoc. Prof. Dr. Tuğba Davran Candan .....
(Yeditepe University)

DATE OF APPROVAL: ..../..../2020

ACKNOWLEDGEMENTS

I would like to express my thanks to my MSc supervisor Assoc. Prof. Nihan Çelebi Ölçüm for her valuable guidance, help and support during my MSc study. Apart from this study, being involved in her research group helped me get a valuable experience and courage to work harder in my studies.

This study was supported by TUBITAK (113Z614). The calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Finally, I would like to thank my dear family for their most valuable, lifetime support.

ABSTRACT

COMPUTATIONAL SCREENING OF NATURAL PROTEINS FOR MORITA-BAYLIS-HILLMAN ACTIVITY

Morita-Baylis-Hillman (MBH) reaction is a versatile C-C coupling reaction between an α,β-unsaturated carbonyl compound and an aldehyde, activated ketone or other carbon electrophiles in the presence of a nucleophilic catalyst. MBH adducts are densely functionalized molecules displaying wide range of biological activities. However, even in the presence of the most efficient organocatalysts, reaction rate is exceedingly slow for substrates of pharmaceutical interest.

The aim of this project is to computationally explore natural proteins that contain the required catalytic machinery in the optimal three-dimensional arrangement to catalyze the MBH reaction. Two main theozyme models were used as templates to construct various catalytic atom maps (CAMs) which were later used to screen the protein databank. Promising matches (2BDB, 2NW6, 1FFE) were subjected to molecular dynamics (MD) simulations and the catalytic contacts were evaluated in a dynamic solvated environment. MD simulations showed that 2BDB could potentially show promiscuous activity for the MBH substrates that can be enhanced using active site redesign in a further study.

ÖZET

MORITA-BAYLIS-HILLMAN AKTİVİTESİ İÇİN DOĞAL PROTEİNLERİN HESAPSAL TARANMASI

Morita-Baylis-Hillman (MBH) reaksiyonu α,β-doymamış karbonil bileşeni ve bir aldehit, aktive edilmiş keton ya da diğer karbon elektrofilleriyle bir nükleofilik katalist varlığında meydana gelen çok yönlü bir C-C bağlanma reaksiyonudur. MBH ürünleri geniş çeşitlilikte biyolojik aktivite gösteren yoğun biçimde fonksiyonel hale getirilmiş moleküllerdir. Ancak, en etkili organokatalistlerin bile kullanımında reaksiyon hızı, ilaç kullanımına ait bir hammaddenin üretimi söz konusu olduğunda oldukça yavaştır.

Bu projenin amacı, MBH reaksiyonunu katalize eden, üç boyutlu dizilimlerinde gerekli katalitik mekanizmayı barındıran doğal proteinleri hesapsal olarak araştırmaktır. Çeşitli katalitik atom haritalarını (KAM) oluşturmak için iki ana teozim modeli şablon olarak kullanıldı ve daha sonra bunlar protein data bankasını taramak için kullanıldı. En fazla katalitik potansiyel gösteren protein eşleşmeleri (2BDB, 2NW6, 1FFE), moleküler dinamik simülasyonlarında (MD) kullanıldı ve katalitik kontaktlar dinamik çözeltili bir ortamda değerlendirildi. MD simülasyonları, 2BDB'nin potansiyel olarak aktif bölge yeniden tasarlamasıyla hızı arttırılabilecek MBH substratlarına yönelik katalitik aktivite sahibi olduğunu gösterdi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

ΔG	Free energy of reaction (kcal/mol)
ºC	Degree centigrade
α	Alpha
β	Beta
ALA	Alanine
ASN	Asparagine
ASP	Aspartame
CAM	Catalytic atom map
CSA	Catalytic site atlas
Ea	Activation energy
GLU	Glutamine
GLY	Glycine
HIS	Histidine
MBH	Morita-Baylis-Hillman
NMR	Nuclear magnetic resonance
PDB	Protein data bank
pKa	Acid dissociation constant
RMS	Root mean square
RMSD	Root mean square deviation
SABER	Selection of active/binding sites for enzyme redesign
SER	Serine
THR	Threonine
TS	Transition State
TYR	Tyrosine
WCN	Weighted contact number
MeCN	Acetonitrile

1. INTRODUCTION

Morita-Baylis-Hillman (MBH) is a C-C bond formation reaction which yields multifunctional compounds from simple reactants by integrating aldol and Michael reactions in one catalytic cycle. (Figure 1.1.)

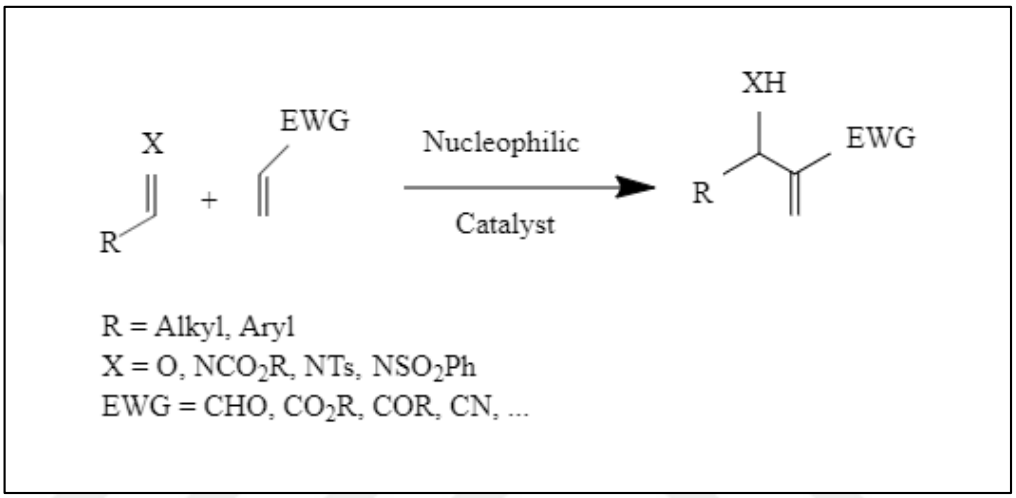


Figure 1.1. MBH Reaction

MBH products display wide range of biological activities such as antitumoral activity, antifungal activity, antibacterial activity (Figure 1.2.) [1, 2]. However this bimolecular, multistep reaction requires long reaction times especially for substrates of pharmaceutical interest in the presence of expensive chemical catalysts to give low to moderate yields. It can take up to several days or weeks to achieve a decent yield of products which prevents it from being incorporated at industrial level or even in smaller scales [3, 4]. DABCO, pyrrocoline, quinuclidine are some of the most effective chemical catalysts currently used but there is still room for improvement both for reaction rate and selectivity of the catalysts. Biocatalysts emerge as important green alternatives for this purpose. Biocatalysts are basically protein structured enzymes that increase the rate of a chemical reaction. They can be modified to increase the efficiency for a non-natural reaction thanks to protein engineering studies and provide very good selectivity due to their large, 3-D structure that surrounds the substrate.

Biocatalysts are manufactured by using renewable, affordable resources and they are biodegradable which makes them ideal in terms of green chemistry and sustainability [5]. Considering the significant advantages of the MBH reaction and biocatalysts, studies have been conducted for identification of naturally occurring protein structures of the MBH reaction to improve the rate and yield in recent years. Reetz et al. obtained little conversion rate (10 percent after 5 days) as a result of using lipases for MBH reaction and 15 percent conversion with BSA (fraction V) after 2 days which is somewhat higher. Jiang and Yu reported that the MBH reaction between 4-nitrobenzaldehyde and methyl vinyl ketone achieves a yield of 46 percent with E.coli biotin esterase in acetonitrile after 96 hours. M. Kapoor et al. confirmed that lipases catalyze the MBH reaction with the right reaction medium. Among the lipases tested, Burkholderia cepecia lipase has the best conversion rate with 96 percent inside 50 percent (v/v) DMSO as a result of series of experiments and H NMR, HPLC analysis of the experiment results [6].

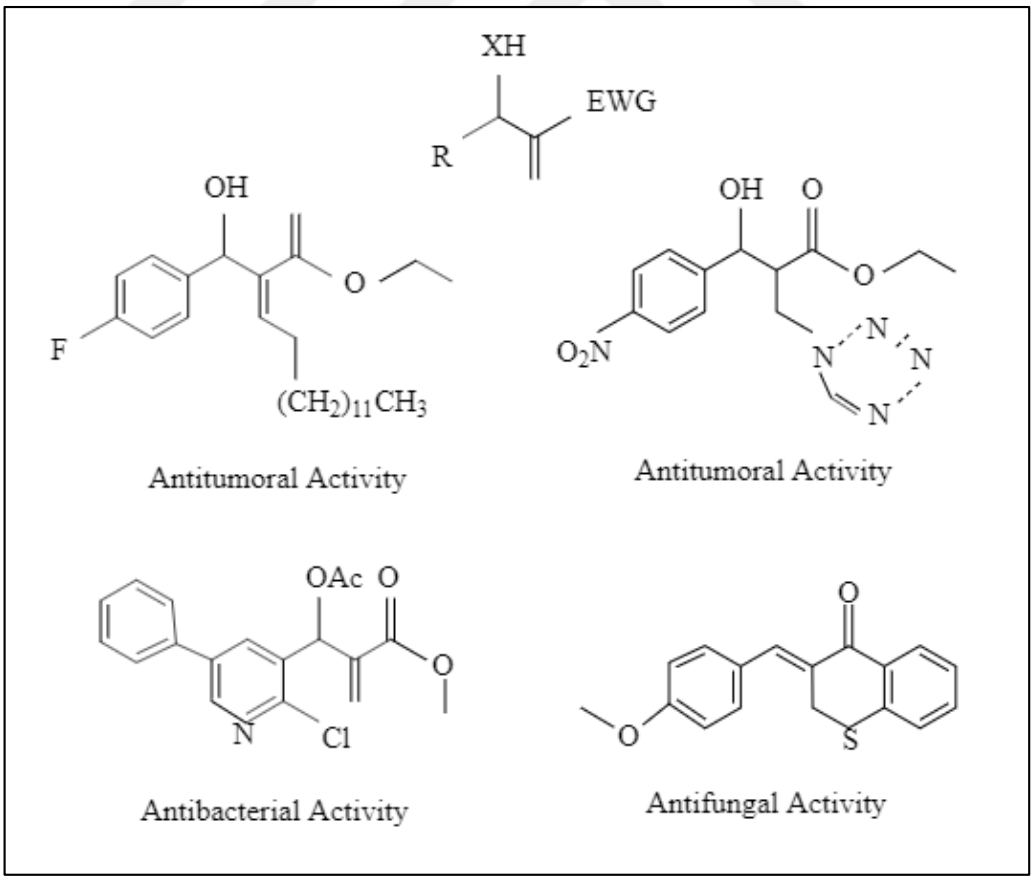


Figure 1.2. Various biological activities of MBH products

This project introduces a new computational protocol to identify proteins that have potential to show catalytic activity and promiscuous enzyme activity towards the MBH reaction. Motivation for this study is enhanced by the fact that the functional groups required to catalyze the MBH reaction is already present in the active sites of many natural enzymes that work with a nucleophilic mechanism, for instance; lipases, esterases and proteases. (Figure 1.3.)

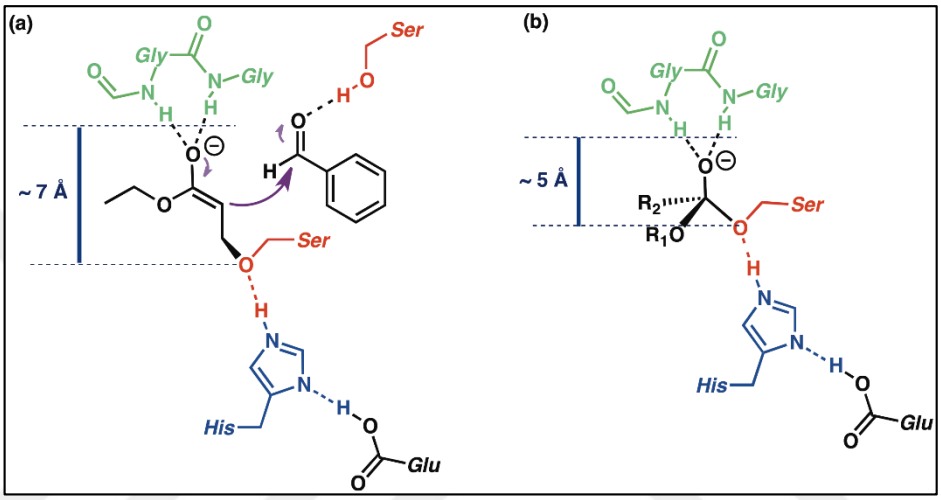


Figure 1.3. (a) Catalytic functional groups essential for MBH reaction (b) Catalytic amino acid groups present in serine esterase active site

Enzymes are protein structured biological catalysts that increase the rate of a reaction by maintaining a lower activation energy without interfering with reaction's equilibrium. They leave the reaction without going through any change themselves or being consumed during the process. Therefore they can be used to catalyze a specific reaction numerous times. Reactants which enzymes act upon are called substrates [7]. In all chemical reactions, substrates are transformed into an unstable structure with high energy called the activated complex in the transition state. The reactant molecules must carry enough energy to form the activated complex so that the reaction can be completed. However transition states occur very fast with a life span of a few femtoseconds which roughly makes $1/10^{13}$ secs. The activated complex deforms very fast to form the products. Therefore spectroscopic or physical methods are unable to identify the characteristics of the transition states. Enzymes enhance the reaction rate by lowering the energy reactants must bare in order to form the activated complex in transition state [8]. They accomplish this by altering the substrate's electron distribution.

Enzymes can change the electronic arrangement by proton detachment, electron removal, proton addition, hydrophobic detachment, geometric deformation and Lewis acid-base contact. These processes are achieved by arrangement-wise changes in sequential protein and substrates. These forces themselves may be too small to affect the substrate's structure but when united, they form a substantial energy enough to move electrons in order to break existing bonds and make new ones to form the products [9].

Enzymes enable the biochemical reactions to occur at a feasible rate such as without the presence of enzymatic catalysts 1 second of enzymatic reaction could take up to years [9]. Considering the human life and ecology mostly depends on biological reactions, enzymes are of vital importance. They are a critical part of key reactions such as digestion, muscle movements, nervous system etc [7].

Enzyme catalysis is a subject of great interest due to enzymes' efficient, enantioselective and environmentally friendly nature. These qualities make enzymes preferable to non-enzymatic catalysts but despite all their advantages natural enzymes, on their own, do not offer catalytic activity for many important industrial processes. In order to be used in these processes, enzymes are expected to have strong enantioselectivity and catalytic activity against some important substrates. Also, they are expected to maintain their stability while stored and while being exposed to high temperature, pH levels, substrate and product concentrations. In order to meet all these expectations enzyme engineering is applied to natural catalysts. [10] It seems as once the enzymatic operations are thoroughly understood then we can design synthetic (*de novo*) enzymes for many industrial processes which are currently insufficiently and synthetically catalyzed [11]. In vitro experiments are one of the first options that come to mind however they require a lot of repeating which means having to spend a lot of time, money and labor. By using in silico methods, efficiency of in vitro experiments can be considerably increased since the number of target molecules will be reduced [12].

Enzyme promiscuity means that the active site of an enzyme is able to catalyze a distinct chemical reaction of other substrate(s) than the one it is mainly used for. One of the earliest studies about promiscuity is from 1965 by Mark et al. and the topic still has been a point of interest to this date. Even though the promiscous enzyme activity was not considered to be common in the beginning, an increasing number of this activity is being observed in a wide range of reactions such as the carbon-carbon and carbon-heteroatom bond formations [12]. A current study by Feng et al. explains the enzymatic catalysis of decarboxylative aldol

reaction and decarboxylative Knoevenagel reaction which are both catalyzed by acrylic resin immobilized Candida antarctica lipase B, CAL-B, with very good yields at a time range between 12 – 72 hours [13]. Another study explains the phosphodiesterase and phosphotriesterase behaviour of some aminopeptidases in addition to their natural hydrolytic activity for peptides. Phosphotriesterase activity is especially important here because this enzyme class is able to hydrolyze manufactured formulations that are used as insecticides and chemical warfare agents which means that they show enzymatic promiscuity by catalyzing a reaction that does not naturally occur [14, 15]. Enzymatic promiscuity identification can often be made based on similarity of structure, sequence or mechanism. For example, it is mostly identified in proteins that belong to distinct superfamilies and share common scaffolds along with inherited enzymatic strategy. Another way of identifying promiscuity is usage of cofactors in common. Soo et al. reported the identification of mutual enzymatic promiscuity between pyridoxal 5-phosphate dependent enzymes so that alanine racemase shows promiscuity for cystathionine β-lyase and also in reverse.

Enzyme promiscuity is important as it initiates the new enzyme evolution and provides input for protein engineering to be used in synthesis of pharmaceuticals or various chemicals. The catalytic activity measured for a promiscuous activity may be small when compared to an enzyme's main function however it can provide accelaration rate up to 10^{26} and can be increased even more with even a little mutation. A very significant example is the improvement of promiscuous o-succinylbenzoate synthase functionality in the muconate lactonizing enzyme II more than 10^6 times by only replacing Glu323 with Gly [16].

SABER (Selection of Active/Binding sites for Enzyme Redesign) is a software tool used to investigate the protein data bank according to proteins' functional groups among many computational methods used for rational enzyme redesign. SABER finds the residues that might possibly have catalytic properties and therefore provides the information about active sites that are suitable for computational redesign. One of the major accomplishments of this tool is the design of Kemp eliminase by determining the enzymes which have o-succinyl benzoate synthase catalytic residues [17].

In this study, SABER tool is used to examine the Protein Data Bank according to specific functional groups required for the catalysis of the MBH reaction by use of theozymes (short for theoretical enzymes) [5]. SABER uses a number of modules to search the protein database to find matching protein structures based to the theozyme model used [18]. Its

output is analyzed in a spreadsheet program and the best catalyst candidates from natural enzyme active sites are determined by using a few criteria such as Root Mean Square Deviation (RMSD) value and the protein type. Finally, these catalyst candidate protein structures are tested in Molecular Dynamics (MD) simulations to evaluate their enzymatic performance and stability under given circumstances.

2. METHODOLOGY

This study involves computational screening of the protein databank (RSCB) using SABER followed by molecular dynamics analysis of some of these protein structures that can be used as biocatalysts for the Morita-Baylis-Hillman (MBH) reaction.

A theozyme is a theoretical enzyme model that consists of the 3D constellation of amino acid side mimics around the transition state of a target reaction that allows maximum TS stabilization [1]. Theozymes here calculated using density functional theory at the B3LYP/6-31G(d) level and they were used as templates for finding catalysts for a certain reaction such as MBH in this case.

Heteroatoms from the theozymes were selected to construct the catalytic atom maps (CAM) by using their xyz coordinates. An xml file is obtained from CAM (xyz) after format conversions. Tolerance of the screening (delta) was set as 2 Å and protein databank was screened for each CAM structure using SABER. SABER results were analyzed in a spreadsheet program by first filtering according to root mean square deviation (rmsd) and delta. In this case, rmsd was preferred to be less than 1.3 and delta less than 2. Selected proteins were incorporated in MD analysis to observe the interactions between molecules and their stability. In a dynamic and solvated environment, MD simulations were run for 400 ns each. Data collection rate was set as 100 ps.

2.1. SELECTION OF ACTIVE/BINDING SITES FOR ENZYME REDESIGN

Selection of Active/Binding sites for Enzyme Redesign (SABER) is a tool which helps find protein alternatives for enzyme redesign based on functional group similarity. SABER finds the residues that might have catalytic properties for the specific theozyme structure and provides information about active sites that can be applied for a specific reaction.

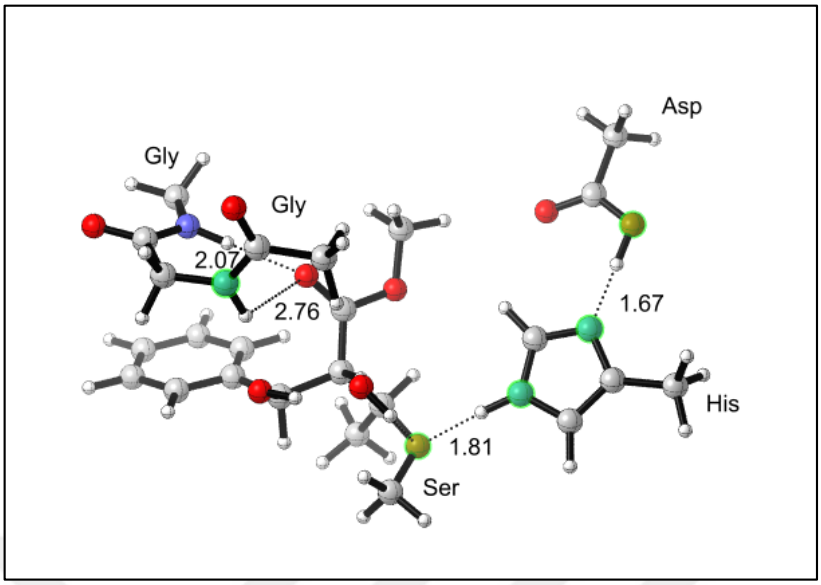


Figure 2.1. CAM atoms highlighted on theozyme model

Once CAM is constructed its atoms' coordinates were saved in xyz format when clicked on "make XYZ" button in Cylview. (Figure 2.2)

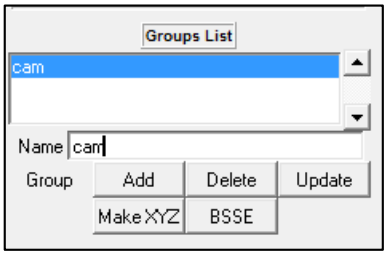


Figure 2.2. XYZ coordinate file creation in Clyview

Figure 2.3 shows the xyz file of selected atoms. This file was the first step of this procedure.

```
catdits3
O          8.0394051          -0.5706203           1.1776582
N          5.4854051          -0.9596203           0.4396582
N          3.4994051          -0.2346203          -0.1993418
O          0.5444051           0.0283797           0.9396582
N         -4.2215949          -0.7456203           2.1646582
```

Figure 2.3. XYZ coordinates of selected atoms

Next step was to convert xyz file to a tess file by using script “xyz2tess”. The xyz file was arranged in a specific format, which is shown in Figure 2.4.

```
ATOM      0  OD2 ASP A 1       8.039  -0.571   1.178
ATOM      0  NE2 HIS A 2       5.485  -0.960   0.440
ATOM      0  ND1 HIS A 3       3.499  -0.235  -0.199
ATOM      0  OG  SER A 4       0.544   0.028   0.940
ATOM      0  N   GLY A 5      -4.222  -0.746   2.165
END
```

Figure 2.4. Example of a tess file

Process is followed by conversion of tess file to an xml file by using “tess2jess filename > filename.xml” command. Example of an xml file obtained was shown below in Figure 2.5.

```
<?xml version="1.0"?>
<jess xmlns="http://www.ebi.ac.uk/~jbarker/Jess">
    <template name="freq-ts-tox6_ts3_w2_sonuncu-5.tess">
        <atom name="0">
            <select>
                and
                (
                    isAtomNamed("_OD1"),
                    inResidueNamed("ASP")
                )
            </select>
            <xyz>8.039 -0.571 1.178</xyz>
        </atom>
        <atom name="1">
            <select>
                and
                (
                    isAtomNamed("_ND1"),
                    inResidueNamed("HIS"),
                    inAnnulus(atom(#0),sub(2.69,$delta),add(2.69,$delta)),
                    inChainOf(atom(#0))
                )
            </select>
            <xyz>5.485 -0.960 0.440</xyz>
        </atom>
        <atom name="2">
            <select>
                and
                (
                    isAtomNamed("_NE2"),
                    inResidueNamed("HIS"),
                    inAnnulus(atom(#0),sub(4.77,$delta),add(4.77,$delta)),
                    inChainOf(atom(#0)),
                    inAnnulus(atom(#1),sub(2.21,$delta),add(2.21,$delta))
                )
```

Figure 2.5. Example of an xml file

```
</select>
            <xyz>3.499 -0.235 -0.199</xyz>
        </atom>
        <atom name="3">
            <select>
                and
                (
                    isAtomNamed("_OG_"),
                    inResidueNamed("SER"),
                    inAnnulus(atom(#0),sub(7.24,$delta),add(7.24,$delta)),
                    inChainOf(atom(#0)),
                    inAnnulus(atom(#1),sub(4.94,$delta),add(4.94,$delta)),
                    inAnnulus(atom(#2),sub(2.83,$delta),add(2.83,$delta))
                )
            </select>
            <xyz>0.544 0.028 0.940</xyz>
        </atom>
        <atom name="4">
            <select>
                and
                (
                    isAtomNamed("_N__"),
                    inResidueNamed("GLY"),
                    inAnnulus(atom(#0),sub(9.94,$delta),add(9.94,$delta)),
                    inChainOf(atom(#0)),
                    inAnnulus(atom(#1),sub(8.78,$delta),add(8.78,$delta)),
                    inAnnulus(atom(#2),sub(7.64,$delta),add(7.64,$delta)),
                    inAnnulus(atom(#3),sub(6.64,$delta),add(6.64,$delta))
                )
            </select>
            <xyz>-4.222 -0.746 2.165</xyz>
        </atom>
    </template>
</jess>
```

Figure 2.5. Example of an xml file

Although this .xml file can be run in Saber program, its results will be very limited since the functional groups in it will only allow the search to look for those specific atom types in specific amino acids. However similar functional groups are contained by various amino acids. So format of xml file was manually edited in a way that it will diversify the results. This new format contains “and/or” options that allows Saber to match more functionally similar amino acids.

For example, Aspartate’s oxygen can be substituted with Glutamate’s oxygen atom. The xml file with edited with “and/or” options is given in Figure 2.6.

```
<?xml version="1.0"?>
<jess xmlns="http://www.ebi.ac.uk/~jbarker/Jess">
    <template name="freq-ts-tox6_ts3_w2_sonuncu.tess">
        <atom name="0">
            <select>
```

Figure 2.6. Example of an xml file in “and/or” format

```
            and
                (
                    or
                    (
                        and
                        (
                            or
                            (
                                isAtomNamed("_0D2"),
                                isAtomNamed("_0D1")
                            ),
                            inResidueNamed("ASP")
                        ),
                        and
                        (
                            or
                            (
                                isAtomNamed("0E2"),
                                isAtomNamed("0E1")
                            ),
                            inResidueNamed("GLU")
                        )
                    )
                )
            </select>
            <xyz>8.039 -0.571 1.178</xyz>
        </atom>
        <atom name="1">
            <select>
                and
                (
                    or
                    (
                        isAtomNamed("_NE2"),
                        isAtomNamed("_ND1")
                    ),
                    inResidueNamed("HIS"),
                    inAnnulus(atom(#0),sub(2.69,$delta),add(2.69,$delta)),
                    inChainOf(atom(#0))
                )
            </select>
            <xyz>5.485 -0.960 0.440</xyz>
        </atom>
        <atom name="2">
            <select>
                and
                (
                    or
                    (
                        isAtomNamed("_ND1"),
                        isAtomNamed("_NE2")
                    ),
                    inResidueNamed("HIS"),
                    inAnnulus(atom(#0),sub(4.77,$delta),add(4.77,$delta)),
                    inChainOf(atom(#0)),
                    inAnnulus(atom(#1),sub(2.21,$delta),add(2.21,$delta))
                )
            </select>
            <xyz>3.499 -0.235 -0.199</xyz>
        </atom>
        <atom name="3">
            <select>
```

Figure 2.6. Example of an xml file in “and/or” format (Continued)

```
    and
            (
                or
                (
                    and
                    (
                        isAtomNamed("_OG_"),
                        inResidueNamed("SER")
                    ),
                    and
                    (
                        isAtomNamed("_OG1"),
                        inResidueNamed("THR")
                    ),
                    and
                    (
                        isAtomNamed("_OH_"),
                        inResidueNamed("TYR")
                    )
                ),
                inAnnulus(atom(#0),sub(7.24,$delta),add(7.24,$delta)),
                inChainOf(atom(#0)),
                inAnnulus(atom(#1),sub(4.94,$delta),add(4.94,$delta)),
                inAnnulus(atom(#2),sub(2.83,$delta),add(2.83,$delta))
            )
        </select>
        <xyz>0.544 0.028 0.940</xyz>
    </atom>
    <atom name="4">
        <select>
            and
            (
                or
                (
                    and
                    (
                        isAtomNamed("_N__"),
                        inResidueNamed("GLY")
                    ),
                    and
                    (
                        isAtomNamed("_N__"),
                        inResidueNamed("ALA")
                    ),
                    and
                    (
                        isAtomNamed("_ND2"),
                        inResidueNamed("ASN")
                    ),
                    and
                    (
                        isAtomNamed("_NE2"),
                        inResidueNamed("GLN")
                    )
                ),
                inAnnulus(atom(#0),sub(9.94,$delta),add(9.94,$delta)),
                inChainOf(atom(#0)),
                inAnnulus(atom(#1),sub(8.78,$delta),add(8.78,$delta)),
                inAnnulus(atom(#2),sub(7.64,$delta),add(7.64,$delta)),
                inAnnulus(atom(#3),sub(6.64,$delta),add(6.64,$delta))
            )
        </select>
        <xyz>-4.222 -0.746 2.165</xyz>
```

Figure 2.6. Example of an xml file in “and/or” format (Continued)

```
                    inAnnulus(atom(#4),sub(2.79,$delta),add(2.79,$delta))
                )
            </select>
            <xyz>-4.222 -0.746 2.165</xyz>
        </atom>
    </template>
</jess>
```

Figure 2.6. Example of an xml file in "and/or" format (Continued)

SABER [17] consists of six stages. Stages in between were briefly described below.

2.1.1. Jess

Jess rapidly analyses large data series by using geometric hashing to search for protein structures and find matches to atom arrangements such as CAMs. Running speed of the program depends on the limitations set in the beginning. For example when strict limitations are set, number of results obtained will be smaller than the case where less limitations were set. Although loose limitations increase the number of results they could also easily fill the storage quota so, the program must be monitored regularly to avoid this. Jess program was run by using "[user@system~]$ jess -t CAM.*xml* -M filelist.*txt* -f -p -d" and delta variable as 2.0 [17].

2.1.2. Summary Builder

Jess output is the input of Summary Builder program which is a bash script that transforms its input into a dense text for further examination. Summary builder was run by using "[user@system~]$ bash summary_builder_v4.s <*Jess_file*.txt >*output_file*.txt". The output of this program was directly forwarded to the next stage unless any errors were met. [17]

2.1.3. Active Site Finder

ActiveSiteFinder uses SummaryBuilder's output as an input. Through its input, it locates information according to the residues identified by Jess program. It searches the protein's entry in the Catalytic Site Atlas (CSA) and compares the residues in the match to the known active site residues defined in the CSA. Result of this process, protein name and EzCat identifier was added to the output file to be forwarded to the next step in the Saber analysis. This program was run by using "[user@system~]$ perl as_finder_v6.pl" [17].

2.1.4. Active Site Score

ActiveSiteScore is a Perl script that uses the residues spotted in Jess search and compares them to the protein's known active site residues which were recognized from data obtained from Catalytic Site Atlas by ActiveSiteFinder. This program was run by entering "[user@system~]$ perl saber as_score_v3.pl" [17].

2.1.5. Binding Site Finder

BindingSiteFinder program supplies an alternative way to identify the active and/or binding sites in a protein structure. It looks for nonwater PDB heteroatoms within 5 Å of the CAM's residues. Matches are labeled accordingly if any nonwater heteroatom is present. Additionally it contains the name of the heteroatom, the ligand and the closest distance between the heteroatom and the catalytic residues. This program was run by entering "[user@system~]$ perl bsf_v6.pl" [17].

2.1.6. SABER_wcn

SABER_wcn module which is a Perl script, was written to perform weighed contact number (wcn) analysis. This module works with a given PDB identifier or set of identifiers and requires a WCN summary file for the proteins of concern which is done by another program called wcn_filegen. By using this file, saber_wcn adresses the highest and the average catalytic atom WCN, and the Cα WCN for a given residue.This program was run by using

"[user@system~]$ perl saber_wcn_v10.pl". For each residue identified by the Catalytic Atom Map, this module reports three WCN values. These are added to the output as three per residue in the same order that the residues are in the input file. The first WCN value is the highest catalytic atom WCN, the second is average and the final one is Cα WCN. This order is always followed so which values belong to each residue can be determined easily.

2.1.7. SABER Output

Output of saber_wcn file can be examined by excel easily since every field is separated by "|" symbol. Each field was always listed in the same way as in Table 2.1 up to 10. The other numbers depend on the number of catalytic residues. In Table 2.1, only one catalytic residue in the CAM was assumed [17].

Table 2.1. Data arrangement of SABER output

Column	**Contents**
1	PDB identifier
2	RMSD vs CAM (Å)
3	Maximum displacement from CAM (Å)
4	Protein name
5	ActiveSiteScore
6	EzCat identifier
7	BindingSiteFinder flag: yes/no
8	BindingSiteFinder data: Closest residue name, number and atom type
9	BindingSiteFinder data: Heteroatom residue name, number and atom type
10	BindingSiteFinder data: Heteroatom distance
11	WCN value of residue, heavy atom maximum
12	WCN value of residue, heavy atom average
13	WCN value of residue, Cα
14	Residues identified by the CAM

15	Predicted pKa of residue
16+	Known active site residues for this PDB ID, taken from the Catalytic Site Atlas

2.2. MOLECULAR DYNAMICS

MD simulations calculate the physical properties such as equilibrium and transport characteristics of a classical many body systems over a certain period. Classical meaning that system particles follows the classical mechanics laws. MD simulations have a lot in common with regular laboratory experiments. Such as they both measure a specific property during a time range and see how it develops. In both cases, accuracy of the results improves with larger time span [19].

In an MD simulation, Newton's equations of motion for a model system sample is solved until the system specifications are stable with respect to time which means the system is equilibrated. Once the equilibration is achieved properties of the system can be measured. In order to do this measurement, first this property must be written in terms of the momenta and positions of the system's constituents [19]. As an example, kinetic energy of each degree of freedom in a classical many body system is measured as follows;

$$\frac{1}{2}mv_{\alpha}^{2}=\frac{1}{2}k_{B}T \tag{2.1}$$

This equation can be used to describe the temperature of a simulation whereas in real life dividing the kinetic energy of a system by the number of degrees of freedom, N_f, gives the instantaneous temperature. In order to get a correct temperature evaluation, a high number of fluctuations must be analyzed and averaged [19].

$$T(t)=\sum_{i=1}^{N}\frac{m_i v_i^2(t)}{k_B N_f} \tag{2.2}$$

2.2.1. MD Algorithms

First step of an MD simulation is to define the parameters of the run such as starting temperature, time intervals, count of particles etc [19]. In order to do this, we first need an algorithm such as the one given in Figure 2.7.

Algorithm 3 (A Simple Molecular Dynamics Program)

```
program md                      simple MD program

call init                       initialization
t=0
do while (t.lt.tmax)            MD loop
   call force(f,en)             determine the forces
   call integrate(f,en)         integrate equations of motion
   t=t+delt
   call sample                  sample averages
enddo
stop
end
```

Figure 2.7. A basic MD algorithm

First step of a simple algorithm is selecting starting velocities and positions for the particles. Then the forces applied to them are calculated. Next step along with the force calculation are the main steps of the program which is the integration of Newton's equations of motion. These last two steps iterate along the given time period for the simulation. The line "call sample" is run to obtain the mean values of pressure or temperature which is the final step of the algorithm. This algorithm simply shows the overall steps of a MD simulation. However, each of these steps here are actually run by separate algorithms [19]. Figure 2.8 is the algorithm used to initiate the simulation.

Algorithm 4 (Initialization of a Molecular Dynamics Program)

```
subroutine init              initialization of MD program
sumv=0
sumv2=0
do i=1,npart
  x(i)=lattice_pos(i)        place the particles on a lattice
  v(i)=(ranf()-0.5)          give random velocities
  sumv=sumv+v(i)             velocity center of mass
  sumv2=sumv2+v(i)**2        kinetic energy
enddo
sumv=sumv/npart              velocity center of mass
sumv2=sumv2/npart            mean-squared velocity
fs=sqrt(3*temp/sumv2)        scale factor of the velocities
do i=1,npart                 set desired kinetic energy and set
  v(i)=(v(i)-sumv)*fs        velocity center of mass to zero
  xm(i)=x(i)-v(i)*dt         position previous time step
enddo
return
end
```

Figure 2.8. Initialization algorithm

First step of the initialization is to define the velocities and locations of each particle of the system at the starting point. Particle locations should be defined in such a way that it will be suitable with the simulation arrangement so that it will not cause any overlapping of the particles. This is usually accomplished by locating the atoms in a cubical frame such as in the initialization algorithm shown in Figure 2.8 [19].

In this case, the assumption that the starting temperature and density values cause the cubical system to be unsteady and heat up fast will be taken into consideration. Each fragment is first placed on their own location in the cubicle frame and a velocity value in [-0.5, 0.5] range is assigned to them. Then, total momentum is reduced to zero by shifting particle velocities and these velocity values are adjusted to obtain the target kinetic energy value. Assuming that following equation applies when the system is thermally equilibrated [19] (where v_α represents the α constituent of a fragment's velocity);

$$<v_\alpha^2>=k_B\,T/m \tag{2.3}$$

by using equation 2.3.3, temperature at any time t can be defined as follows;

$$k_B T(t) = \sum_{i=1}^{N} \frac{m v_{\alpha,i}^2(t)}{N_f} \tag{2.4}$$

As seen from this equation, temperature at any time T (t) can be changed to a specific value by adjusting the velocities with the factor of $(T(t))^{1/2}$ however the starting temperature here is not very important since the temperature will vary along equilibration [19].

In order to solve Newton's Equation of Motion, we make use of the current position (x) and the previous position (x_m) of the particles and the force (f) that acts on them to foresee their next position rather than using their velocities. Considering only linear momentum conservation law of mechanics, a particle position at time i can be estimated as follows;

$$x_m(i) = x(i) - v(i)\,dt \tag{2.5}$$

Once the positions are estimated, next part is the force calculation which is one of the most important and laborious steps of an MD simulation. (Figure 2.8) [19]. Total potential energy of a system consisting of N interactive molecules can be described as below;

$$V(r_1, r_2, \ldots, r_N) = \sum_{i<j}^{N} V_2(r_i, r_j) + \sum_{i<j<k}^{N} V_3(r_i, r_j, r_k) + \ldots + \sum_{i<j<k,\ldots,n}^{N} V_n(r_i, \ldots, r_n) \tag{2.6}$$

Here, V_n represents the interaction between n body. The first term of the above equation is the approximation of pairwise additive interactions meaning that the total potential of the system as a result of all the 2-body interactions between particles. Even though this approach usually provides qualitatively accurate results, many body terms other than 2-body interactions should be taken into account for also quantitatively accurate results. However, in practice, mostly only 2-body interactions are considered because of the uncertainty about many body terms' effect on system potential and also practicality of using only the pairwise interactions [20].

Algorithm 5 (Calculation of the Forces)

```
subroutine force(f,en)              determine the force
en=0                                and energy
do i=1,npart
   f(i)=0                           set forces to zero
enddo
do i=1,npart-1
  do j=i+1,npart                    loop over all pairs
     xr=x(i)-x(j)
     xr=xr-box*nint(xr/box)         periodic boundary conditions
     r2=xr**2
     if (r2.lt.rc2) then            test cutoff
       r2i=1/r2
       r6i=r2i**3
       ff=48*r2i*r6i*(r6i-0.5)      Lennard-Jones potential
       f(i)=f(i)+ff*xr              update force
       f(j)=f(j)-ff*xr
       en=en+4*r6i*(r6i-1)-ecut     update energy
     endif
  enddo
enddo
return
end
```

Figure 2.9. Force calculation algorithm

First step of calculating pairwise additive interactions is computing the distance between 2 molecules (i and j) in x, y and z directions and show it as xr. While using periodic boundary conditions, a cutoff distance of r_c is used where r_c value is set to be less than half of the diameter of the hypothetical box that the simulation is run in. In this way we can ensure that when we calculate the interactions between molecules i and j, we make the calculation for i and the closest j among its other recurrent reflections.

In order to calculate the force between i and j, let's call the diameter of the periodic box as D. Considering simple cubic boundary conditions, the maximum length between i and the closest recurrent reflection of j cannot be more than half the diameter (assuming absolute values). Nearest integer function (nint(x)) of FORTRAN (which tells the nearest integer to a real number) can be used in order to calculate the distance between i and the closest recurrent reflection of j. For example the x-distance between these molecules xd can be calculated as =xd-D*nint($\frac{xd}{D}$) . After calculating all Cartesian componens of d_{ij}, value of d_{ij}^2 is calculated instead of $|d_{ij}|$, because calculating the square root takes more effort and it is not needed for this case. Once the value of d_{ij}^2 is known, it should be compared to the value of d_c^2 to see if it is less. In case it is more, another value of j should be considered until a smaller than d_c^2 value is found.

Force calculation between two molecules that are in interaction and their addition to the total potential energy should be considered and calculated for all components. For example, the x component can be calculated as follows;

$$f_x(d) = -\frac{\partial u(d)}{\partial x}$$

$$= -\left(\frac{x}{d}\right)\left(\frac{\partial u(d)}{\partial d}\right) \tag{2.7}$$

According to Lennard-Jones potential, this can also be described as following;

$$f_x(d) = \frac{48}{d^2}\left(\frac{1}{d^{12}} - 0.5\frac{1}{d^6}\right) \tag{2.8}$$

2.2.2. Equations of Motions Calculations

Once force calculations are completed for all components, Newton's equations of motion can be integrated for the system by using different algorithms. The algorithm used in this case is the Verlet algorithm which is one of the most useful and simple ones. (Figure 2.10)

Algorithm 6 (Integrating the Equations of Motion)

```
subroutine integrate(f,en)          integrate equations of motion
sumv=0
sumv2=0
do i=1,npart                        MD loop
   xx=2*x(i)-xm(i)+delt**2*f(i)     Verlet algorithm (4.2.3)
   vi=(xx-xm(i))/(2*delt)           velocity (4.2.4)
   sumv=sumv+vi                     velocity center of mass
   sumv2=sumv2+vi**2                total kinetic energy
   xm(i)=x(i)                       update positions previous time
   x(i)=xx                          update positions current time
enddo
temp=sumv2/(3*npart)                instantaneous temperature
etot=(en+0.5*sumv2)/npart           total energy per particle
return
end
```

Figure 2.10. Equations of motion integration algorithm

Derivation of the algorithm begins with Taylor expansion calculation of a molecule's coordinate at a time t. For $t + \Delta t$ we have;

$$r(t+\Delta t)=r(t)+v(t)\Delta(t)+\frac{f(t)}{2m}\Delta t^2+\frac{\Delta t^3}{3!}\dddot{r}+\sigma\left(\Delta t^4\right) \tag{2.9}$$

For $t - \Delta t$ we have;

$$r(t-\Delta t)=r(t)-v(t)\Delta(t)+\frac{f(t)}{2m}\Delta t^2-\frac{\Delta t^3}{3!}\dddot{r}+\sigma\left(\Delta t^4\right) \tag{2.10}$$

Adding 2.3.9 to 2.3.10 results as following;

$$r(t+\Delta t)+ r(t-\Delta t)=2r(t)+\frac{f(t)}{m}\Delta t^2+\sigma\left(\Delta t^4\right) \tag{2.11}$$

$$r(t+\Delta t)\approx 2r(t)-r(t-\Delta t)+\frac{f(t)}{m}\Delta t^2 \tag{2.12}$$

Considering that Δt is the time step in this MD simulation, eliminating $\sigma(\Delta t^4)$ causes the result to have an error amount of Δt^4 but this acceptable neglection simplifies the equation.

As seen from the equation 2.3.12, Verlet algorithm does not include velocity for position calculation. But velocity can still be obtained by using the trajectory information and by arranging the equation as follows;

$$r(t+\Delta t)- r(t-\Delta t)= 2v(t)\Delta(t)+\sigma\left(\Delta t^3\right) \tag{2.13}$$

$$v(t)=\frac{r(t+\Delta t)- r(t-\Delta t)}{2\Delta(t)}+\sigma\left(\Delta t^2\right) \tag{2.14}$$

Velocity definition in 2.3.14 is correct for only order of Δt^2. In order to evaluate better velocity estimations a Verlet like algorithm can be obtained for velocity.

Using Verlet algorithm for Equations of Motion calculations is advantageous in terms of calculation speed. However, shorter time steps must be defined in order to get accurate results. In other words, force calculation for all constituents of the system must be done frequently. Verlet algorithm requires very little memory to work with which is a plus while working with extensive systems and it shows only a small amount of diversion for the value of long-term energy due to being time reversible.

2.2.3. MD Protocol

Molecular Dynamics (MD) simulations consist of three main steps. First step is setup which starts with parametrization of the substrate. Antechamber is used to extract partial charges from Gaussian output. Parmchk is used to produce the parameters and check them and finally to produce the substrate pdb.

antechamber -i FILENAME.out -fi gout -o FILENAME.prepin -fo prepi -c resp

After this step, substrate file is obtained inside the prepin file and it is manually renamed as MAC for s-trans and MSC for s-cis conformations.

parmchk -i FILENAME.prepin -f prepi -o FILENAME.frcmod

With this step, parameter file, frcmod, was created and checked to make sure it contains parameters for a particular bond, angle or dihedral. If not, file should be manually edited.

antechamber -i FILENAME.prepin -fi prepi -o FILENAME.pdb -fo pdb

Finally substrate pdb file is created as a result of this third step.

As the second step of the setup the local directory is prepared and prepin, frcmod, pdb files that were initially created are copied to this directory where the MD simulation will be run.

As the third step of setup, protein file is prepared for MD by removing waters and hydrogens from it. All the water removed from protein is saved as another object in a visualization program and hydrogens are also removed from it. Next step is to merge protein with the substrate. Using a text editor, first protein constituents are listed then ligand constituents are added to the file and finally all the water molecules are added and these three items are separated with TER then END is added to the end of the file. Some manual alterations are done for protonation state of the residues of vital importance such as HID residues were converted to HIE to make sure epsilon-N is protonated and Cysteine residue is converted to CYX from CYM for disulfide bridge formation. As last step of the setup, system is neutralized and tleap file is run in order to generate prmtop and inpcrd files with the following line;

tleap -f FILENAME.tleap

Water was used as an explicit solvent and the protein-ligand-water complex was covered with a water box in a 10 Å radius.

After setup is completed, next step is the minimization of the system in order to make sure that the initial configuration does not contain any high energy interactions that could cause instabilities in the simulation.

Once minimization is done, next step is heating of the system to room temperature, from 0 K to 300 K. Heating is divided into six steps with increments of 50 K instead of going to 300 K from 0 K directly. After heating is completed, equilibration step is done and MD production run is finally started. MD simulations in this study are run for 400 ns with a data collection rate of 100 ps. Once an MD run is completed, result files are generated to analyze the molecular interactions along the simulation with data of distance vs time, distance vs angle, residue fluctuations, pdb snapshots along the run etc.

2.3. FORCE FIELDS

Biological molecules can be analyzed effectively by using computer simulations tools. Theoretically, all features of a molecular design and molecules' interactions with its surroundings can be envisioned by using quantum molecular dynamics calculations. However, it is not the most preferable tool due to its remerkably high computational costs and that is why we need to reduce computational calculations to a more simplified pattern in order to analyze the design of biological molecules [21].

Force field method also known as molecular mechanics method provides a simplification enough to analyze large molecular structures practically [22]. Force field explains the relevance of a system's energy to its particles' coordinates mathematically, which is an indicator of the molecular geometries in equilibrium also of the corresponding energies among different conformers or among various molecules [23]. It is defined by the sum of interatomic potential energies combined with specific parameters obtained from ab initio or semi-empirical QM calculations or by using experimental data such as NMR, electron diffraction, X-ray, Raman or infra-red spectroscopy [22]. Force field methods are all based on the Born-Oppenheimer approximation. This approximation divides nuclear and electronic motions in a molecule and basically, neglects the degrees of freedom of electrons in a

molecule and only considers the nucleic motions. Therefore it allows us to define the system's energy only by considering nuclear coordinates. Other than the Born-Oppenheimer approximation, most force fields include two more assumptions which are additivity and transferability. Additivity proposes that a system's potential energy can be calculated by adding all of the system's potential energies which contain a simple physical correspondance such as bond stretching, bending, electrostatics etc. Transferability explains that the smaller and wider versions of the similar chemical groups of molecules can use the potential energy expressions. The most common potential energies with a physical correspondance mentioned in additivity are bond or angle deformations such as stretching or bending, torsional movements. Angle and torsional terms are much less rigid than bond stretchings and they maintain the molecules rigidity at the right level, Electrostatic (Coulomb) and dispersion (Van Der Waals) interactions are included in the non-bonded terms. Polarizability of atoms and cross coupling are some of the forces that more complicated force fields include.

Force fields are empirical expressions therefore dividing the system's potential energy into terms [21]. Simply, Force Field (FF) represents the potential energy that holds the atoms of a defined system together by rythmic forces. The most common mathematical expression of a force field is defined as follows [17];

$$U=\sum_{\text{bonds}}\frac{1}{2}k_b(r-r_0)^2+\sum_{\text{angles}}\frac{1}{2}k_a(\theta-\theta_0)^2+$$

$$\sum_{\text{torsions}}\frac{V_n}{2}[1+\cos(n\varphi-\delta)]+\sum_{\text{improper}}V_{imp}+\sum_{LJ}4\epsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}}-\frac{\sigma_{ij}^{6}}{r_{ij}^{6}}\right)+\sum_{\text{elec}}\frac{q_iq_j}{r_{ij}} \qquad (2.15)$$

From the first to four expressions represent the energy contributions either locally or intermolecularly such as bending and streching motions. The last two expressions describe the Van der Waals and electrostatic interactions [22].

The origin of force fields go back to 1930's when D.H.Andrews first introduced concept of spectroscopic force field into biomolecular mechanics in *Phys. Rev*. In 1940, F.H.Westheimer achieved the first and only manual molecular mechanics calculation in order to find out a tetra-substituted biphenyl's transition state. Despite these pioneer developments, first major developments were achieved in 1960's when computers became accessible with the elementary goal of forecasting molecular arrangements, vibrational

spectra and enthalpies of single molecules. J.B. Hendrickson achieved conformational examination of rings which have more than six components in 1961. K.B. Wiberg issued the first program which is able to find the minimum energy wing molecular mechanics in 1965. The most important work on force fields of biomolecules were achieved by N.L.Allinger, H.Scheraga and S.Lifson. Most of the molecular mechanics potentials established by this group were mainly focused on small biomolecules but some of those could be applied more generally and these potentials such as MM2, MM3 and MM4 can even be used today These force fields were first constructed to analyze hydrocarbons but in time they extended to cover ethers, amides, alcohols etc [23].

Recently force fields have been developed to be applied to much more complicated systems. For instance Dreiding and Universal force fields enable the parametrization of all the atoms in periodic table. Some of the most popular force fields such as AMBER, GROMACS and CHARMM are mostly applied to molecular dynamics simulations of biomolecules while COMPASS and OPLS force fields are applied more to the simulations of condensed matter physics. These force fields are being developed constantly, so more recent versions may be used in simulations in order to get advanced results [22].

Polarizable force fields have been developed in the late 20th century. Some of the most popular polarizable force fields include PIPF which means polarizable intermolecular potential function, DRF90 and AMOEBA. Some of the regular force field models mentioned earlier such as AMBER, GROMOS, OPLS or CHARMM also have polarizable forms [24].

Numerous specific potentials have been formed other than force fields, in order to represent only a specific system or a combination of compounds. Water is one of the most important specific potential models and even more models are being proposed since the first Monte Carlo simulation performed by Barker and Watts. Some of the most popular water models include TIP3P, TIP4P, TIP5P, SPC and SPC/E). According to a current study carried out by Vega et al. TIP4P/2005 provides the best results for 90 percent of the experimental characteristics examined among the most widely used rigorous water potentials that are non polarizable [24].

Comparing the efficiencies of the popular force fields is a quite challenging task since the outcome of the simulation performed depends mainly on the constituents of the structure and their features. Basically, every force field provides the best result for a specific system and

the parametrization applied. Nevertheless, all commonly used force fields present acceptable outcomes for a broad scope of characteristics of single molecules, aqueous solutions and pure liquids as expressed by Jorgensen and Tirado-Rives [24]. However there are some comparisons regarding commonly used force field models in the literature. For example, Price and Brooks [25] stated that using AMBER, CHARMM or OPLS in the framework of simulations performed for biomolecules yielded resembling outcomes regarding the arrangement and dynamics of three types of protein molecules. In another study, Yeh and Hammer discovered [26] that there was an important amount of distinctness in the arrangements of two peptide molecules whose simulations were performed by AMBER and CHARMM.

Configurations of small sized peptides with open chains were also found to be reliant to the type of force field used by Aliev and Courtier-Murias [27]. A conclusion resembling to this was obtained in the molecular simulation of insulin performed with AMBER, OPLS, CHARMM and GROMOS where various conformations were obtained in each simulation. One of the most extensive researches on the mostly used force fields was conducted by Paton and Goodman [28] where they have used seven types of force field models (AMBER, OPLS, OPLSAA, MM2, MM3, MMFF94, MMFF94s) in order to investigate the interaction energies of 22 molecular compounds of various sizes and nuclear acid bases and amino acids with a total number of 143, then used these as a reference point to all the energy values acquired from high level *ab initio* calculations. In the end, they have reached to the conclusion entirety of potentials examined and most importantly MMFF94s and OPLSAA, demonstrated fairly correct interactions of van der Waals and electrostatic forces [22].

Table 2.2. List of the most commonly used force fields along with their target systems

Name	**Target**
AMBER	biomolecules, organics
CHARMM	biomolecules
CHARMm	biomolecules, organics
CFF/CVFF	organics, biomolecules
DREIDING	main group organics, inorganics
ECEPP	proteins
ESFF	general
GROMOS	biomolecules
MM2	organics
MM3	organics, biomolecules

MM4	hydrocarbons
MMFF	organics, biomolecules
MOMEC	transition metal compounds
OPLS	biomolecules, organics
SHAPES	transition metal compounds
SYBYL/Tripos	proteins, organics
UFF	general
VALBOND	transition metal compounds

2.4. AMBER MD PACKAGE

AmberTools is a cumulative name for a collection of computer tools which allows conducting MD simulations and analysis especially on biomolecules. The simulation subroutines used with Amber include sander, nab, mdgx and pmemd. The workflow followed when performing a simulation using Amber is shown in Figure 2.10. First of all the cartesian coordinates of every atom in the system must be listed. They can be obtained experimentally from X-ray crystallography, NMR spectroscopy (from protein data bank form as pdb files) and model building. In order to conduct these modelings, LEaP can be used as well as other modeling programs. LEaP allows us to prepare the coordinate, parameter abd topology files to start MD simulations [29].

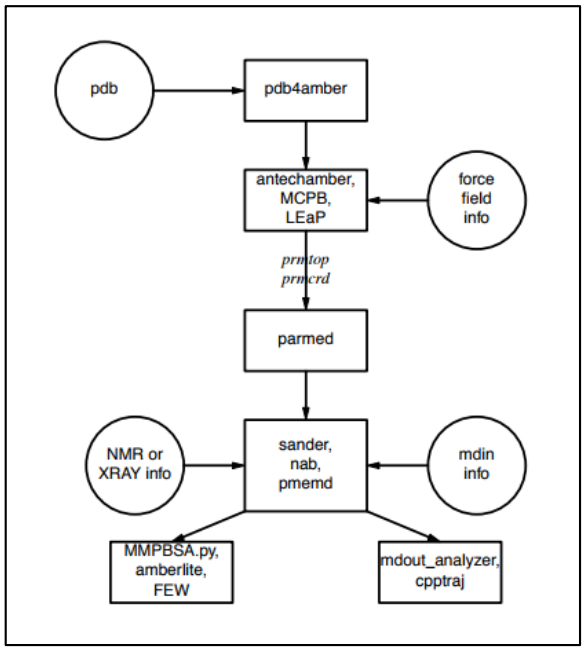


Figure 2.11. MD simulation flowchart by Amber

3. RESULTS AND DISCUSSION

MBH is a bimolecular reaction between an activated alkene and an aldehyde commonly catalyzed by amines or phosphines. Even though the reaction mechanism is still a point of controversy [30, 31, 32], four major steps are known to be involved in the catalytic cycle (Figure 3.1.).

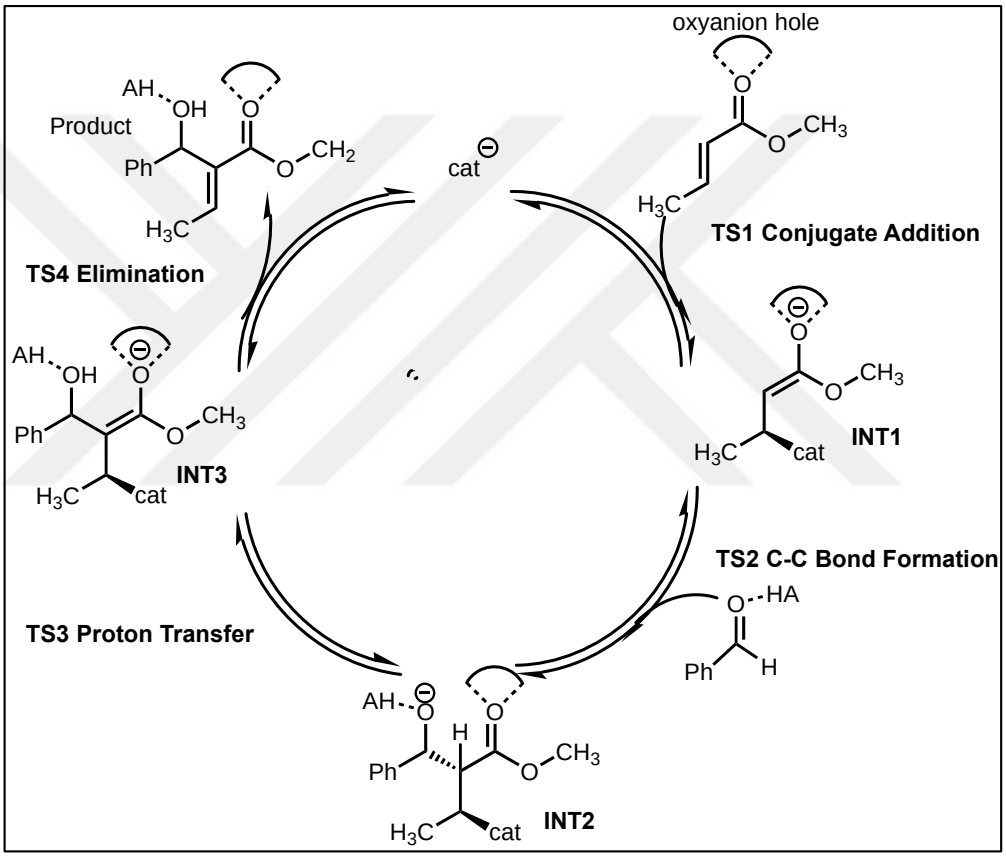


Figure 3.1. Mechanism of the MBH Reaction

Michael addition of a nucleophilic catalyst to the alkene (**TS1**) initiates the reaction, forming an enolate intermediate (**INT1**), which in turn attacks the electrophilic carbon of an aldehyde (**TS2**). This C-C bond formation step is shown to be the rate determining step in the presence of protic solvents or when autocatalysis become dominant at the later stages of the reaction [33, 34]. This step gives the second intermediate (**INT2**). The following proton transfer step (**TS3**) is considered to be rate determining in aprotic solvents [35, 36]. Presence of protic solvents significantly increased the rate of the reaction by decreasing the activation energy

of the proton transfer state [37, 38]. Computations suggest a six-membered TS involving a proton shuffle mechanism for TS3. The elimination step (**TS4**) release the products and regenerate the nucleophilic catalyst.

Recently, Ütnier [39] explored the catalytic effect of different amino acids on the MBH reaction by the help of theozyme models constructed as shown in Figure 3.2.

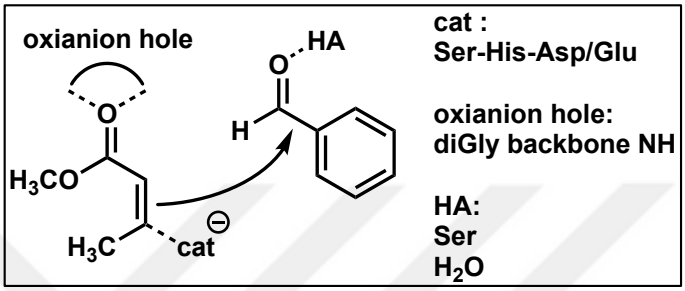


Figure 3.2. Theozymes constructed for the MBH Reaction [39]

Calculations suggest that an acid-base co-catalyst decreases the activation barrier by 30-35 kcal/mol and rate determining step is proton transfer for all theozyme models as seen in Figure 3.3 [39].

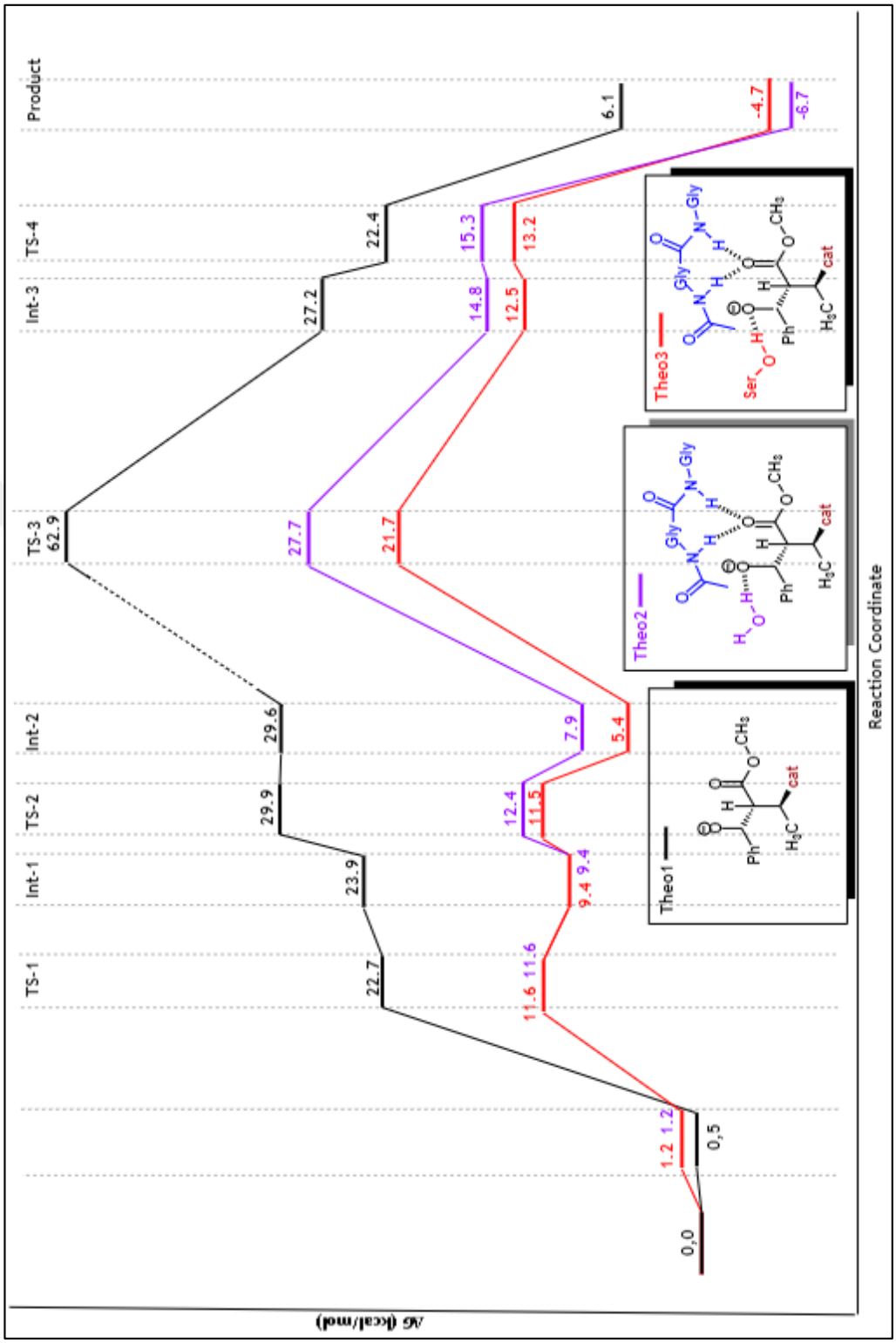


Figure 3.3. Free energy profiles of Theozyme1 (**Theo1**, black), theozyme 2 involving oxyanion hole and water as acid-base co-catalyst (**Theo2**, purple) theozyme 3 involving oxyanion hole and methanol as acid-base co-catalyst (**Theo3**, red) [42]

In this thesis, two of the theozyme models (**Theo2** and **Theo3**) were selected as templates for the screening of the protein databank to identify similar active sites occurring in natural proteins. Both of the theozymes involved Ser_His_Asp/Glu catalytic triad as the nucleophile, and diglycine backbone NH groups as the oxyanion hole motif. The difference is the acid-base co-catalyst, which is represented by a water molecule in **Theo2** and by methanol in **Theo3**. For **Theo3**, however, an additional model was generated by changing the conformation of the substrate from s-trans to s-cis. Although the catalytic residues remain the same, this conformational change altered their arrangement in space with respect to each other.

3.1. SCREENING OF THE PROTEIN DATABANK

Theozymes were first incorporated in SABER analysis to find similar protein structures that could be analyzed as potential enzymes.

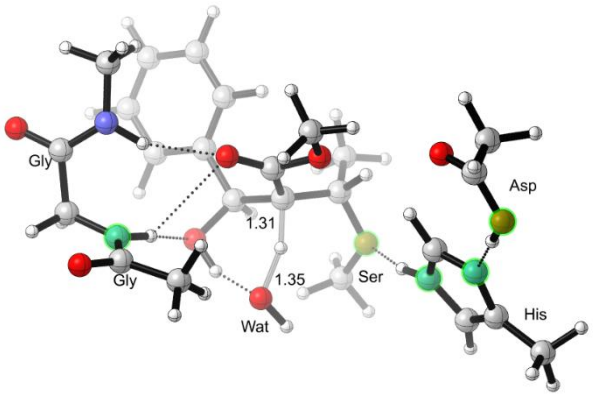
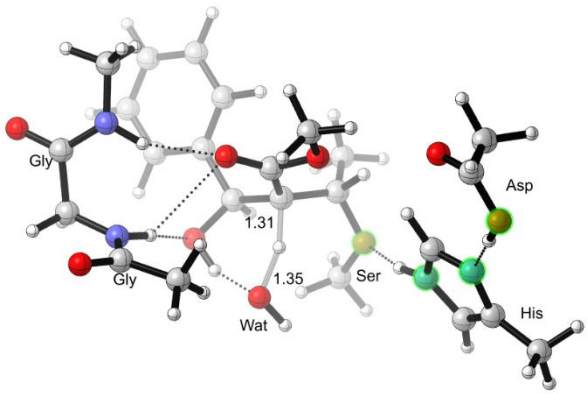
The proton transfer transition state (**TS3**) was chosen to generate CAMs as it was predicted to be the rate determining step according to the energy profile given in Figure 3.3. Different CAMs were constructed by changing the number and position of the atoms selected. Increasing the number of atoms limited the search results since it is more challenging to find the similar atom sequence for 6 atoms rather than 4 atoms.

Three different CAMs were generated based on **Theo2**:

- **Theo2**: Proton Transfer TS (**TS3**); Ser-His-Asp/Glu as nucleophilic triad ; diglycine backbone NHs as oxyanion hole; H_2O as acid-base co-catalyst; substrate in s-trans conformation.
    - **CAM1**: 5 atoms (OD2 (ASP), NE2 (HIS), ND1 (HIS), OG (SER), N (GLY) Highlighted in Table 3.1(a))
    - **CAM2**: 4 atoms (OD2 (ASP), NE2 (HIS), ND1 (HIS), OG (SER) Highlighted in Table 3.1(b))
    - **CAM3**: 6 atoms (OD2 (ASP), NE2 (HIS), ND1 (HIS), OG (SER), N (GLY), N (GLY) Highlighted in Table 3.1(c))

Two different CAMs were generated based on **Theo3**:

- **Theo3**: Proton Transfer TS (**TS3**); Ser-His-Asp/Glu as nucleophilic triad ; diglycine backbone NHs as oxyanion hole; MeOH (SER) as acid-base co-catalyst; substrate in s-trans conformation.
    - **CAM4**: 5 atoms - (OD2 (ASP), NE2 (HIS), ND1 (HIS), OG (SER), N (GLY)) (Highlighted in Table 3.2(a))
- **Theo3**: Proton Transfer TS (**TS3**); Ser-His-Asp/Glu as nucleophilic triad ; diglycine backbone NHs as oxyanion hole; MeOH (SER) as acid-base co-catalyst; substrate in s-cis conformation.
    - **CAM5**: 5 atoms - (OD2 (ASP), NE2 (HIS), ND1 (HIS), OG (SER), N (GLY)) (Highlighted in Table 3.2(b))

(a)	
(b)	

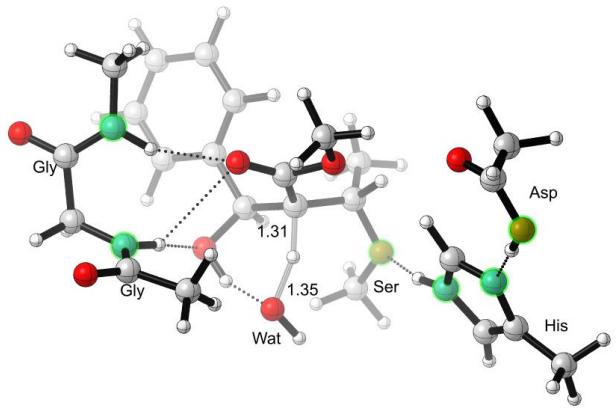
(c)	

Figure 3.4. Atoms selected for (a) **CAM1** of **Theo2** (b) for **CAM2** of **Theo2** (c) for **CAM3** of **Theo2**

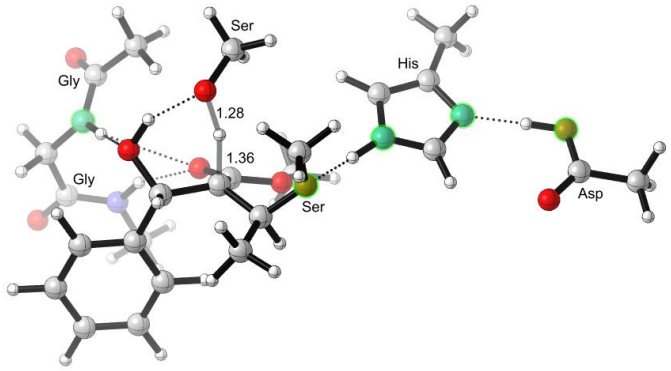
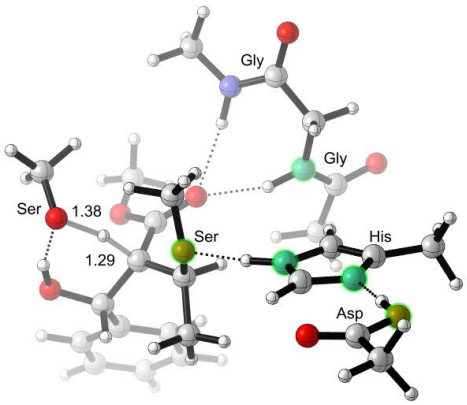
(a)	
(b)	

Figure 3.5. (a) Atoms selected for CAM4 of Theo3 (b) Atoms selected for CAM5 of Theo3

Completion of SABER run took a few days depending on the magnitude of delta and the number of the atoms selected. Increasing delta value means a larger area will be scanned to find matching proteins therefore larger amount of results were obtained but it took longer to get the results. Also having more results did not increase the number of good matches. Instead, it resulted in an increased computational cost as well as difficulties in analyzing a huge set of data.

Selecting more atoms for the **Theo2 CAM1** than **Theo2 CAM2** decreased the running length of the job as well as the number of matches found. In addition, these results were more specific and easier to eliminate. Some of the good matches of **Theo2 CAM1** can be seen in Table 3.3. These proteins were selected since they have low RMSD values and known catalytic mechanisms. These were pair fit with the original theozyme and once the closest RMSD values were achieved, that match was used to perform MD simulations. Figure 3.4 and 3.5 show the overlays of active site of proteins 2IXT (green structure) and 2SEC (pink structure) with **Theo2**. An RMSD value of 0.293 for 2IXT and 0.304 for 2SEC, which are very close to the original values.

Table 3.1. SABER matches for **Theo2 CAM1**

PDB ID	**Name of the Protein**	**RMSD (Å)**	**Matching Residue**
2B61	Homoserine o-acetyltransferase	0.22	ASP A 304 HIS A 337 SER A 143 GLY A 142
1E5T	Prolyl endopeptidase	0.27	ASP A 641 HIS A 680 SER A 554 GLY A 553
2IXT	36kda protease	0.29	ASP A 34 HIS A 71

			SER A 250 GLY A 141
2Z3W	Dipeptidyl aminopeptidase iv	0.30	ASP A 678 HIS A 710 SER A 603 GLY A 606
2SEC	Subtilisin carlsberg	0.30	ASP E 32 HIS E 64 SER E 221 GLY E 127
2NW6	Lipase	0.7	ASP A 264 HIS A 286 SER A 87 GLY A 90
2BDB	Elastase-1	0.88	ASP A 102 HIS A 57 SER A 195 GLN A 192
1FFE	Cutinase	0.55	ASP A 175 HIS A 188 SER A 120 ALA A 123

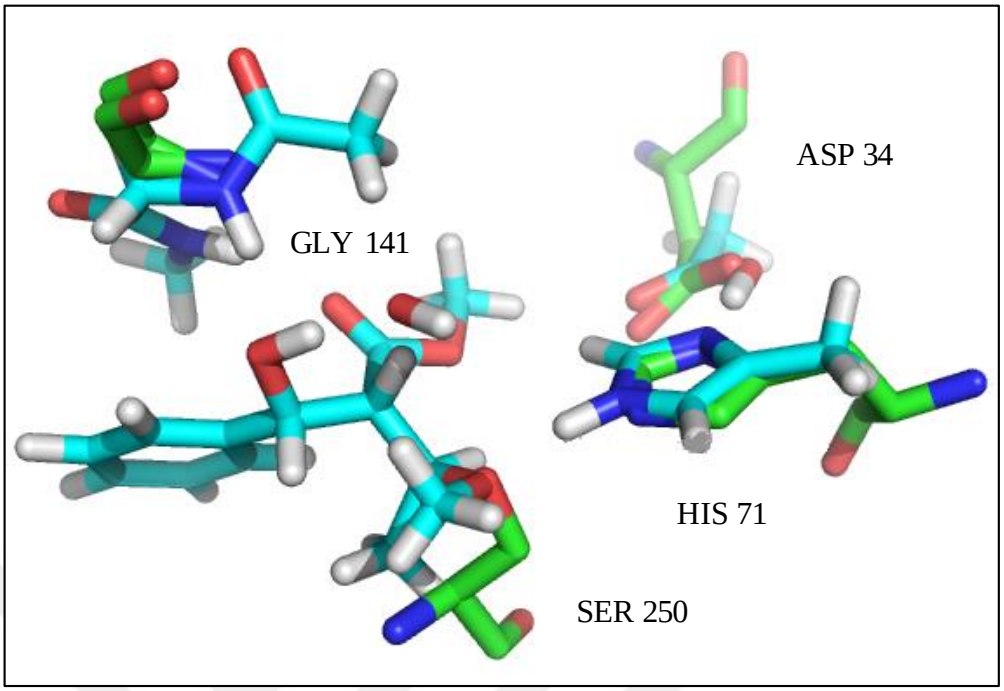


Figure 3.6. Overlay of 2IXT (green) with **Theo2 CAM1** (Blue)

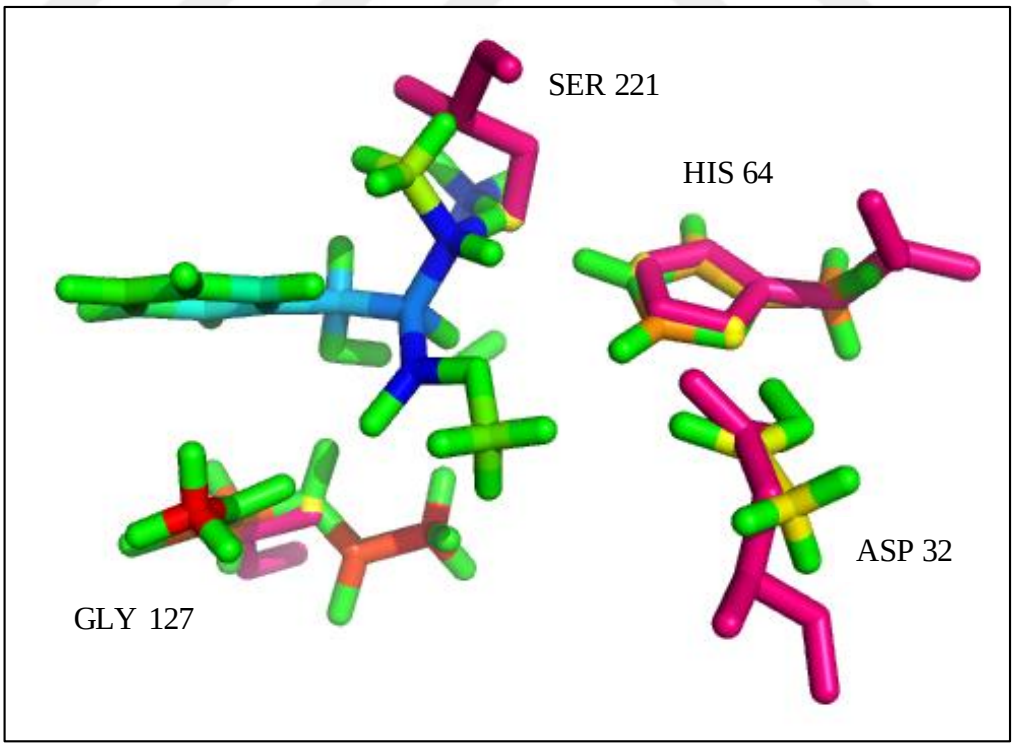


Figure 3.7. Overlay of 2SEC (magenta) with **Theo3 CAM1** (green)

SABER for **Theo2 CAM2** was also run and results were analyzed. Some of the good protein matches for this search are listed in Table 3.4. These proteins were again chosen considering their low RMSD values and catalytic properties. Overlay of matching residues in native protein and the theozyme for 1QNP and 2NW6 are shown in Figures 3.6 and 3.7. RMSD values of 0.136 and 0.152 were obtained respectively for 1QNP and 2NW6.

Table 3.2. Protein matches for **Theo2 CAM2**

PDB ID	Name of the Protein	RMSD (Å)	Matching Residue
2NW6	Lipase	0.15	SERA87 ASPA264 HISA286
2BDB	Elastase-1	0.05	HISA57 ASPA102 SERA195
1FFE	Cutinase	0.19	ALAA42 SERA120 GLNA121 ASPA175 HISA188

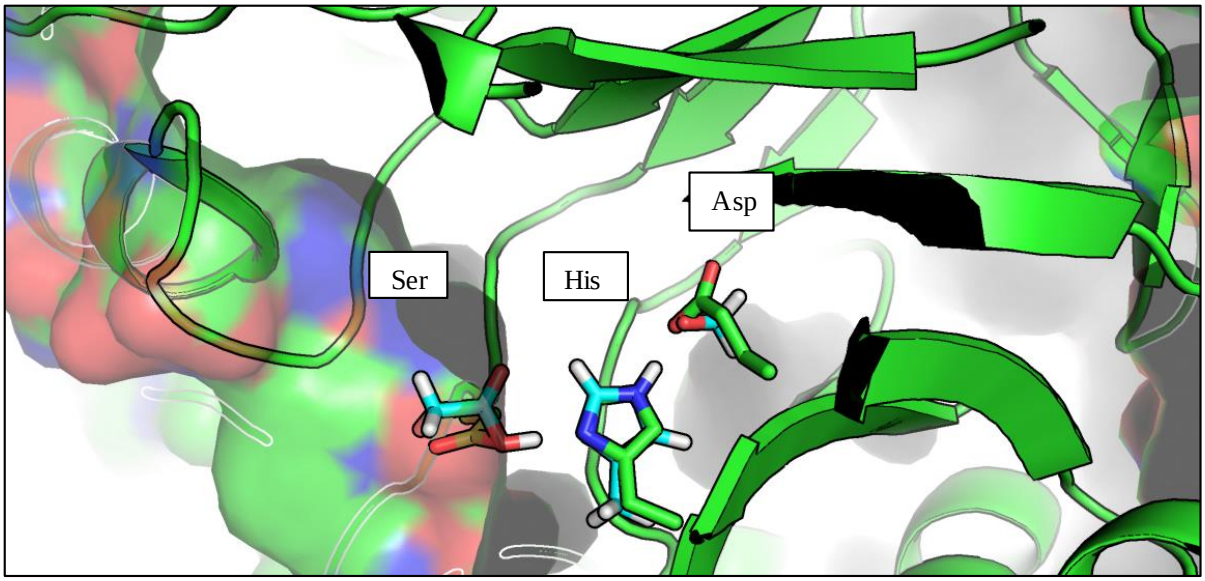


Figure 3.8. Overlay of 1QNP for **Theo2 CAM2**

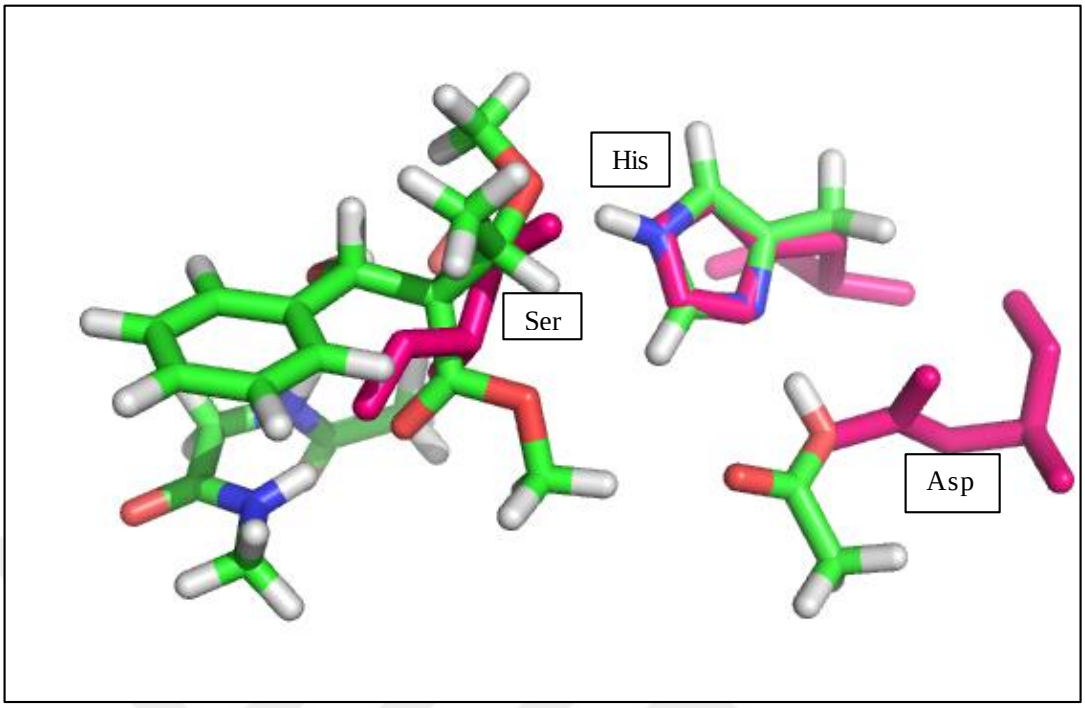


Figure 3.9. Overlay of 2NW6 for **Theo2 CAM2**

No results could be obtained from SABER run for **Theo2 CAM3** because selected number of atoms (6 atoms) limited the results. Finding matches for 6 functional groups at specific locations rather than 4-5 atoms was proved to be more difficult.

Some of the good protein matches for **Theo3 CAM4** are listed in Table 3.4.

Table 3.3. Protein matches for **Theo3 CAM4**

PDB ID	**Name of the Protein**	**RMSD (Å)**	**Matching Residue**
2NW6	Lipase	0.53	ASP A 264 HIS A 286 SER A 87 GLY A 89

2BDB	Elastase-1	0.88	ASP A 102 HIS A 57 SER A 195 GLY A 193
1FFE	Cutinase	0.5	ASP A 175 HIS A 188 SER A 120 GLY A 122

Lastly, **Theo3 CAM5** confirmation SABER was run for 5 atoms. Some of the good matches can be seen in table 3.6.

Table 3.4. **Theo3 CAM5** SABER matches

PDB ID	Name of the Protein	RMSD (Å)	Matching Residue
2D0D	Crystal structure of a meta-cleavage product hydrolase	0.92	ASPA224 HISA252 SERA103 GLYA33
1HMU	Chondroitinase ac	1.05	GLUA371 HISA225 TYRA234 ASNA175
2W2B	P-coumaric acid decarboxylase	0.33	GLUB27 HISB15 TYRB173 ALAB143
1K7C	Rhamnogalacturonan acetylesterase	0.52	ASPA192 HISA195 SERA9 GLYA42
2NW6	Lipase	0.92	GLUA289 HISA86

			TYRA29 GLYA16
2BDB	Elastase-1	0.51	ASPA102 HISA57 SERA195 ALAA1002
1FFE	Cutinase	0.51	ASPA175 HISA188 SERA120 ASNA84

3.2. MOLECULAR DYNAMICS ANALYSIS

Once the output of Saber analysis for all three models were obtained, next step was to decide which proteins to focus on for further analysis by using MD simulations. Options were already narrowed down by selecting particular proteins from each SABER result. This time an overall filtering was applied considering all selected proteins. As a result of this filtering following three proteins were found in common in the SABER results of all three theozyme structures as good matches which are 2NW6, 2BDB and 1FFE. All three of them have RMSD values less than one in all screening results and they are fairly distinctive (Table 3.7) protein structures therefore they were used in MD simulations for further analysis. MD analysis for these 3 proteins were done for three variations as below;

1) apo MD: only protein
2) s-cis MD: protein + s-cis conformation of ligand (MSC)
3) s-trans MD: protein + s-trans conformation of ligand (MAC)

Table 3.5. Characteristics of MD proteins

Protein	Class	Fold	Superfamily	Family
2BDB	All beta proteins	Trypsin-like serine proteases	Trypsin-like serine proteases	Eukaryotic proteases
2NW6	Alpha and beta proteins (a/b)	alpha/beta-Hydrolases	alpha/beta-Hydrolases	Bacterial lipase
1FFE	Alpha and beta proteins (a/b)	alpha/beta-Hydrolases	alpha/beta-Hydrolases	Cutinase-like

3.2.1. 2BDB – MD Analysis

2BDB - Porcine pancreatic elastase complexed with Ala-Ala and Asn-Pro-Ile residues is a hydrolase molecule. It includes 240 residues.

The interactions between catalytic triad, ligand and the other significant residues during this simulation are investigated via distance vs time and distance vs angle plots.

Fluctuation of distance with time and angle between O1 atom of s-cis ligand-MSC 242 and H atom of GLY186 seems to be disordered and not stable which indicates that H-bond could not be maintained between these molecules so they drift away from each other easily.

Figure 3.8 shows that the two molecules were very close to each other at the initial stages of MD but around 50 ns, they started drifting away from each other and gathering at a distance of 3 Å. This behavior shows that substrate (MSC) was constantly moving loosely in the active site but also keeping in touch with backbone NH atom of GLY since they are always at a certain distance.

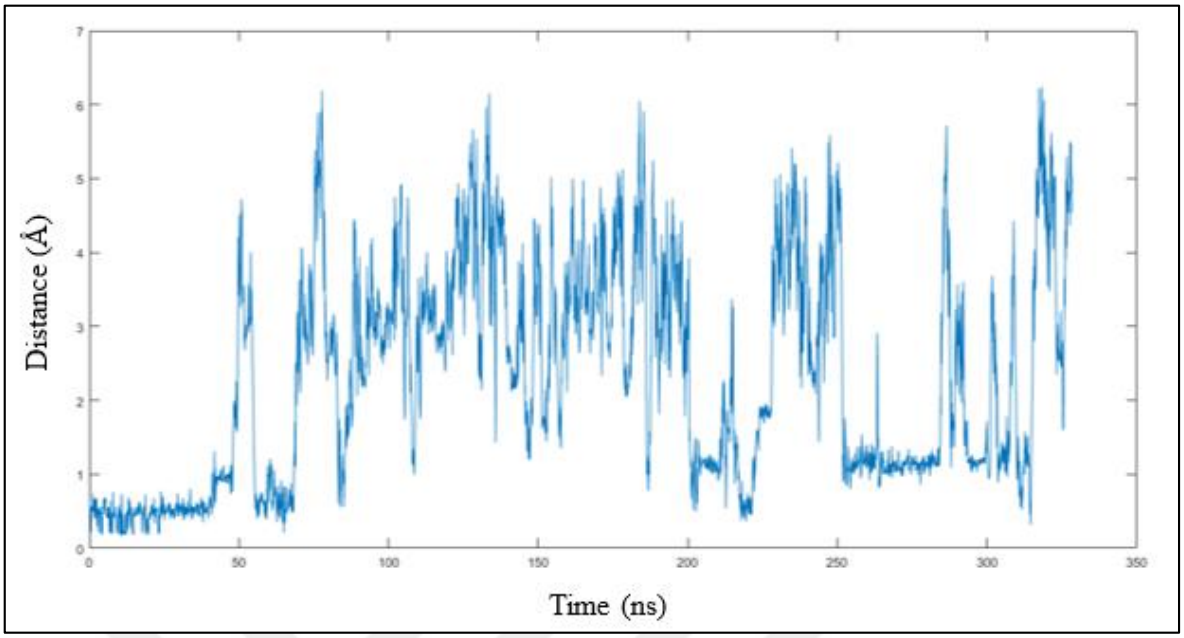


Figure 3.10. Distance vs time plot of MSC242 O1 – GLY186 H (2BDB s-cis MD)

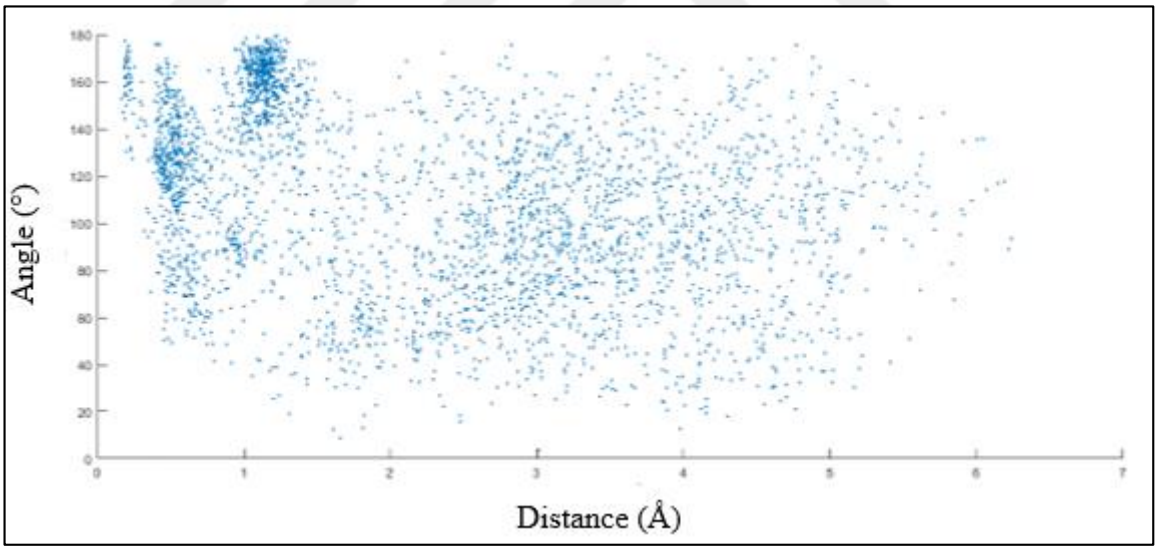


Figure 3.11. Distance vs angle plot of MSC242 O1 – GLY186 H – GLY186 N (2BDB s-cis MD)

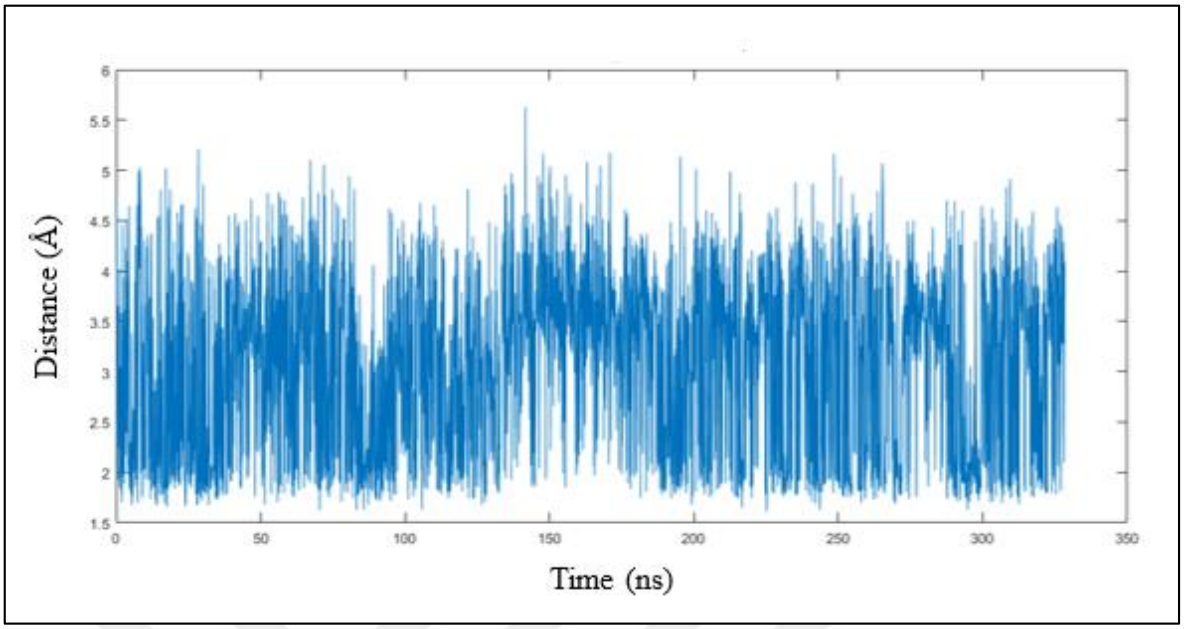


Figure 3.12. Distance vs time plot of HID45 NE2 – SER188 HG (2BDB s-cis MD)

Interactions between HID45 NE2 – SER188 HG are one of the most important molecule interactions because they construct the hydrogen bonding in ASP-HIS-SER catalytic triad. Figure 3.15 shows the distance to fluctuate between 2.0 Å – 4.5 Å which shows the H-bond was kept during the simulation time. The distance vs angle plot between these residues shows that the interactions are mainly focused at around 2 Å distance - 160°.

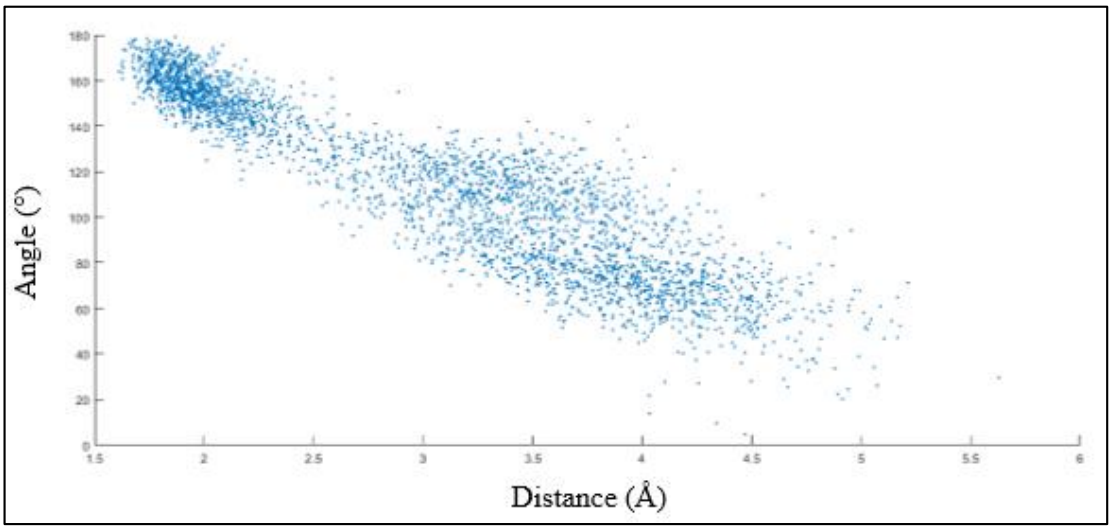


Figure 3.13. Distance vs angle plot of HID45 NE2 – SER188 HG – SER188 OG (2BDB s-cis MD)

MD simulation was also run for 2BDB-apo option, where ligand was not included in the process and only the protein was run. Similar results were obtained for the catalytic triad. Some of the most important reactions between catalytic residues can be seen in Figure 3.12 and Figure 3.13 where ASP and HIE residues made a strong hydrogen bond as expected, so the distance between them is mostly around 2 Å without much diversion.

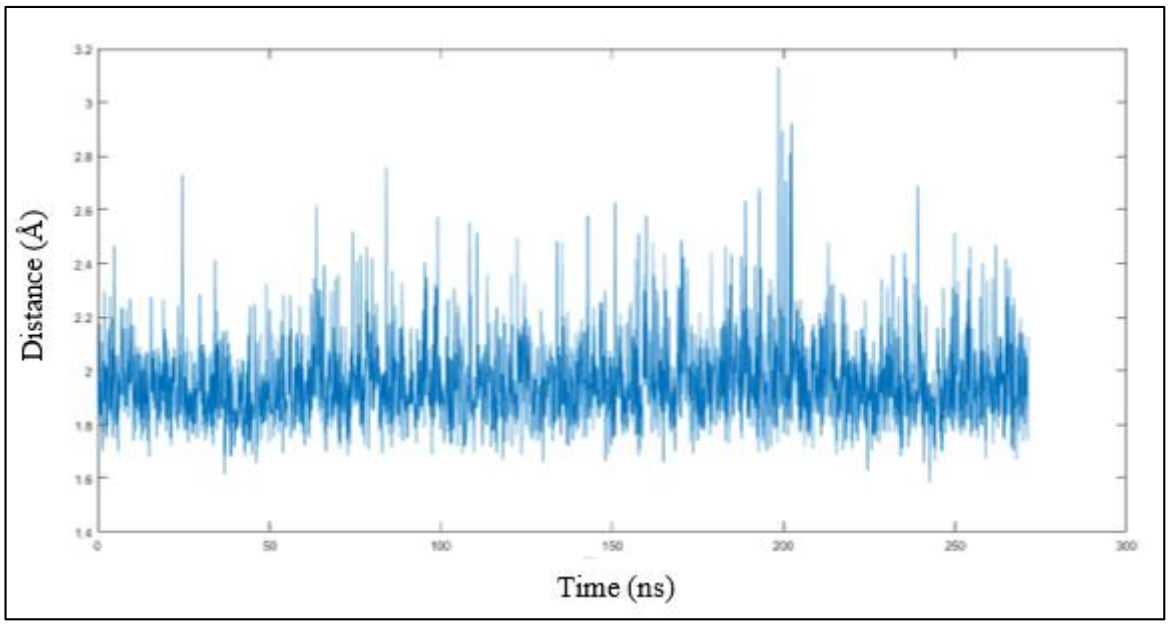


Figure 3.14. Distance vs time plot of ASP102 OE1 – HIE57 H (2BDB apo MD)

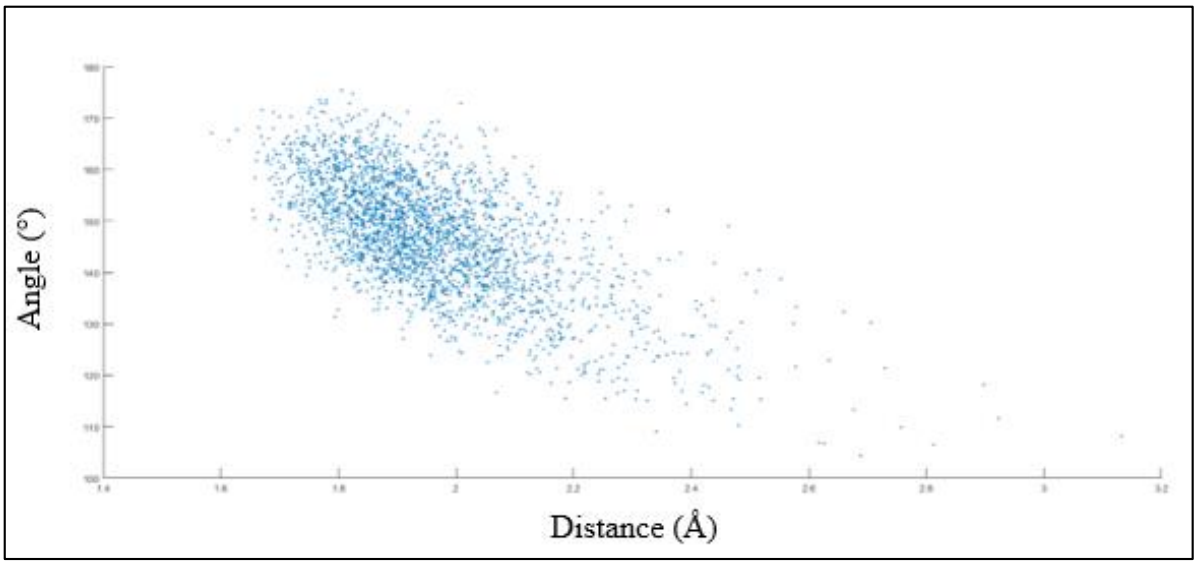


Figure 3.15. Distance vs angle plot of ASP102 OE1 – HIE57 H (2BDB apo MD)

Other part of the catalytic triad which is the HID-SER interaction shows more fluctuation then ASP-HIS interaction however the distance is still mainly stable around 2 Å as expected as seen in Figure 3.14 and Figure 3.15.

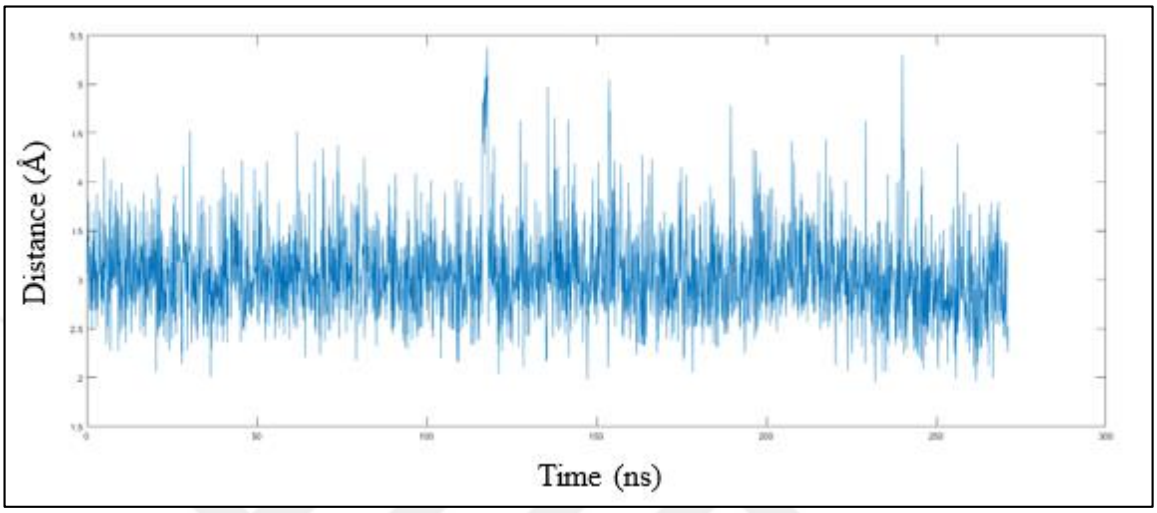


Figure 3.16. Distance vs time plot of HID57 NE2 – SER195 HG (2BDB apo MD)

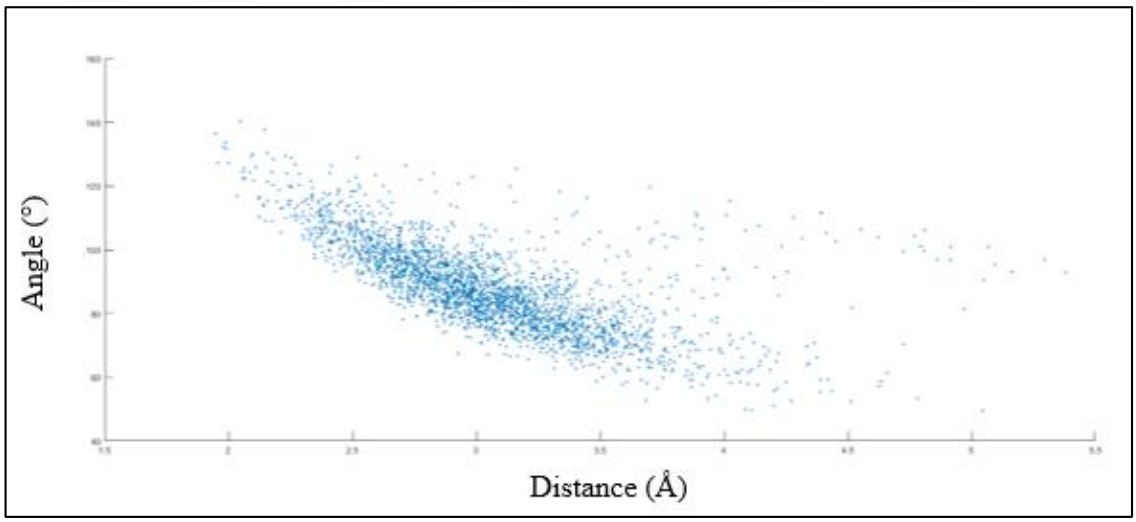


Figure 3.17. Distance vs angle plot of HID57 NE2 – SER195 HG (2BDB apo MD)

3.2.2. 2NW6 – MD Analysis

2NW6 - Burkholderia cepacia lipase complexed with S-inhibitor is a hydrolase molecule that also includes a hydrolase inhibitor. It has one polypeptide chain with 320 residues. A

better match was obtained when theozyme was fit to the protein's natural substrate rather than matching residues.

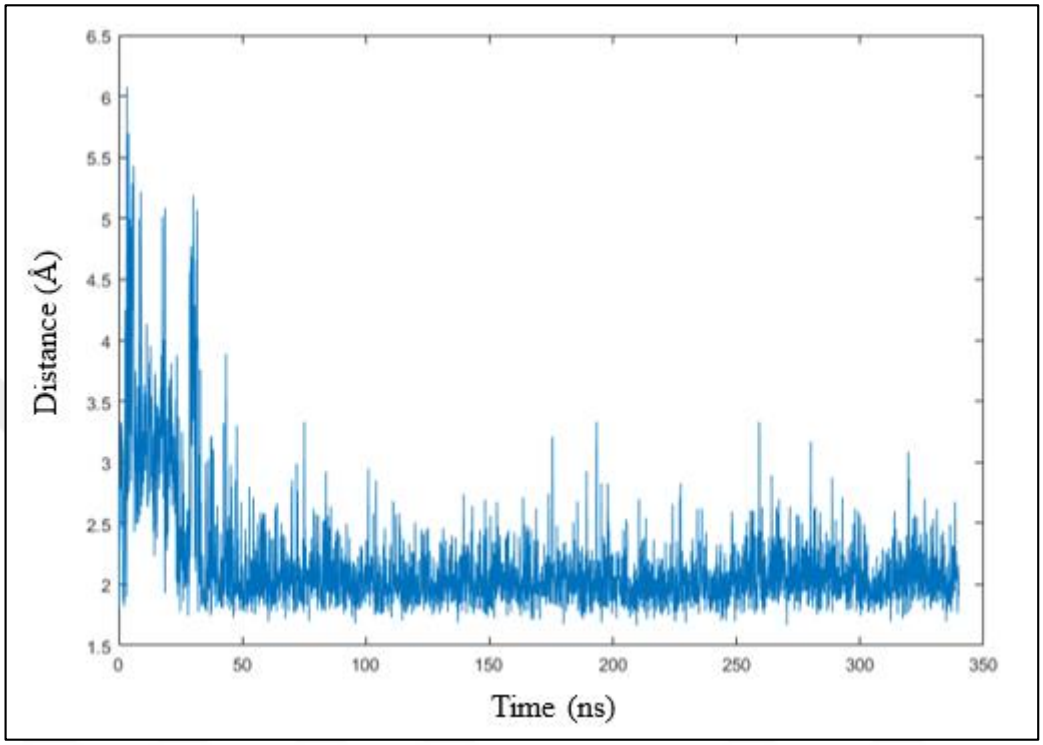


Figure 3.18. Distance vs time plot of GLY111 O – HIE286 HE2 (2NW6 s-trans MD)

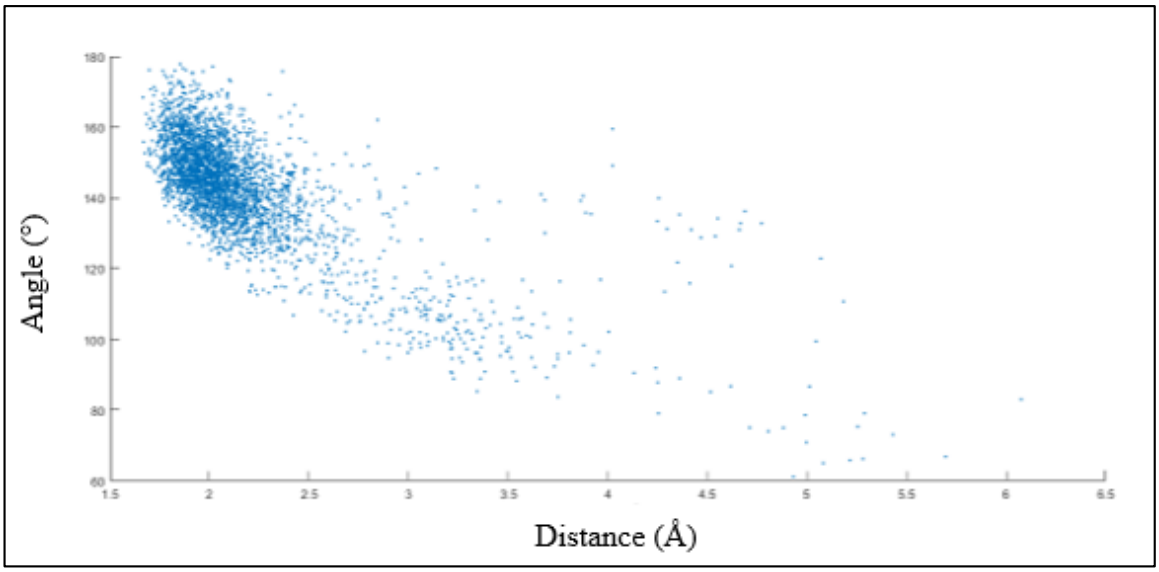


Figure 3.19. Distance vs angle plot of GLY111 O – HIE286 HE2 – HIE286 NE2 (2NW6 s-trans MD)

One of the most stable interactions of the 2NW6 s-Trans MD was established between GLY111 O – HIE286 HE2 as seen in the distance vs time plot in figure 3.26. After an initial settling for the first 25 ns, distance only fluctuates between 1.25 Å – 2.25 Å and the angle fluctuates between 120° – 150° degrees at the same distance range as seen in figure 3.27. This interaction shows that histidine did not interact with aspartate as expected but it made a strong hydrogen bond with backbone OH of glycine.

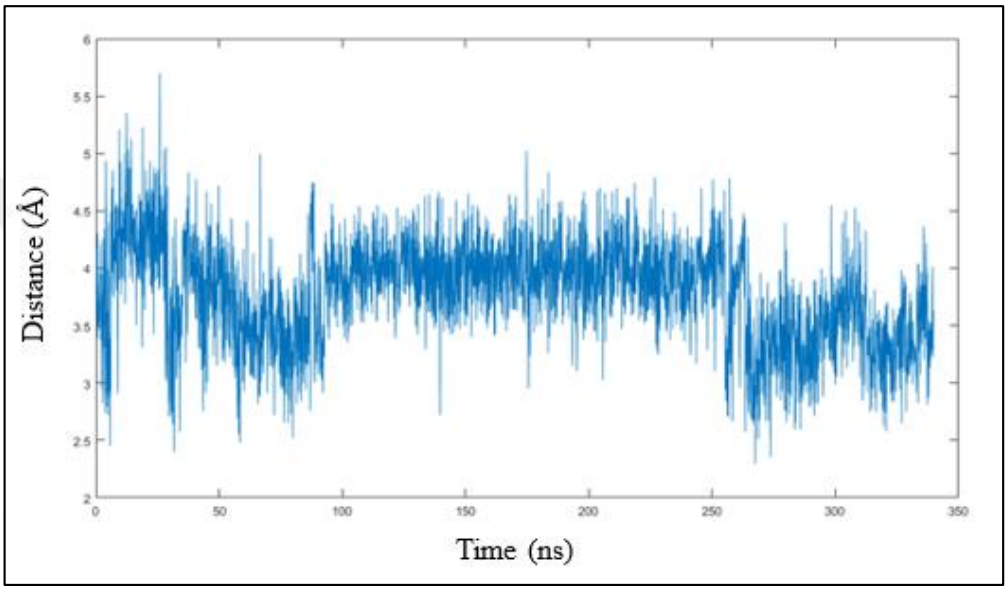


Figure 3.20. Distance vs time plot of SER87 OG – HIE286 H2 (2NW6 s-trans MD)

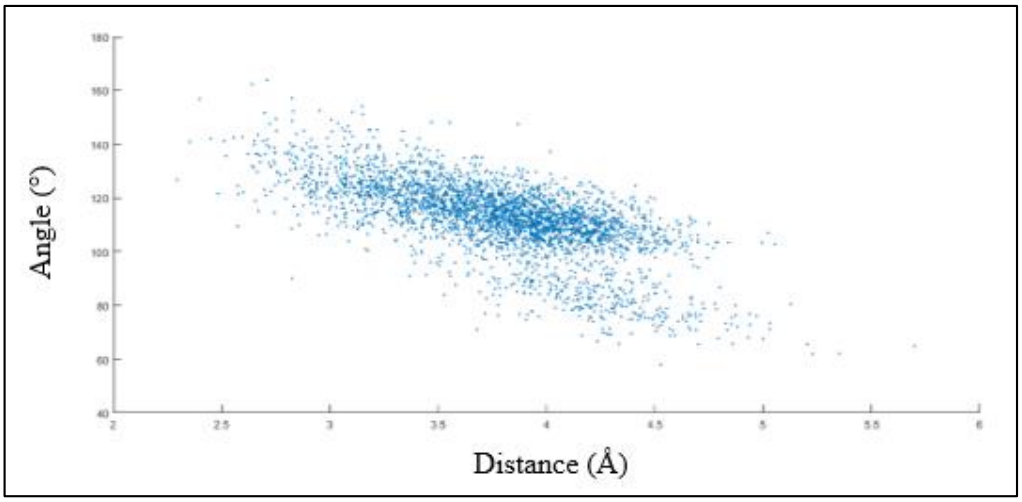


Figure 3.21. Distance vs angle plot of SER87 O – HIE286 H – HIE286 NE (2NW6 s-trans MD)

Interaction between the matching residues SER87 OG – HIE286 H2 fluctuated around 3.5 - 4 Å and no H-bond was established between Ser-His of the catalytic triad as can be seen from figures 3.19 and 3.20.

The interaction between HID-SER is very strong as seen in Figure 3.20 and 3.21 because the distance between them is almost always constant at 2 Å with minimum fluctuation which means the hydrogen bonding was very strong.

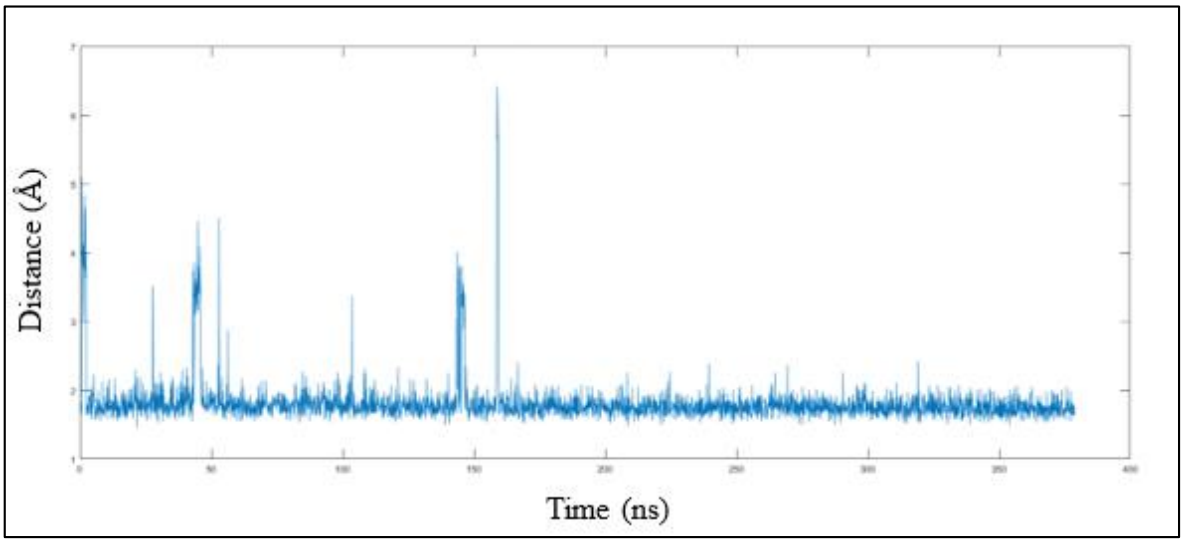


Figure 3.22. Distance vs time plot of HID286 NE2 – SER87 HG (2NW6 apo MD)

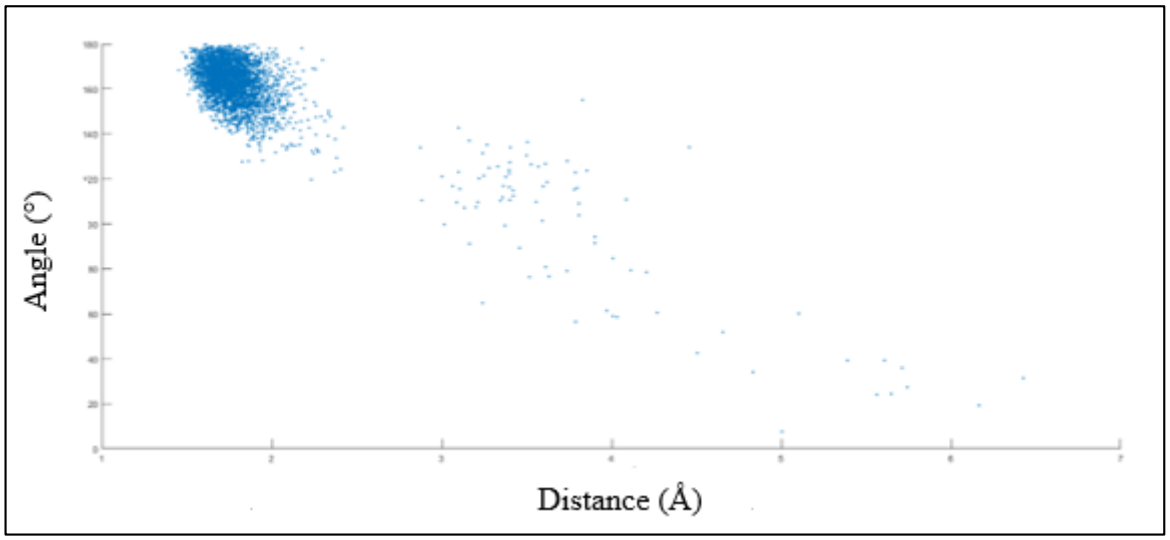


Figure 3.23. Distance vs angle plot of HID286 NE2 – SER87 HG (2NW6 apo MD)

Interaction between ASP-HIS does not seem to be as strong as HIS-SER as shown in Figure 3.22 and Figure 3.23. The reason might be the very strong bond between HIS-SER that left ASP looser.

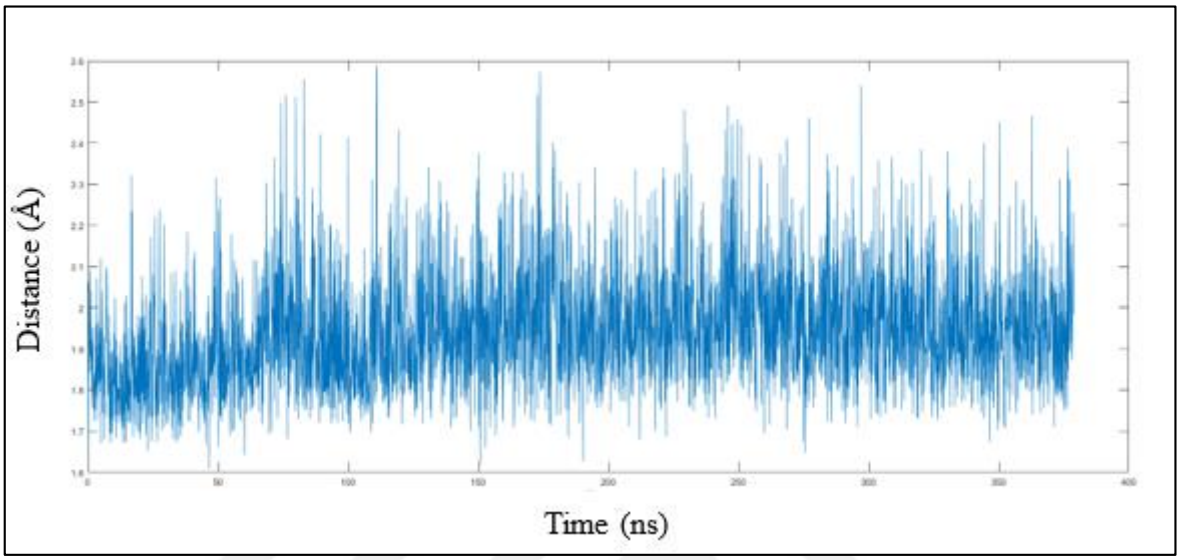


Figure 3.24. Distance vs time plot of ASP264 OD1 – HID286 H (2NW6 apo MD)

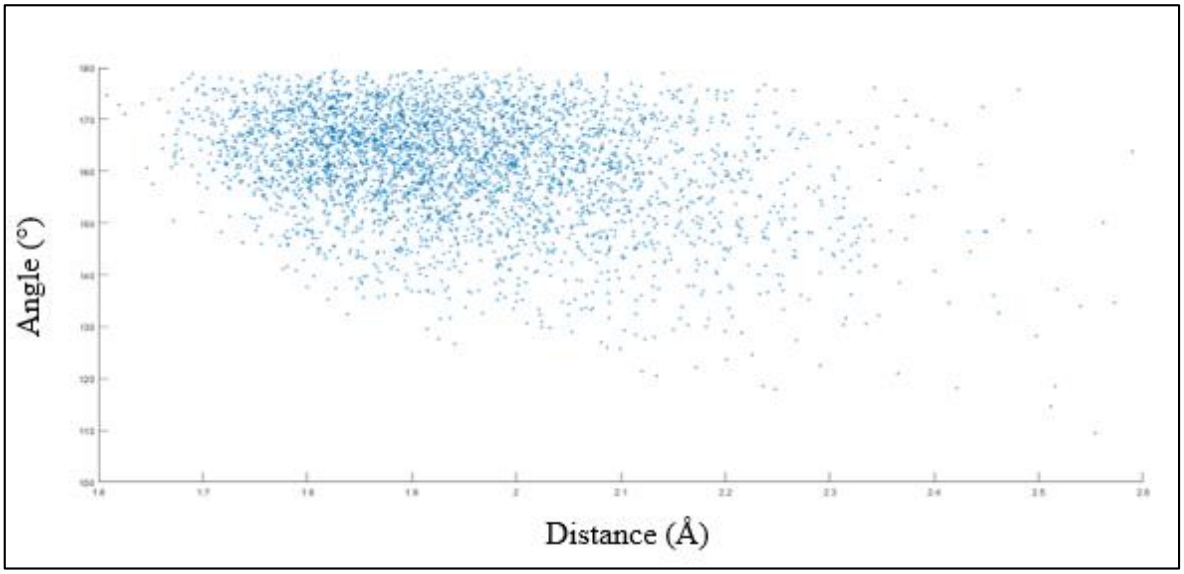


Figure 3.25. Distance vs angle plot of ASP264 OD1 – HID286 H (2NW6 apo MD)

3.2.3. 1FFE – MD Analysis

1FFE is a cutinase with side chain length of 214. Some of the analysis results from MD with s-Trans ligand are listed below.

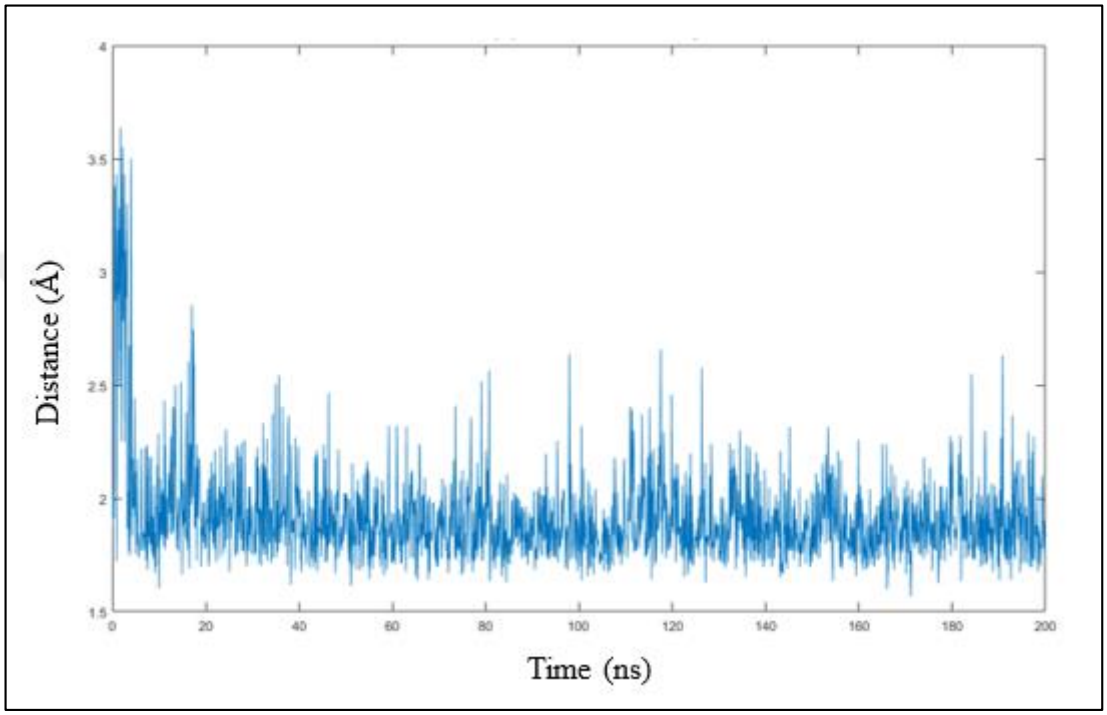


Figure 3.26. Distance vs time plot of ASP159 OD1 – HID172 HD1 (1FFE s-trans MD)

One of the most important results show a strong interaction between ASP159 OD1 – HID172 HD1 which are the residues belonging to the catalytic triad. Distance vs time plot in figure 3.35 show they maintain a stable hydrogen bond for almost all of the simulation length with a residue fluctuation between 1.5 Å - 2.0 Å. As explained earlier, normally catalytic triad contains histidine residue (HIS), but it was converted manually to HID here because at this case the delta hydrogen needs to be protonated instead of the epsilon one.

The distance vs angle plot in Figure 3.36 also demonstrates a stable scattering between 150° – 175° angles at the same distance interval as in Figure 3.35.

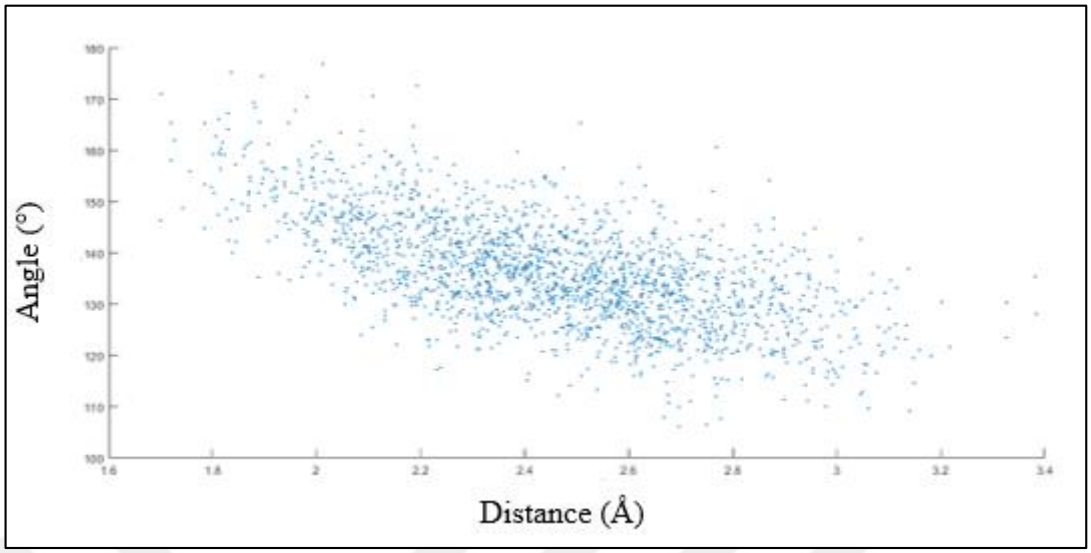


Figure 3.27. Distance vs angle plot of ASP159 OD – HID172 HD1 (1FFE s-trans MD)

The hydrogen bonding between HIS-SER is very strong as expected as they are a part of the catalytic triad. Only very little fluctuation is noticed between molecule distances, for the most part it is stable at 2 Å as seen in Figure 3.26 and Figure 3.27.

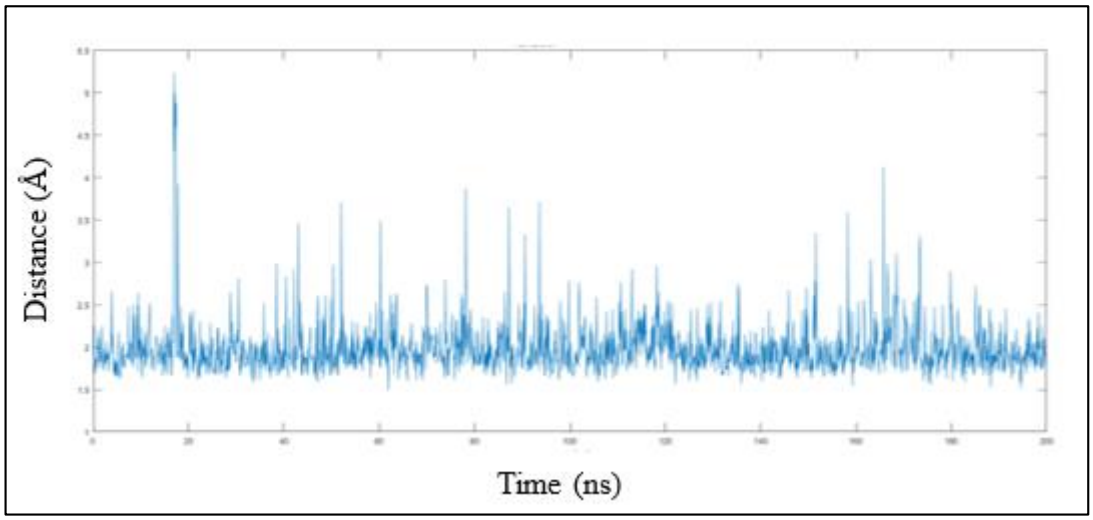


Figure 3.28. Distance vs time plot of HID188 H – SER120 OG (1FFE apo MD)

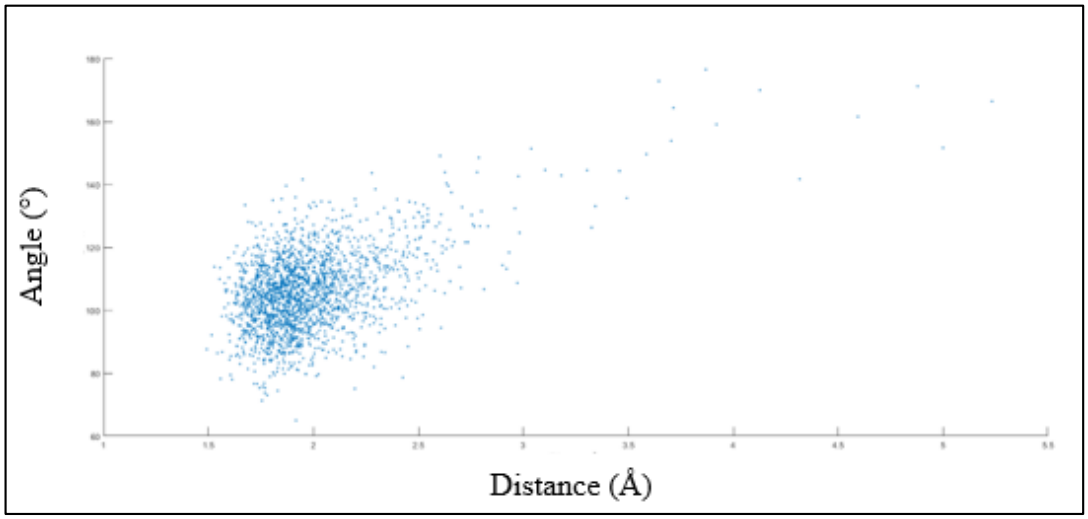


Figure 3.29. Distance vs angle plot of HID188 H – SER120 OG (1FFE apo MD)

SABER jobs were run with same delta level of two for all CAMs. Most successful SABER results were obtained for five atoms because six atoms was a very limiting search criteria that no protein matches were found in the same order. On the contrary, SABER protein match results of CAM 1 with 4 atoms led to too many results that it was more difficult to go through and a lot of irrelevant protein matches were obtained. Therefore optimal number of 5 atoms were used to construct CAMs for Theo3.

Even though results were filtered to see the best protein matches, still too many protein results were obtained to analyze with MD. It was noticed that three proteins were found in common of all three theozymes' SABER results which are 2NW6, 2BDB and 1FFE. These proteins had low rmsd values and matched well with their theozyme models, so MD analysis were done for these three proteins under three conditions.

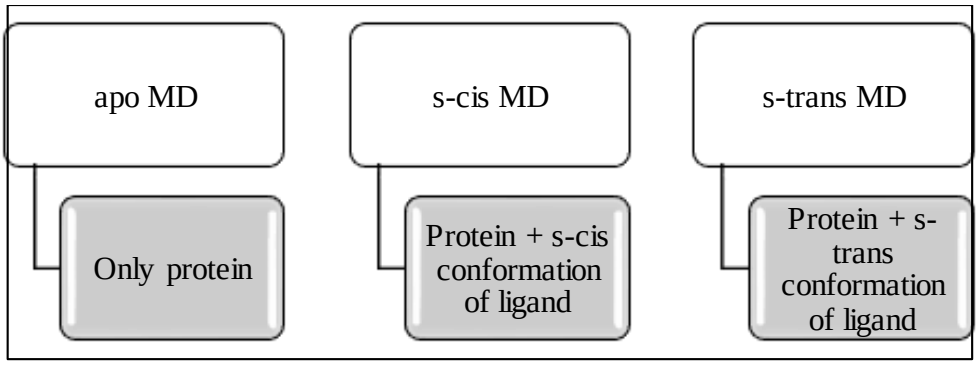


Figure 3.30. MD analysis conditions

MD analysis was simply done to observe the behaviours of these potentially enzymatic proteins at conditions closest to real life in computational environment. MD simulations are chosen because less expensive and time consuming then real life experiments, so it is more advantageous especially at earlier stages of a study. Once most promising proteins are found out via MD analysis, they can be tested experimentally.

4. CONCLUSION

In this study, protein data bank was screened by using SABER. Three previously constructed theozyme models were used. First theozyme including water with TS3 and s-trans and s-cis conformations of second theozyme including methanol with TS3 were used. Heteroatoms of theozyme models were selected to make the xyz files. Series of conversions from xyz to tess and to xml file which was the main input of a SABER run was completed. Xml file was manually edited in order to increase the number of matches by adding residue substitutes to the file. Optimal results were obtained when five atoms (OD2, NE2, ND1, OG, N) were selected for all theozymes. No results were obtained for six atoms and results were too much for four atoms that it could be misleading.

When all SABER results were analyzed, three important proteins were found in common with good results for all three theozymes which are 2BDB, 2NW6, 1FFE. As a next step, MD simulations were run for these three proteins for 400 ns. Two conformations of the ligand, s-cis and s-trans, were used along with apo MD which does not contain any ligand.

Best results in terms of the stability of the atomic interactions of catalytic triad were found in results of 2BDB. Distance vs time plot of ASP187 OD2 – CYX184 H and ASN65 OD1 – ASN66 H were given as examples to this. So this protein might be investigated further in experimental studies for further information about its enzymatic activity for this reaction.

REFERENCES

1. Tantillo DJ, Jiangang C, Houk KN. Theozymes and compuzymes: Theoretical models for biological catalysis. *Current Opinion in Chemical Biology*. 1998;2:743-745.

2. Shi M, Wang F, Zhao M, Wei Y. *The chemistry of the Morita Baylis Hillman reaction.* London: Royal Society of Chemistry; 2011.

3. Lima-Junior CG, Vasconcellos ML. Morita-Baylis-Hillman adducts: Biological activities and potentialities to the discovery of new cheaper drugs. *Bioorganic and Medicinal Chemistry*. 2012;20(13):3954-3971.

4. Liu Y. Organocatalyzed Morita Baylis Hillman Reaction: Mechanism and Catalysis [dissertation]. Department Chemie, Ludwig-Maximilians-Universitat München, 2011.

5. Damborsky J. Brezovsky J. Computational tools for designing and engineering enzymes. *Current Opinion in Chemical Biology*. 2014;19:8-16.

6. Kapoor M, Majumder AB, Gupta MN. Promiscuous lipase-catalyzed C–C bond formation reactions between 4 Nitrobenzaldehyde and 2-Cyclohexen-1-one in biphasic medium: Aldol and Morita–Baylis–Hillman Adduct Formations. *Catalysis Letters.* 2015;145:527–532.

7. Mäntsälä P, Niemi J. Enzymes: The biological catalysts of life. *Physiology and Maintenance.* 2009;2:22-25.

8. Schramm VL. Enzymatic transition states, transition-state analogs, dynamics, thermodynamics, and lifetimes. *Annual Review of Biochemistry*. 2011;80:703–732.

9. Schramm VL. Enzymatic transition-state analysis and transition state analogs. *Methods Enzymol*. 1999;308:301–55.

10. Steiner K, Schwab H. Recent advances in rational approaches for enzyme engineering. *Computational and Structural Biotechnology Journal*. 2012;2:e201209010.

11. Linder M. Computational enzyme design: Advances, hurdles and possible ways forward. *Computational and Structural Biotechnology Journal*. 2012;2:e201209009.

12. Sebestova E, Bendl J, Brezovsky J, Damborsky J. Computational tools for designing smart libraries. *Methods in Molecular Biology*. 2014;1179:291-314.

13. Feng XW, Li C, Wang N, Li K, Zhang WW, Wang Z, et. al. Lipase-catalysed decarboxylative aldol reaction and decarboxylative Knoevenagel reaction. *Green Chemistry*. 2009;11:1933-1936.

14. Madalinska L. Kwiatkowska M. Cierpial T. Kielbasinski P. Investigations on enzyme catalytic promiscuity: The first attempts at a hydrolytic enzyme-promoted conjugate addition of nucleophiles to α,β-unsaturated sulfinyl acceptors, *Journal of Molecular Catalysis B: Enzymatic*. 2012;81:25-30.

15. Pattanaik S. Mohapatra P. Origin, evolution and diversity of phosphotriesterases-an organophosphate degrading enzyme. *The Ecoscan*. 2013;3:23-26.

16. Copley SD. Shining a light on enzyme promiscuity. *Current Opinion in Structural Biology*. 2017;47:167-175.

17. Nosrati GR. Houk KN. SABER: a computational method for identifying active sites for new reactions. *Protein Science*. 2012;21(5):697-706.

18. Saber. 2015 [cited 2017 3 December]. Available from: http://med.stanford.edu/tanglab/software/saber.html

19. Frenkel D. Smit B. *Understanding molecular simulation: from algorithms to applications*. San Diego: Academic Press; 2002.

20. Dahlke EE. Truhlar DG. Assessment of the pairwise additive approximation and evaluation of many-body terms for water clusters. *The Journal of Physical Chemistry*. 2006;110(22):10595-10601.

21. Monticelli L. Tieleman DP. Force fields for classical molecular dynamics. *Biomolecular Simulations*. 2013;924:197-213.

22. Gonzalez M. Force fields and molecular dynamics simulations. *Les Journées de la Neutronique (JDN)*. 2011;12:169-200.

23. Sherrill CD. Introduction to molecular mechanics. [Lecture] Georgia Institute of Technology. 2018.

24. Jorgensen WL. Tirado-Rives J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proceedings of the National Academy of Sciences*. 2005;102:6665–6670.

25. Price DJ. Brooks III CL. Modern protein force fields behave comparably in molecular dynamics simulations. *Journal of Computational Chemistry*. 2002;23:1045–1057.

26. Yeh IC. Hummer G. Peptide loop-closure kinetics from microsecond molecular dynamics simulations in explicit solvent. *Journal of American Chemical Society*. 2002;124:6563–6568.

27. Aliev AE. Courtier-Murias D. Experimental verification of force fields for molecular dynamics simulations using Gly-Pro-Gly-Gly. *Journal Physical Chemistry*. 2010;114:12358–12375.

28. Paton RS. Goodman JM. Hydrogen bonding and π-stacking: How reliable are force fields? A critical evaluation of force field descriptions of nonbonded interactions. *Journal of Chemical Information and Modeling*. 2009;49:944–955.

29. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE, Cruzeiro VWD, et.al. (2018), AMBER 2018[cited 2019 17 October], Available from: http://ambermd.org/contributors.html.

30. Plata RE. Singleton DA. A case study of mechanism of alcohol-mediated Morita Baylis Hillman reactions. The Importance of Experimental Observations. *Journal of the American Chemical Society*. 2015;137:3811-3826.

31. Robiette R. Aggarwal VK. Harvey JN. Mechanism of the Morita-Baylis-Hillman reaction: A computational investigation. *Journal of the American Chemical Society*. 2007;129:15513-15525.

32. Aggarwal VK, Fulford SY, Llyod-Jones GC. Reevaluation of the mechanism of the Baylis-Hillman reaction: Implications for Asymmetric Catalysis. *Angewandthe Chemie International Edition*. 2005;44:1706-1708.

33. Price KE. Broadwater SJ. Jung HM. McQuade DT. Baylis-Hillman mechanism: A new interpretation in aprotic solvents. *Organic Letters*. 2005;7:147-150.

34. Roy D. Sunoj RB. Ab-initio and density functional theory evidence of the rate-limiting step in the Morita-Baylis-Hillman reaction. *Organic Letters*. 2007;9:4873-3876.

35. Sheldon RA. Woodley JM. Role of biocatalysis in sustainable chemistry. *Chemical Reviews*. 2018;118(2):801-838.

36. Truppo MD. Biocatalysis in the pharmaceutical industry: The need for speed. *ACS Medicinal Chemistry Letters*. 2017;18(5):476-480.

37. Hu W, Guan Z, Deng X, He YH. Enzyme catalytic promiscuity: The papain-catalyzed Knoevenagel reaction. *Biochimie*. 2011;94:656-661.

38. Salomon-Ferrer R. Goetz AW. Poole D. Grand S. Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle Mesh Ewald. *Journal of Chemical Theory and Computation*. 2013;9(9):3878-3888.

39. Utnier T. Modeling Optimal Catalytic Active Sites for the Morita-Baylis-Hillman Reaction [dissertation]. Department of Chemical Engineering. Yeditepe University. 2017.