# Evaluacija aktivnih mjesta enzima i usporedba sa kataliticki aktivnim peptidima koji kataliziraju hidrolizu estera

**Babić, Marko**

**Master's thesis / Diplomski rad**

**2021**

*Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj:* **University of Rijeka / Sveučilište u Rijeci**

*Permanent link / Trajna poveznica:* https://urn.nsk.hr/urn:nbn:hr:193:230088

SVEUČILIŠTE U RIJECI
ODJEL ZA BIOTEHNOLOGIJU

Diplomski sveučilišni studij

Istraživanje i razvoj lijekova

Marko Babić

# Evaluacija aktivnih mjesta enzima i usporedba sa katalitički aktivnim peptidima koji kataliziraju hidrolizu estera

Diplomski rad

Rijeka, 2021.

SVEUČILIŠTE U RIJECI
ODJEL ZA BIOTEHNOLOGIJU

Diplomski sveučilišni studij

Istraživanje i razvoj lijekova

Marko Babić

# Evaluacija aktivnih mjesta enzima i usporedba sa katalitički aktivnim peptidima koji kataliziraju hidrolizu estera

Diplomski rad

Rijeka, 2021.

Mentor: dr.sc. Daniela Kalafatović

UNIVERSITY OF RIJEKA

DEPARTMENT OF BIOTECHNOLOGY

Graduate program

Drug research and development

Marko Babić

# Evaluation of active sites of enzymes and their comparison to catalytically active peptides involved in ester hydrolysis

Master thesis

Rijeka, 2021.

Mentor: dr.sc. Daniela Kalafatović

Diplomski rad obranjen je dana

29.9.2021.

Pred povjerenjstvom:

1. Doc. dr. sc. Jelena Ban

2. Doc. dr. sc. Željko Svedružić

3. Doc. dr. sc. Daniela Kalafatović

Rad ima 70 stranica, 18 slika, 6 tablica i 56 literaturnih navoda.

## Saţetak

Racionalni dizajn proteina i peptida je polje koje zahtijeva razumijevanje funkcije proteina do razine aminokiselina. Ovo znanje obuhvaća razumijevanje geometrije i sastava aktivnih aminokiselinskih ostataka. Ovaj rad će se fokusirati na aktivna mjesta prirodnih enzima i pokretače katalitičke funkcije u njima. Izvršena analiza ima za svrhu razradu seta kriterija za izgradnju sintetičkih peptida s istom katalitičkom funkcijom, te općenito produbljivanje razumijevanja enzima.

Izgrađen je skup podataka EC 3.1. podklase (hidrolaze esterskih veza) kako bi se analizirao sastav i geometrija aminokiselina aktivnog mjesta. Uzorci u podatcima istraženi su analizom 96 esteraza. Tijekom tog procesa, 22 enzima izdvojeno je na kriteriju posjedovanja katalitičke trijade za daljnju analizu. Temeljeno na primarnoj sekvenci odabranih enzima stvoreni su profili sastava za (1) cijelu sekvencu enzima, (2) "dugo aktivno mjesto" koje je uključivalo sve aminokiseline od prvog do zadnjeg aktivnog bočnog lanca, (3) "kratko aktivno mjesto" koje je sadržavalo isključivo bočne lance katalitički aktivnih aminokiselina i (4) "katalitičku petlju" koja je sadržavala 4 aminokiseline sa svake strane aktivnog bočnog lanca.

Korištena su dva pristupa od kojih su prvim analizirani sastavi i kemijska svojstva dok su se u drugom koristile kristalne strukture PDB datoteka dobivenih X-zrakama rezolucija ispod 3,0 Å kako bi se izmjerile geometrije bočnih lanaca.

Rezultati su pokazali kako su udaljenosti aktivnih bočnih lanaca unutar kojih prirodni enzimi vrše funkciju manji od 1Å variacije te kutevi između njih ne variraju više od 10% od prosječnog kuta aktivne aminokiseline u svom interkvartalnom rasponu. Nadalje, kvantitativna analiza aminokiselina blizu aktivnih bočnih lanaca pokazuje povećanje non-polarnih i smanjenje u bazičnim, hidroksilnim i polarnim aminokiselinama. Ovi rezultati prikazuju strogu funkcionalnu geometriju za optimalnu enzimatsku aktivnost, kao i postojanje specifičnog aminokiselinskog sastava koji aktivna mjesta

toleriraju. Ovo istraživanje može potencijalno razviti algoritme za predviđanje i optimizaciju enzimatskih funkcija i nove teoretske modele za modularni dizajn peptida.

**Ključne riječi:** Dugo aktivno mjesto, kratko aktivno mjesto, katalitička petlja, trijada, esteraze

# Summary

Rational design of proteins and peptides is a growing field that requires knowledge of protein function down to the amino acid level. Understanding protein geometry and amino acid composition is crucial for understanding molecular and supramolecular chemistry in protein design. In this thesis, the focus is put on active sites of natural enzymes and the drivers of catalytic function in said active sites. The analysis of natural enzyme active sites was done with the purpose of having a set of criteria for building synthetic peptides with identical catalytic function and to broaden enzyme understanding.

A data set was built to analyze enzymes from the EC 3.1. (Hydrolases acting on ester bonds) subclass looking at both amino acid content and geometry of active site residues. Patterns in existing data were searched by statistically analyzing 96 esterases. In this process, 22 enzymes with known catalytic triads were selected for further evaluation. Based on the primary structure of the selected enzymes, the composition profiles were created for: (1) the full sequence, (2) the "long active site" including the residues between the first and the last active amino acid, (3) the "short active site" containing only active residues and (4) the "catalytic loop" which looked at 4 residues from each side of the sequence from the catalytic residues.

To analyze the dataset two approaches were used. The first approach was based on composition and chemical property analysis. The second approach consisted in using crystal structure PDB files obtained via X-ray, only considering a resolution below 3,0 Å and in measuring the residue geometries in PyMol.

The results showed that the distances within which natural enzymes function are less than 1Å in variation and that the angles conform within a 10% variation in the interquartile range compared to the average angle.

Moreover, the qualitative amino acid analysis showed an increase of non-polar and a decrease in basic, hydroxylic and polar residues near the catalytically active residues. This suggests a strict functional geometry for the optimal enzymatic activity as well as a specific local amino acid content tolerated in active sites. These findings will allow the development of algorithms for prediction and optimization of enzyme functions and of new theoretical models of modular peptide design.

**Key words:** Active site, catalytic triad, catalytic loop, long active site, short active site

# Contents

# 1 Introduction

Enzymes are molecules that catalyze chemical reactions. This catalysis is done by lowering the activation energy of reactions, changing the rate of the reaction significantly without influencing the position of the chemical equilibrium. Enzymes exploit the aforementioned activation energy using the amino acid chemistry and the specific 3D protein structure.[1,2] Because of this, enzymes can be used for creating efficient chemical synthesis processes that require less solvents and steps compared to conventional chemical synthesis based solely on chemical reactants. However, enzymes are difficult to synthesize. Therefore, there is a need for further studies of catalytic sites to both create smaller synthetic peptides as alternatives to large proteins and to better understand enzyme biochemistry in general. However, the parametrization of active site geometry and qualitative analysis of enzyme amino acid composition are sparse. This is reflected in the research of synthetic peptides where the importance of active site chemistry and geometry are acknowledged but not directly applied. The literature mostly contains cut out portions of natural active sites or active site residues implanted in self-folding scaffolds (See heading 1.5). The former expecting that a catalytic loop will retain its geometry and the latter that the residues will self-orientate. While these peptides managed to perform catalysis their measured $k_{cat}$ values were significantly lower compared to their enzyme progenitors. The other challenge is that not all enzymes have single catalytic loops that can be simply cut out. Therefore, in depth understanding of geometric and chemical parameters of active sites is important to create better catalytic peptides and expand design methods to include more catalytic mechanisms. To expand this large field, this thesis starts with an enzyme subclass that has well researched active sites and for which many synthetic peptides with equivalent catalytic function were found. This subclass is the EC 3.1 subclass of ester hydrolases, enzymes that hydrolyze ester bonds.

The aim of this work was to analyze the geometry and composition of EC 3.1 active sites to gain insight into parameters important for understanding enzyme catalysis and synthetic peptide design. Specifically, this work was focused on a subdivision of EC 3.1 that contain catalytic triads and oxyanion holes.

## 1.1 EC numbers and divisions of EC 3.1

Enzyme commission (EC) number is a numerical classification system for enzymes. It is based on the reactions the enzymes perform, for example, two completely different enzymes with different evolutionary background can have the same EC number if they catalyze the same reaction.

There are four levels to the EC classification. The 1st level is related to the general reaction the enzyme catalyzes. There are seven 1st level codes:

1. Oxidoreductases which perform redox reactions,
2. Transferases which transfer functional groups,
3. Hydrolases which form two products from a substrate by hydrolysis,
4. Lyases which add or remove groups from substrates,
5. Isomerases which rearrange molecules,
6. Ligases that join together molecules by breaking down ATP and
7. Translocases which catalyze movement of molecules or ions across membranes.

The 2nd level of the classification is based on the type of chemical bonds the enzymes act on. For example, EC 2.6 are transferases that transfer nitrogenous groups, while EC 2.9 are transferases that transfer selenium containing groups. The 3rd and 4th level vary based on the class of enzyme. The 3rd level is related to the nature of the acceptor molecules or the type of molecule transformed on the substrate by the enzyme. For example, the EC 2.1 subclass of transferases transferring one carbon group is subdivided into methyltransferases, hydroxymethyl-, formyl- and related transferases,

carboxyl- and carbomoyltransferases, amidotransferases and methylenetransferases; from 2.1.1 to 2.1.5 respectively.

Or in the case of the 1.1 subclass based on acceptors like NAD, quinone, oxygen, *et cetera*. The 4th level is associated to a specific substrate molecule such as the 1.1.1.20 which is specifically a glucuronolactone reductase.

## 1.2 Hydrolases

In this work we focused on the EC 3.1 subclass of enzymes, the hydrolases that act on ester bonds. According to NC-IUBMB there are 31 sub-subclasses of hydrolases.[3] These sub-classes are distinguished in the 3rd EC level by the type of substrate molecules they act on: carbohydrates, phosphates, ribonucleases, deoxyribonucleases or thio/sulfur compounds. To form the 31 3rd level list, the previously mentioned bonds are further differentiated based on whether they are monoester, diester or triester hydrolases, exo- or endo- nucleases, 3' or 5' acting (deoxy)ribonucleases, site specific or non-specific enzymes. The last, 4th level of EC class has hundreds of categories further going into specific compounds they act on. An example is the cutinase which has an EC classification of 3.1.1.74. The 3rd level 3.1.1 means cutinase is a carboxylic ester hydrolase and the 4th level 3.1.1.74 means the enzyme is a carboxylic ester hydrolase that hydrolyses cutin polymer molecules.

This work will focus on 96 EC3.1. hydrolases that have unique and well described catalytic sites in the European Molecular Biology Laboratory (M-CSA) database.[4]

## 1.3 Active sites

Active sites are enzyme regions that bind to a substrate and where the substrate chemically changes with the help of the enzyme. The active site consists of a binding site and of the catalytic site. In natural enzymes, the binding site is usually substrate specific and consists of amino acids that

create a chemically and sterically favorable environment for a substrate to bind and orient towards the catalytic site. The catalytic site performs the catalysis by lowering the activation energy of the reaction. Catalytic sites contain a dyad or a triad and a variety of cofactors, coordinated metal ions, modified amino acids or other molecules to aid or perform the catalysis.[5]

The EC 3.1 hydrolases show different reaction mechanisms. This work will address a subset of EC 3.1 enzymes containing the catalytic triad . This category of active sites always contains three catalytic residues where one acts as a nucleophile, the second acts as a base deprotonating the nucleophile and the third is an acid molecule. In addition, this type of active sites always contains oxyanion holes that stabilize the tetrahedral intermediate of the nucleophile binding to the substrate. The nucleophile is usually a residue with a strongly nucleophilic atom on the side chain, like oxygen or sulfur. Histidine is the most common base and aspartate and glutamine, carboxylic side chain containing residues, are usually the acids. Triads will be further referenced in section 1.4.

However, when discussing EC 3.1 enzymes, the triad mechanism is not the most numerous of the subclass. Metalloenzymes are the most numerous of the hydrolases. These enzymes use metal cations to perform catalysis, mostly calcium, zinc and magnesium. An example is the isotuberculosinol synthase which uses its $Mg^{2+}$ to anchor the substrate tuberculosinyl diphosphate while an arginine donates its proton to activate the diphosphate for hydrolysis[6]. Metal ions coordinate the substrates and reduce the pKa of water molecules near the coordination sphere to allow the nucleophilic attack of the substrate[7–13]. These metalloenzymes vary in terms of the number of bound metals, with one, two or three bound atoms, and based on the proton donor which activates the substrate. The most common proton donors being histidine, followed by aspartate, glutamate, lysine, arginine and threonine.

Another common type of enzyme mechanism in EC 3.1 are catalytic dyad mechanisms. An example is phospholipase A2 that contains a serine/aspartate dyad in which aspartate activates the serine by deprotonating it so it can attack the carbonyl carbon of the substrate. The Aspartate later deprotonates a water molecule which attacks the remaining substrate bound to serine[14]. These dyads always consist of a nucleophile and acid or base, of which nucleophiles are usually serine[14,15], histidine[16–18], cysteine[19,20] and sometimes glutamate[21] or aspartate[22]. The acid or base residue is mostly aspartate[18,18,19] or histidine[15], with some exceptions such as in the case of fructose-2,6-biphosphatase enzymes which have a glutamate as acid.[17]

There are many other mechanisms that differ from all three mentioned groups. The EC 3.1 classification also contains active sites that use modified amino acids[23,24], proton relays[25] and reactions started by the substrate activating enzyme residues to perform nucleophilic attack[26].

## 1.4 Catalytic triad and oxyanion holes

### 1.4.1 Catalytic triad

The catalytic triad is composed of a tri-molecule group of amino acids which, with their side chains, perform an SN2 bimolecular nucleophilic substitution. The three residues do this in five steps and each residue has its specific and irreplaceable role in the reaction being the nucleophile, the base and the acid. The acid molecule serves to increase the basicity of the base, modifying its pKa via hydrogen bonds. The base molecule uses this increased basicity to activate the nucleophile secondly to protonate the first leaving group and lastly to deprotonate another molecule, usually water, so it can bind to the second leaving group and regenerate the active site with this taken proton. The nucleophile role is performed by an amino acid with a strongly electronegative atom at the top of its side chain, usually an oxygen or sulfur atom. This atom, when activated, performs a nucleophilic attack and covalently attaches to the carbonyl carbon of the substrate. This

action creates a tetrahedral intermediate which collapses when the base residue protonates the first leaving group. The rest of the substrate is cleaved from the nucleophile by another deprotonated molecule from the solution, usually water, binding to the remaining substrate. In the last step, the nucleophile residue is regenerated by the base .[27,28]

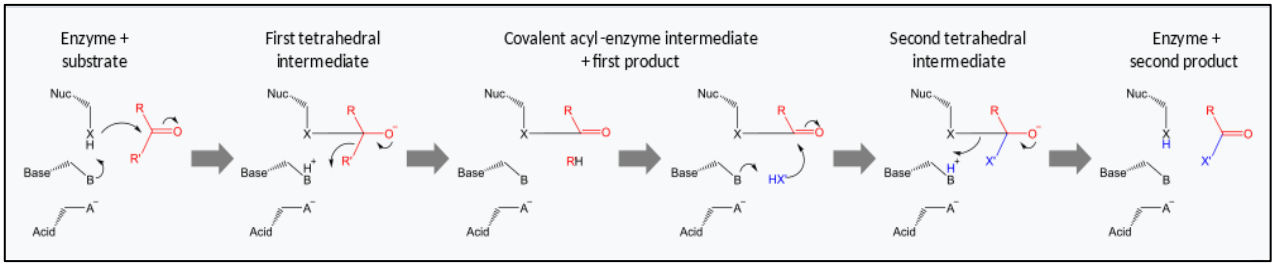


**Figure 1 - Reaction scheme of catalytic triads – This scheme represents the general reaction mechanism for all catalytic triads. The picture is taken from Thomas, Shafee, (2014).[27]**

The acid residues are amino acids like aspartate, glutamate, histidine or asparagine. The base residues are mostly histidines but also lysines or in some cases N-terminal amides [15]. The nucleophiles are usually serine, cysteine or threonine and sometimes tyrosine.[29]

This thesis will focus on the most common triad being the Ser-His-Asp with a few exceptions, namely Ser-His-Glu, Cys-His-Asp and Ser-His-Trp (C-main) triads.

### 1.4.2 Oxyanion hole

The main stabilizers of ester hydrolysis reactions are the amino acids that form the oxyanion hole. The oxyanion hole is an integral part of the triad mechanism[29] and mutating the residues can cause a 100 to 1000 fold decrease in catalytic function[30–33]. The term oxyanion hole means a hole for a negatively charged (anion) oxygen which is formed by a nucleophilic attack onto the carbonyl carbon. The electrons from the C-O double bond

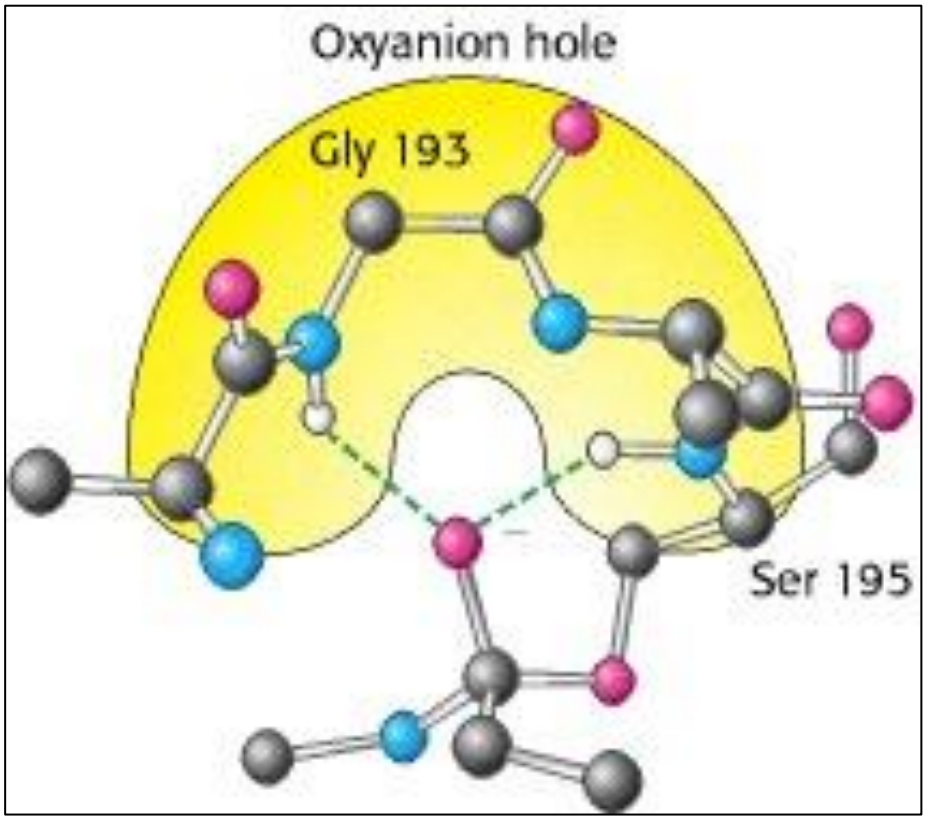


**Figure 2 - Oxyanion hole - picture from: Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. Section 9.1, Proteases: Facilitating a Difficult Reaction. This picture shows how positively charged nitrogen groups stabilize the negative oxygen in the tetrahedral intermediate.**

localize on the oxygen while in the tetrahedral intermediate form. They are then stabilized by the positively charged atoms of the oxyanion hole members. These atoms are usually nitrogens that form a slightly positive hole in which the oxyanion can stabilize in. This intermediate forms twice

in every triad reaction these holes help to reduce the activation energy in both instances.

## 1.5 Synthetic catalytic peptides

The literature contains a number of papers that report on the synthesis and characterization of designed peptides that mimic the reactions of esterases, mostly in pNPA assays.[34–42] While their activity is significantly slower than that of natural enzymes with equivalent function their measurable activity shows that abstracting some of the core catalytic residues into a folding or self-assembling peptide can result in catalytic function.

The peptides that have esterase activity range from small 7 amino acid molecules to large multi-chain structures. However, the reported active molecules that directly contributed to the reaction were always under 5 amino acids. If there was a larger structure reported in the literature, it was always a scaffold onto which catalytically active amino acids were bound to. For example, the sequence called 'VK2H' (H2N-HSGVKVKVKVKV$^{D}$PPTKVKVKVKV-CONH2). The sequence has alternating valine and lysine amino acids with a D-proline turn in between the two sheet-forming segments. The structure contains a set of catalytically active residues held at the end of the antiparallel sheet (the HSG- in the sequence).[42] The VK2H structure is also called MAX1.[36] Another example is the MIDI-zinc complex (or using other names; 3SCC and Zn(II)-His3O), a heterodimer structure of three alfa helices. Each helix monomer contains a histidine molecule that stabilizes the central zinc ion which performs the catalysis.[38,43] Other often used scaffolds for protein design are polyproline scaffolds, the TIM barrel, periplasmic binding protein, lipocalin, jellyroll, and β-propellers.[44]

These scaffolds incorporate either catalytic triads or metal ions in between amino acids that tend to coordinate with the metal, mostly histidine. We focused on the peptides that are up to 25 amino acids in length, have the highest activity and that contain a catalytic triad. Namely, 3 were chosen:

1) A hydroxyproline scaffold with a Cys3-His6-Asp9 triad configuration. Formula Ac-OO**C**OO**H**OO**D**OOGY-NH2 (Hyp-C**3**H**6**D**9**)
2) A polyproline scaffold with a Ser3-His6-Asp9 triad configuration. Formula Ac-PP**S**PP**H**PP**D**PPGY-NH2 (S**3**H**6**D**9**).
3) Another polyproline with a Cys3-His6-Asp9 triad configuration. Formula Ac-PP**C**PP**H**PP**D**PPGY-NH2 (C**3**H**6**D**9**).

## 1.6 Research problem and aims

Knowing the geometry and surrounding sequence of active sites is important for understanding enzymes and for creating synthetic active sites.[27,45] However, exact residue geometries and active site composition still remain to be fully explored. Papers that address the fine tuning of these geometries or specific compositions in synthetic peptides are under-represented in the literature. Peptide self-assembly and synthetic biology has, however, advanced significantly in recent years.[46–49] The lack of mentioned parameters is becoming apparent in these rapidly growing fields that now have new self-assembling peptides, advanced scaffolds and peptide synthesis methods but are unable to replicate the high efficiency of natural proteins.[50]

This thesis aims to expand that knowledge by analyzing the geometric and qualitative parameters of active sites in natural enzymes that can be paired to existing synthetic peptide variants from the literature. These parameters are the distances and angles between catalytically active residues as well as the amino acid composition of active sites and neighboring amino acids of active residues. The focus is put on catalytic triad containing enzymes because of their well-researched active mechanisms. The objective is to identify patterns in distances and angles between catalytically active residues in EC 3.1 enzymes which contain triads. The objective is also to identify and characterize the properties of amino acids in the active site and near catalytic residues in all EC 3.1 subclasses of enzymes with

characterized active sites to see what parameters should be considered when designing a synthetic catalytic sequence.

## 1.7 Significance of the research and limitations

This work investigates specific qualitative and geometric parameters for applications in peptide design. In view of the rapidly growing new fields such as peptide self-assembly, *de novo* peptide construction and emerging *in silico* technologies for rational design of biomolecules, this thesis has the scope address the gap in research that exists between what is known to be important for enzyme function and what parameters need to be emulated by synthetic biology.

This research, however, has its limitations. The scope of study was narrow, looking at a subgroup of a small portion of enzymes within a set EC class. This class, and the Rossman fold containing enzymes within the EC 3.1 are relatively rigid in active site conformation and mostly do not change significantly based on substrate binding.[51] The EC 3.1 contains only a limited number of cofactors none of which directly participate in reaction and presents catalytic residues spaced out throughout the sequence. Therefore, it is expected that the methods and conclusions from this work will not apply to other classes of enzymes or even to the different proteins within the EC 3.1 subclass. Our methodology could have benefited from a more in-depth analysis by simulating the analyzed molecules in solution as well as having crystal structures of the catalytic peptides from the literature. In future, the analysis of amino acid properties could be expanded to consider the structural context and the proximity to catalytic residue in space instead of just in sequence. This work outlines important design parameters when formulating new catalytic peptide sequences while also pronouncing how conventions like active sites are difficult to be defined and analyzed.

## 1.8 Structural outline

In the introduction chapter the context and research objectives have been identified and the value of such research argued. The most important factors of enzyme catalysis have been established as geometric and composition parameters of the active site.

In the next chapter the methodologies used to analyze these factors are described. This section is divided into two parts. A part which discusses geometry measurements and the other which discusses qualitative analysis of the active site composition.

In the third chapter the results are reviewed by looking at the key points of these measurements, establishing average values and ranges these factors fall into. This is also divided into two parts; geometry analysis and qualitative analysis.

The fourth chapter places the results into context and debates their significance and reliability in the framework of this thesis.

# 2 Methods

## 2.1 Building a data set

A set of catalytic peptides was taken from the literature, while data on natural enzymes was collected via publicly available databases. For this data collection step, the M-CSA database from EMBL-EBI was chosen.[4] From here the catalytic mechanisms, active site catalytic residues and their roles were copied, as well as links to other databases detailed further in the text.

The 3.1. EC classification of enzymes was filtered out using the "Browse" interface on the site and selecting the respective subclass filter. Using this filter on 27$^{th}$ of November 2020, 96 enzymes were listed on the site.

The M-CSA entries were qualified based on "star" ratings from the site. Enzymes with fully characterized mechanisms and entries without any mechanism were both considered. All of them had X-ray crystallography analysis with 3.0 Å or greater resolution, therefore a resolution cutoff was not necessary.

A dataset of EC 3.1 enzymes was made from the M-CSA list of 96 proteins (called **Dataset 1** or **D1**). The dataset contained M-CSA ID, enzyme name, Uniprot ID, full EC number, PDB ID, enzyme sequence length, the full sequence, position of active site in primary sequence, the sequence of the 'long' active site, catalytic triad/dyad residues, other stabilizers, oxyanion hole residues, metal binding sites, bound metal ions, 'short' active site sequence, species its derived from and CATH number.

The M-CSA site provided the M-CSA ID, enzyme name, Uniprot ID, full EC number, PDB ID, position of active site in primary sequence, species its derived from, what metal ion they use and CATH number, as well as enough information to label active residues into categories of metal binding ligands, if they are base, acid or nucleophile in the catalytic triad or dyad, oxyanion hole or other stabilizer. Protein sequences were copied from Uniprot, via links found on their respective M-CSA entries. The PDB resolutions, EC numbers, CATH numbers, species origin and mechanism distribution were analyzed by counting using the COUNTIF function in Excel.

A dataset containing sequences of **long active sites (LAS)** and **short active sites (SAS)** was built. LAS and SAS were derived from the M-CSA and Uniprot sites and with the use of a text editor to highlight, select and cut out portions of the protein. These were written down into the datasets.

From the EC 3.1 enzymes, a smaller dataset was derived of EC 3.1 enzymes containing only catalytic triads and oxyanion holes (called **Dataset 2** or **D2**). It contains color coded triads, a PDB coordinate column and loop sequences. The colors were used to differentiate between the most common Ser-His-Asp triad and other triad types: the Ser-His-Glu, Cys-His-

Asp and Ser-His-Trp (main-C) triads. The PDB coordinate column contained PDB residue number that could be copy pasted into PyMol, into the "show sticks, resi" command so that the program could find and mark active residues.

Finally, "**4x4 loop**" sequence column was the sequence of amino acids found adjacent to their respective triad member residues. This sequence contained 4 amino acids from each side of the active residue, excluding the active residue itself.

## 2.2 Analysis of composition properties

The analyzed properties were composition subcategories based on amino acid properties. The properties are named; tiny, small, aliphatic, aromatic, non-polar, polar, charged, basic, acidic, sulfur and hydroxylic. These amino acid properties are categorized as seen in Table 1, indicated by amino acids single letter codes. The properties assigned to a certain amino acid are expressed in mole percentage (mol%). The program from which mol% data was gathered from also provides: Cruciani polarity, Cruciani hidrophobicity, Cruciani H-bonding, hydrophobicity, hydrophobic moment, aliphatic index, isoelectic point, net charge, instability index and Boman index data, that was also gathered.

**Table 1 - Mole percentage categories – Classification of amino acids into 11 categories based on their composition**

| | |
|---|---|
| Tiny | (A+C+G+S+T) |
| Small | (A+B+C+D+G+N+P+S+T+V) |
| Aliphatic | (A+I+L+V) |
| Aromatic | (F+H+W+Y) |
| NonPolar | (A+C+F+G+I+L+M+P+V+W+Y) |
| Polar | (D+E+H+K+N+Q+R+S+T+Z) |
| Charged | (B+D+E+H+K+R+Z) |
| Basic | (H+K+R) |
| Acidic | (B+D+E+Z) |
| Sulfur | (M+C) |
| Hydroxylic | (S+T) |

These properties are analyzed by a software that counts the number of amino acids that have a defined characteristic. It then divides the number of residues in the full protein with its number of counted residues. The composition properties mentioned here as 'Tiny', 'Small' or other are based on their side chain (R-group) size and chemical properties and are counted as mol percentage (mol%). The other properties such as Cruciani hydrophobicity, H-bonding and polarity are scored based on residue identity and then calculated from the average score[52]. Similarly, this is done for hydrophobicity, hydrophobic moment, aliphatic index, isoelectic point, net charge, instability index and Boman index. [53,54] The characteristics are calculated from the individual residues.

**Cruciani properties** are regarded as principal amino acid properties; **polarity**, **hydrophobicity** and the ability to form **H-bonds**. The three scales characterize side chains in the sequence based on the interaction of each amino acid residue with several chemical groups (or "probes"), such as charged ions, methyl, hydroxyl groups, and so forth.

**The Bowman index** is equal to the sum of the solubility values for all residues in a sequence, it might give an overall estimate of the potential of a peptide to bind to membranes or other proteins as receptors, to normalize it is divided by the number of residues. A protein has high binding potential if the index value is higher than 2.48.

**Instability index** predicts the stability of a protein based on its amino acid composition, a protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable.

**Isoelectic point** (pI) is the pH at which a particular molecule or surface carries no net electrical charge.

**Net charge** is the function the computes the net charge of a protein sequence based on the Henderson-Hasselbalch equation.

**Hydrophobicity index** is calculated adding the hydrophobicity of individual amino acids and dividing this value by the length of the sequence.

**Hydrophobic moment** is a quantitative measure of the amphiphilicity perpendicular to the axis of any periodic peptide structure, such as the α-helix or β-sheet. It can be calculated for an amino acid sequence of N residues and their associated hydrophobicities.

**The aliphatic index** is defined as the relative volume occupied by aliphatic side chains (Alanine, Valine, Isoleucine, and Leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins.

### 2.2.1 Statistical analysis of generated data

The generated data was analyzed using Excels built in graphs to compare data of active site data to the whole protein. This entire analysis was done to 96 enzymes in dataset 1 and then separately for the 22 proteins in dataset 2 containing triads and oxyanion holes.

#### 2.2.1.1 Mole percentage analysis

Mole percentage (mol%) was presented using two different graphs and tabularly. Next, the mol% data was averaged and a 2D bar graph was created. These were named ‘profiles’ because of the distinctive profile they formed from its bars for each mol% group. This was also done on both D1 and D2, with the exception of an added three profiles for D2. The 4x4 loops for the nucleophile, base and acid roles of triad containing enzymes also had their mol% data averaged and compared to D1 and D2 profiles.

Tabularly the averaged mol% of whole, LAS, SAS and the 4x4 loop sequences are presented. The relevant data was highlighted based on if the mol% difference presented a significant composition change in given amino acid characteristic.

#### 2.2.1.2 Sequence percentage analysis

To further this analysis the LAS and SAS were then divided into categories based on the sequence length percentage of the whole protein they occupy. This meant that LAS that were 90% of its parent enzyme were in a different group than LAS that were a 20% section of the entire parent protein sequence. The general approach was grouping the different enzymes based on a percentage range the enzymes qualified for.

D1 was divided into 5 groups, each group representing a 20% increase in total protein occupancy, starting from 0% to 100% (0-20%, 20-40%, 40-60%, 60-80% and 80-100%). For example, arylesterase, that had a 218 amino acid length LAS and an entire protein length of 359, was categorized into the 80-60% category because 218 out of 359 was 60,72%,

D2 had a much smaller sample size than D1, therefore there were 4 groups. The groups made were: 35-50%, 50-65%, 65-75% and 75-85%.

The SAS sequences of D1 had 5 categories; 4-2,5%, 2,5-2%, 2-1,33% and 1,33-0%, while D2-SAS had 4; 0,6-1%, 1-1,6%, 1,6-2,2%, 2,2-3%.

Following categorization, the sequences were compared tabularly and graphically within each group the same way as described previously for the non-grouped analysis.

Lastly, a sequence analysis was done for the triad containing enzymes. The catalytic triad and oxyanion hole forming residues are numerated by position in sequence. This numeration was subtracted for each combination of the 5-6 residues forming the triads/oxyanion holes within the given enzyme to see how distant in sequence each residue is from one another.

## 2.3 Analysis of the geometry of enzymes and synthetic peptides

### 2.3.1 Measuring distances and angles

The distances and angles between the catalytically active residues of the catalytic triad and oxyanion hole forming residues were measured. This was done using 22 PDB structures from the RCSB PDB web site. These PDBs were linked as a part of the M-CSA database entries. For active site research it is preferable to find apo-structures, proteins that did not contain a bound substrate. However, previous research shown minimal conformational change with substrate binding in alpha-beta fold proteins, of which these triads are members of.[51]

The protein PDBs were opened in PyMol. In this molecular visualization program, the distances between the triad members any oxyanion hole members were measured using the "Measurement Wizard" tool which allows distance and angle measurements via selecting the reference atoms of interest. The wizard tool returns an angstrom value in the case of distance measurement and degrees in case of angle measurement. In each measurement the solvent was removed, all chains except the chain A was deleted, and the cartoon representation was hidden. Atom to atom measurements were done using the same points of refence on all catalytic triads and oxyanion holes.

**Nucleophiles** in the triad were measured from the atom that performs the nucleophilic attack, like the γ oxygen of serine, in example. **The base role** was performed exclusively by histidines in this group of 22 proteins. Histidines δ-nitrogen is the one which forms a hydrogen bond with an acid residue which reduces histidines pKa and allows it to deprotonate the nucleophile. However, in this work the γ-carbon at the stem of the histidine imidazole ring was used as the reference measuring point. The rationale behind this choice is that the stem would be more static and less prone to

His-flipping or crystallography interpretation errors, making measurements more accurate and consistent.

**The catalytic acid** is usually in the form of a carboxylate which can freely rotate and has a fluid negative charge. Therefore, the method of determining which of the oxygen atoms are going to be measured was the acids oxygen atom closest to the histidine δ-nitrogen. Although their distance from the histidine was still measured from the γ-carbon under the imidazole ring.

The **oxyanion hole members** were measured from the nitrogen atoms that created the negative charge for the tetrahedral intermediate stabilization.

The reference points were the same for both distance and angle measurements. The distances measured were in between each of the triad members and between each of the triad members and oxyanion hole members. The angles measured were between each of the triad members as well as between the 2 oxyanion members and the nucleophile or base residues.

### 2.3.2 Statistical analysis of geometry data

Both the angles and distances were further analyzed using the mean, median and interquartile ranges (IQR) calculated by the Whisker-Box plot graphs and by calculating their 3*standard deviations. The IQRs were compared to averages by calculating what percentage of the average measurement is the IQR. The only exclusion from the data was 1-alkyl-2-acetylglycerophosphocholine esterase which was excluded from distance measurements of triad members because it was the outlier.

## 2.4 Synthetic peptide comparison

The resulting analysis yielded an average angle and distance for triad containing enzymes. The average distances and angles were then used to find a representative protein which has the closest angles and distances to

the average, having in mind the priority is distance similarity. This representative enzyme is then used to compare to the catalytic peptide called C3H6D9 which had its catalytic triad distances measured in the literature.[40] The reference points were different for this study, therefore a representative is chosen for D2 enzymes and measured again using the criteria from the C3H6D9 study in PyMol. The synthetic peptide points of measurement were; (1) both of the side chain oxygens of the aspartate to the δ-nitrogen hydrogen of histidine and (2) the cystein sulfur distance from the ε-nitrogen hydrogen of the histidine. However, in the 1agy PDB of cutinase the hydrogen is still attached to the serine oxygen instead of the eta nitrogen of histidine. Therefore, the serine hydrogen was instead taken as the measurement point from the ε-nitrogen of histidine. The distances were compared between the synthetic and natural triads.

# 3 Results

## 3.1 Enzyme data

All proteins used in this thesis have crystal structures characterized by X-ray crystallography and are equal to, or below 3,0 Å resolution and two thirds of all PDBs had a resolution of 2,0 Å or below.

Of 22 proteins in D2, 17 are rated with three stars, two with two stars and two with one star. Three stars denotes a mechanism that is consistent with all evidence in the literature, two stars a mechanism that either does not explain all the evidence or there are other similarly good proposals and one star that has a mechanism which has been contested by more recent data.Of the 96 enzymes, 70 had completely characterized mechanisms in M-CSA, with other proteins having known mechanisms but without having all detailed parts of the reaction. In D1 there are 51 metalloenzymes, 22 triad containing enzymes and about 8 dyad containing enzymes

### 3.1.1 Species, CATH and EC distribution for 96 and 22

Of 96 proteins, most were derived from *E. coli* strains (16,7%), followed by *Homo sapiens* (13,5%), *Bos taurus* (6,3%), *Bacillus subtilis* (5,2%), *Rattus norvegicus* (3.13%) and others (54,2%) of which a large majority have one protein derived from that species. The distribution of species changed when analyzing only the 22 triad containing EC 3.1 enzymes that have mostly proteins derived from *Bacillus subtilis* (18,2%), followed by *Bos taurus* (13,6%) while the remaining 68,2% have a single protein per species. D1 contains many different CATH

numbers, however, there is one CATH repeated 17 times in D1. This is the 3.40.50.1820 alpha/beta hydrolase fold. All the 17 proteins are also part of the D2 containing triad enzymes. Other well represented folds are 3.30.420.10, 3.30.540.10 and 3.40.190.80 joint fold and 3.40.50.1110 domains. All of the three CATH domains are repeated four times in D1. In D2, there are only three CATH domains; the previously mentioned 3.40.50.1820, the 3.40.50.1110 and 3.40.50.180. The latter two are represented in only 5 proteins, four by 3.40.50.1110, and one by 3.40.50.180.

**Table 2 - CATH number distribution in all EC 3.1 enzymes (D1).**

| CATH number | # proteins with CATH |
|---|---|
| 3.40.50.1820 | 17 |
| 3.30.420.10 | 4 |
| 3.30.540.10 and 3.40.190.80 | 4 |
| 3.40.50.1110 | 4 |
| 3.10.450.30 | 3 |
| 3.20.20.190 | 3 |
| 3.40.720.10 | 3 |
| 3.60.21.10 | 3 |
| 3.90.190.10 | 3 |
| Other | 50 |

**Table 3 - The EC 3rd and 4th level distribution in EC 3.1 enzymes (D1).**

| EC | # proteins with EC |
|---|---|
| 3.1.-.- | 5 |
| 3.1.1.- | 4 |
| 3.1.1.3 | 4 |
| 3.1.21.4 | 4 |
| 3.1.1.4 | 3 |
| 3.1.3.16 | 3 |
| 3.1.3.2 | 3 |
| 3.1.3.48 | 3 |
| Has 2 proteins with same EC | 6 |
| Has 1 protein with same EC | 54 |

EC numbers in D1 are distributed with nine of the entries that do not have a known 3rd or 4th level EC classification, EC 3.1.1.3 and 3.1.21.4 have four proteins in the dataset, EC 3.1.1.4, 3.1.3.16, 3.1.3.2 and 3.1.3.48 classify three proteins each and the rest classify either 2 proteins (6 ECs) or 1 protein (54 ECs) per EC classification.

## 3.2 Full protein, LAS and SAS sequences

In D1, protein sequences ranged from 130 to 756 amino acids, with an average of 342 amino acids. The average size of the LAS is 163 and that of SAS is 6 amino acids. LAS and SAS have a range of 3-373 and 3-12 amino acids, respectively.

In D2, protein sequences ranged from 208 to 637 amino acids, with an average of 363 amino acids, 21 amino acids larger than D1 proteins on average. The average size of the LAS and SAS is 216 and 5,3 amino acids and have a range of 123-332 and 3-7 amino acids, respectively. This is a 58 residue larger average sequence size of dataset 2 LAS (D2-LAS) compared to dataset 1 LAS (D1-LAS).

### 3.2.1 Building long, short and loop active sites

The 'long' active site (LAS) is a sting of amino acid residues that starts on the first and ends on the last catalytically active amino acid residue in the primary sequence. This includes the active residues and all residues in between them as shown in Figure 3.

The other definition used was the 'short' active site (SAS) which contained a sequence of active residues combined in the order in which they appeared in the original protein sequence (Figure 3).

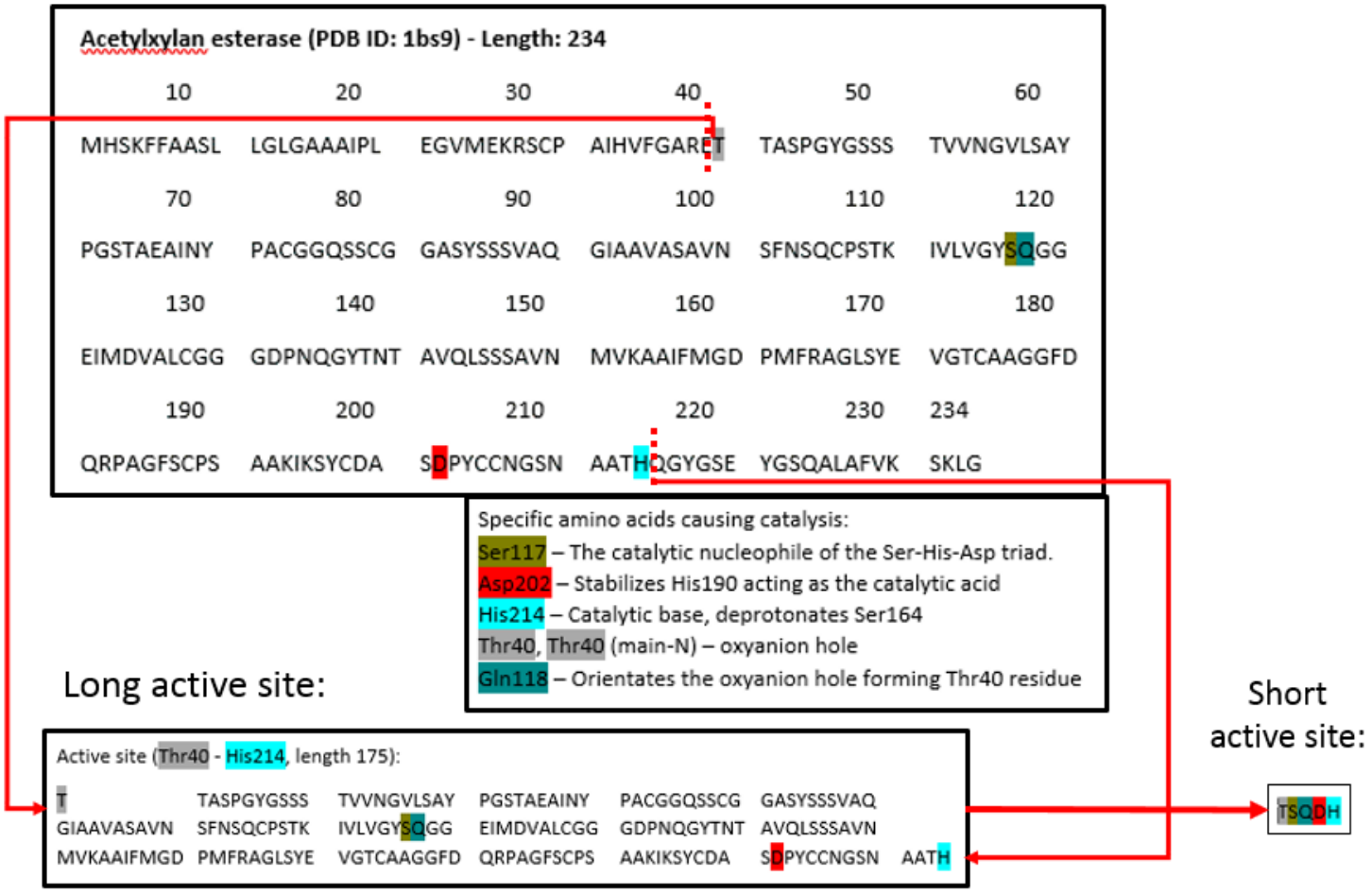


**Figure 3 - Construction of the 'long' and 'short' active site – The red lines represent where the sequence was cut, The sequence was taken from Uniprot while the M-CSA database provided the description of the catalytic amino acids.**

LAS and SAS were generated using the 96 enzymes from the M-CSA database by copying their sequence from Uniprot. The catalytically active residues were highlighted based on what residue was mentioned as catalytically active in the M-CSA database. The SAS sequence was generated by cutting the non-catalytic residues out of the LAS sequence resulting in only catalytic residues adjoined together in sequence respective to how they are numerated in the protein.

The third definition of an active site was used to include 4 amino acids on each side of the active residue. This active site definition was applied only to catalytic triad containing EC 3.1 enzymes and the sequences were built

only for the 22 triad containing enzymes. This was done by finding the three triad residues that have the nucleophile, acid and base roles in the PDB crystal sequence using the previously made PDB coordinates column. The coordinates were copied into the "color red, resi" command in PyMol. The 4 residues before and after the red colored residues were copied into the dataset containing the base and acid and nucleophile loop sequences.

### 3.2.2 Composition data of active site sequences

The compositional and qualitative characteristics of peptide sequences and proteins were analyzed using a web-based tool DeShPet (visited: 2.7.2021., https://deshpet.riteh.hr/). This data was then matched to their respective enzymes in D1 and D2 using the 'sort' tool in Excel.

The sequence data generated included all 96 EC 3.1 enzymes (D1) found in M-CSA and their respective full protein sequence, LAS and SAS. Composition data was also generated for the 4x4 loop sequences, however this step was only reserved for the 22 triad containing enzymes (D2), as previously mentioned, and the data contained only mol% categories.

### 3.2.3 Mole percentage data range, profile and averages

The smallest mol% value for all mol% data gathered is for a sulfur mol% (1,88%) and the highest value is for small mol% (67,52%) for the full protein sequences of D2. For D1, there is a 1,23% sulfur mol% for the minimum value, with the same small mol% maximum as in D2.

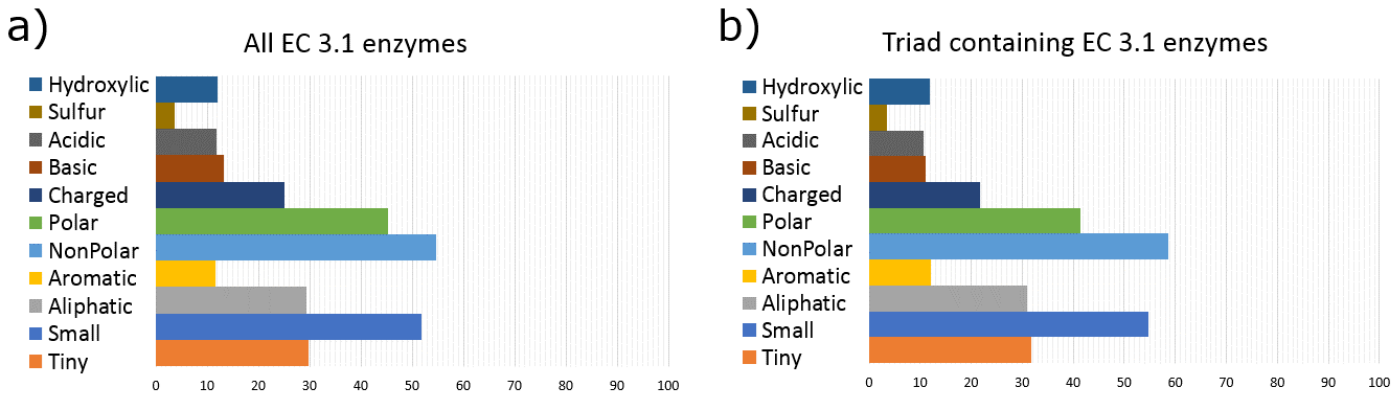


**Figure 4 – EC 3.1 and triad containing EC 3.1 enzyme profiles – graphs showing the distribution of mol% data across different amino acid parameters in the a) EC 3.1 protein dataset (D1) and b) EC 3.1 proteins with a catalytic triad and oxyanion hole (D2).**

The profiles for averaged mol% for D1 and D2 datasets are shown in Figure 4. The D1-LAS sequences have several minimums of 0 mol% and the highest a 75% for a nonpolar mol% in the large unaveraged dataset. The D2-LAS sequences have a minimum 1,23% in sulfur mol% and a maximum value of 73,71% for small mol%. The SAS have many 0 mol% minimums and 100 mol% maximums in both datasets.

The averages of mol% data for full protein, LAS and SAS sequences are shown in Table 4 and in profiles shown in Figure 4, Figure 5 and Figure 6.

The largest differences between D1 and D2 profiles can be seen in the polar/nonpolar mol% difference of 3,7% when subtracting the two percentages and looking at their respective profiles. D1 has 45,24% polar mol% and the triad has 41,49%.

**Table 4 – Distribution table showing averages of mol% properties of: all EC 3.1 enzymes, only EC 3.1 enzymes containing a catalytic triad, all LAS and SAS, only LAS and SAS from enzymes that contain a triad.**

| Mol% | EC 3.1. enzymes (D1) | EC 3.1. with triad (D2) | Long active sites (D1-LAS) | Short active sites (D1-SAS) | Long - triad containing (D2-LAS) | Short - triad containing (D2-SAS) |
|---|---|---|---|---|---|---|
| Tiny | 29.87 | 31.84 | 29.67 | 16.08 | 32.15 | 34.74 |
| Small | 51.82 | 54.71 | 52.08 | 46.57 | 54.99 | 53.92 |
| Aliphatic | 29.37 | 30.91 | 27.17 | 4.23 | 30.31 | 9.68 |
| Aromatic | 11.54 | 11.99 | 13.03 | 26.89 | 12.53 | 29.89 |
| NonPolar | 54.76 | 58.51 | 53.88 | 16.42 | 58.12 | 30.41 |
| Polar | 45.24 | 41.49 | 46.12 | 83.58 | 41.88 | 69.59 |
| Charged | 25.13 | 21.63 | 25.67 | 66.42 | 21.95 | 38.14 |
| Basic | 13.31 | 11.05 | 13.11 | 31.81 | 11.08 | 19.52 |
| Acidic | 11.82 | 10.59 | 12.56 | 34.62 | 10.86 | 18.62 |
| Sulfur | 3.72 | 3.49 | 3.64 | 2.58 | 3.25 | 3.38 |
| Hydroxylic | 11.94 | 11.93 | 12.11 | 9.27 | 11.46 | 22.36 |

Full proteins compared to D1-LAS have no larger difference than 2,2% while the same comparison of full to D2-LAS yielded no greater difference than 0,6%.

All SAS mol% data varied between 1,13% at a minimum to 41,29% at maximum from the full protein in D1 and 0,11% to 28,1% from the full proteins in D2.

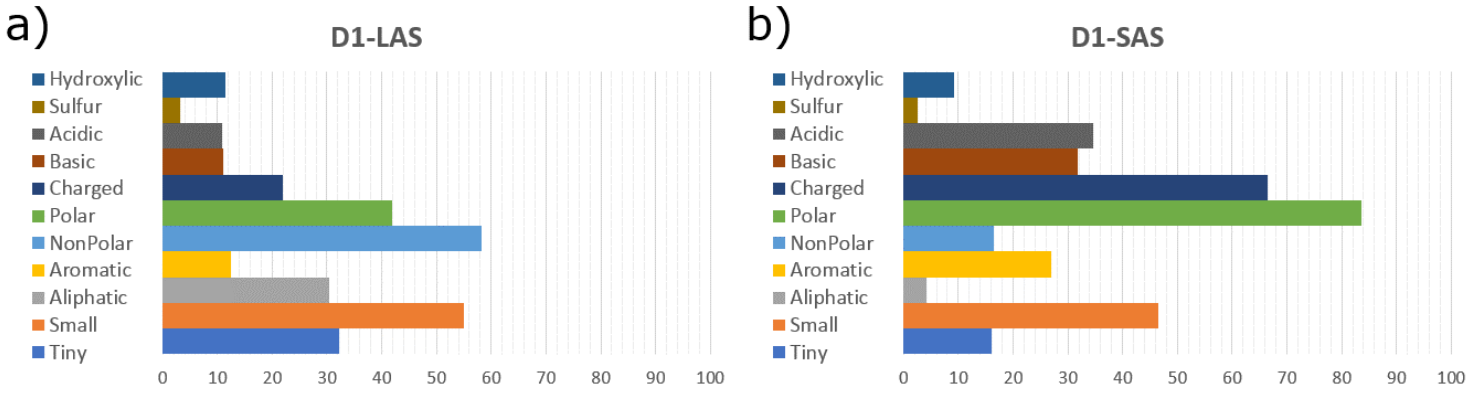


**Figure 5 – EC 3.1 (D1) active site profiles – LAS (a) and SAS (b) mol% data distribution from enzymes in the EC 3.1 subclass dataset. The x axis ranges from 0-100 mol%.**

D1-SAS showed an increase in aromatic charged, basic and acidic mol%, and a decrease in tiny, small, aliphatic, aromatic, sulfur and hydroxylic mol% when compared to D1. D2 to D2-SAS comparison was largely similar to its D1 equivalent, except for having an increase in tiny and hydroxylic mol% instead of a decrease as well as having a different distribution of mol% changes. For example, the charged mol% was increased by 41,29% in D1 comparison while only being increased by 16,51% in D2 comparison. This charged characteristic is also the largest difference in D1 comparison,

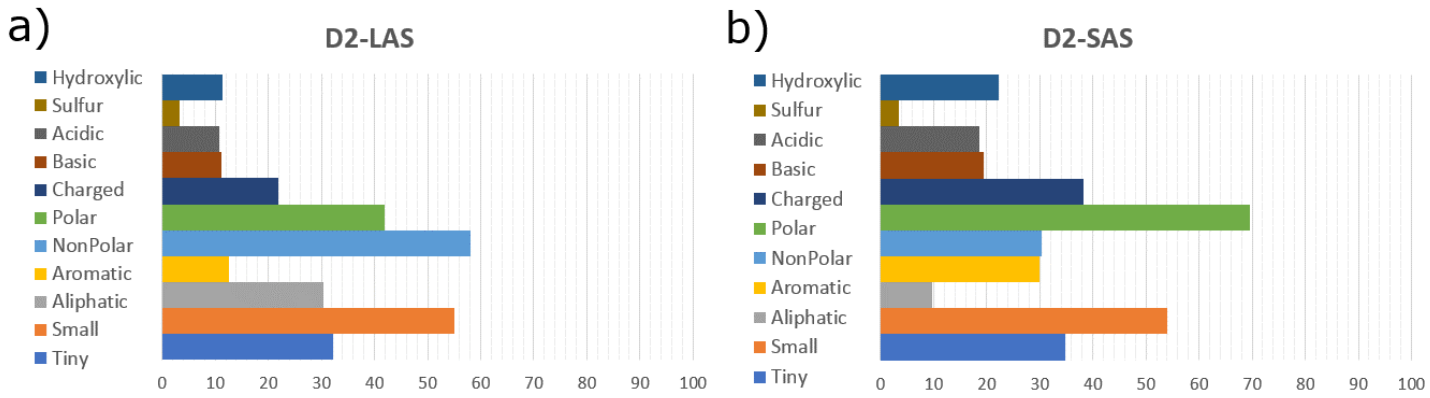


**Figure 6 – Triad containing EC 3.1 (D2) active site profiles – LAS (A) and SAS (B) mol% data distribution from enzymes in the EC 3.1 subclass that contain catalytic triad and oxyanion hole. The x axis ranges from 0-100 mol%.**

while the 28,1% increase in polar mol% was the largest change in the D2 comparison.

Comparing D1-LAS to D2-LAS the largest difference in mol% is in polar and charged parameters, while having almost identical sulfur and hydroxylic mol%. The D1-SAS compared to D2-SAS has large differences in tiny, polar, charged, basic, acidic and hydroxylic mol% characteristics while retaining similar sulfur and aromatic characteristics.

The last part of the analysis yielded Whisker-Box plots for D1 mol% data. This plot allows us to see the distribution and range of data.

As seen in Figure 7 the IQRs of SAS are the largest and a slight increase in IQRs in LAS is seen in comparison to the full protein sequence (referred to as "Full" in the graph). This Whisker-Box plot also shows a relatively larger IQR in tiny, small, charged, polar/non-polar and aliphatic mol% than to the smaller sulfur, acidic, aromatic, hydroxylic and basic mol% in the LAS and full sequences.

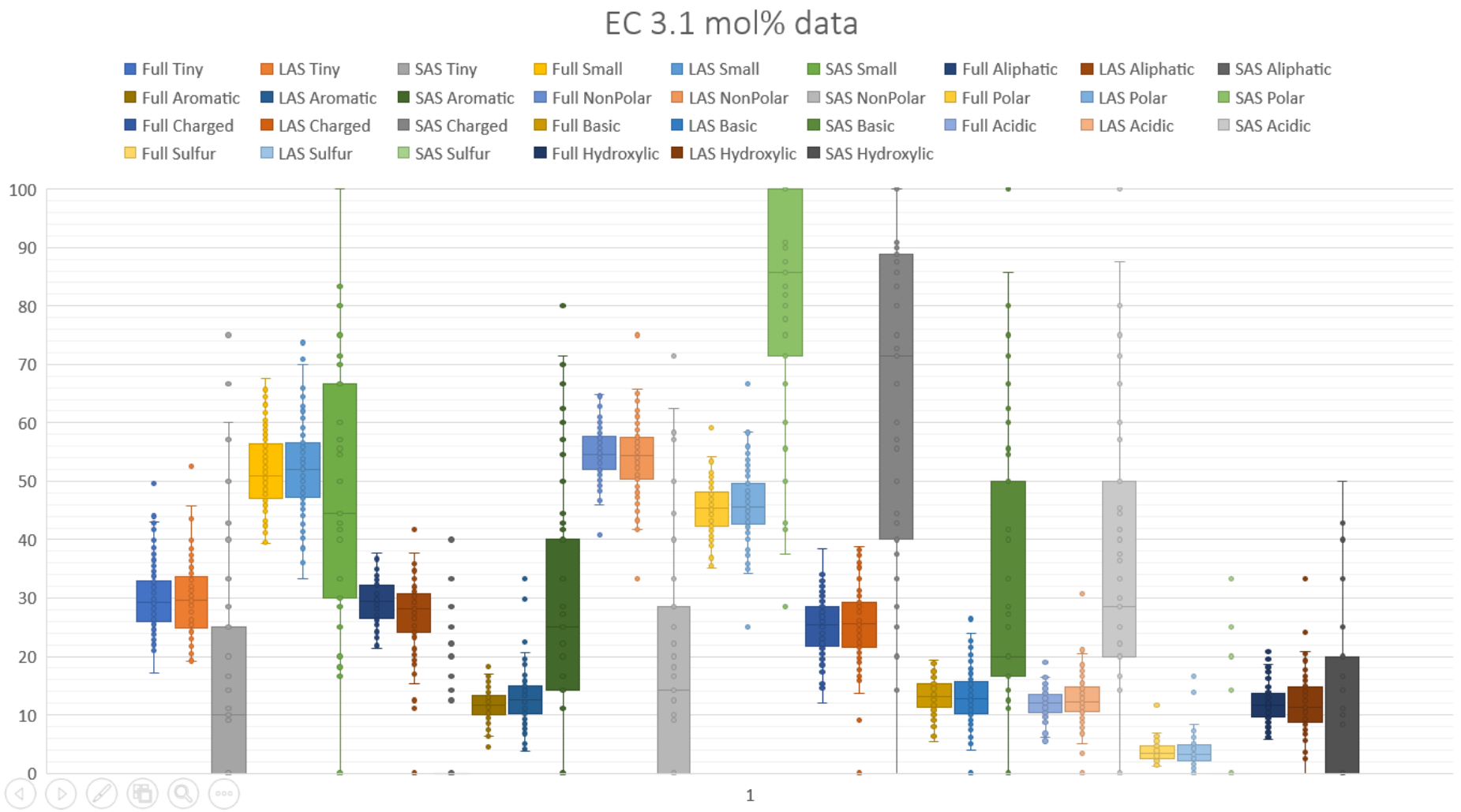


**Figure 7 – EC 3.1 mol% Whisker-Box plot showing the overall distribution of data throughout mol% categories in D1.**

### 3.2.4 Size percentage distribution

Different LAS sequences occupied varying percentages of the corresponding full protein. These percentages were divided into groups; 5 groups for enzymes in D1 (0-20%, 20-40%, 40-60%, 60-80% and 80-100%) and in 4 groups for the enzymes in D2 (35-50%, 50-65%, 65-75% and 75-85%). The distribution of full protein percentage from D1; 0-20% has 17, 20-40% has 16, 40-60% has 23, 60-80% has 33 and 80-100% contains 5 LAS sequences in their respective groups. While these are relatively equally distributed the other 3 groups; D2-LAS, D1 and D2 SAS are not. D1-SAS have 30 sequences in the 1,99-2,73% range, 27 in the 1,25-1,99% range and 20 in the 0,51-1,25% range. The rest of the 19 sequences are distributed in the range from 2,73-5,48%. Shown differently, a 2,5% range difference from 0,0-2,5% accounts for 72 proteins (75% of the entire enzyme dataset) while the same range difference from 2,5-5,0% contains 24 proteins (25% of the entire dataset).

Similarly, D2-LAS data started from a minimum of 35,24% and had 6 proteins in the 35-55% range, while the same 20% range from 60-80% accounted for 14 proteins out of 22. D2-SAS data contained 7 proteins from the 0,61-1,40%, 11 proteins from 1,40-2,19%, but only 4 proteins from 2,19-2,98%.

In summary, these distributions are such that LAS sequences have, on average, larger full protein percentage (60-100%) while SAS sequences have a smaller one (0-2,5%).

### 3.2.5 Sequence percentage LAS and SAS data

This section shows data gained by analyzing LAS and SAS by categorizing them based on what percentage of the full sequence they occupy. For simplicity, here are shown only the largest and smallest percentages of

these weighted LAS and SAS sequences. The data is shown in Table 5 next to the full protein mol% data for reference.

**Table 5 – Long and short active sites by sequence percentage. This table shows LAS and SAS divided int categories based on what percentage of the entire protein do they occupy.**

| Mol% | EC 3.1. enzymes (D1) | Long: 80-95 seq% | Long: 2-20 seq% | Short: 0-1,33 seq% | Short: 2,5-4 seq% |
|---|---|---|---|---|---|
| Tiny | 29.87 | 26.10 | 29.27 | 21.04 | 14.08 |
| Small | 51.82 | 48.36 | 51.14 | 49.95 | 46.64 |
| Aliphatic | 29.37 | 29.10 | 23.62 | 4.68 | 4.92 |
| Aromatic | 11.54 | 12.60 | 13.00 | 26.45 | 28.91 |
| NonPolar | 54.76 | 56.28 | 50.55 | 19.63 | 20.15 |
| Polar | 45.24 | 43.72 | 49.45 | 80.37 | 79.85 |
| Charged | 25.13 | 26.72 | 25.94 | 66.42 | 68.07 |
| Basic | 13.31 | 13.71 | 12.62 | 31.45 | 30.80 |
| Acidic | 11.82 | 13.01 | 13.32 | 34.97 | 37.27 |
| Sulfur | 3.72 | 3.87 | 3.63 | 2.32 | 3.86 |
| Hydroxylic | 11.94 | 8.90 | 13.82 | 8.59 | 7.89 |
| Mol% | EC 3.1. with triad (D2) | Long: 75-85 seq% | Long: 35-50 seq% | Short: 0,6-1 seq% | Short: 2,2-3 seq% |
| Tiny | 31.84 | 29.41 | 31.00 | 40.00 | 23.57 |
| Small | 54.71 | 51.99 | 55.35 | 51.33 | 47.14 |
| Aliphatic | 30.91 | 29.47 | 29.25 | 14.67 | 17.14 |
| Aromatic | 11.99 | 12.76 | 12.84 | 38.00 | 22.14 |
| NonPolar | 58.51 | 57.92 | 57.70 | 34.00 | 34.29 |
| Polar | 41.49 | 42.08 | 42.30 | 66.00 | 65.72 |
| Charged | 21.63 | 23.30 | 23.13 | 44.00 | 40.72 |
| Basic | 11.05 | 11.57 | 11.46 | 22.00 | 18.57 |
| Acidic | 10.59 | 11.74 | 11.67 | 22.00 | 22.14 |
| Sulfur | 3.49 | 3.80 | 3.57 | 0.00 | 8.57 |
| Hydroxylic | 11.93 | 10.11 | 11.12 | 22.00 | 15.00 |

In summary, in previous analysis the largest difference in D1 and D2 enzymes is polar/nonpolar mol% (3,7%). When comparing D1 and D2 LAS the maximum difference observed was in polar/non-polar mol% (4,24%) ) LAS groups are divided into 4 or 5 groups and the difference between the smallest and largest group might show an amino acid environment preference near active residues. This is because a reduction in active site size means less surface that is not close to active residues.

Comparing D1-LAS in the 0-20% and 80-100% group has given a single outlying percentage of 5,72% in polar/non-polar mol%. D2-LAS comparisons result in only a 3,36% difference in the small mol% characteristic. D2-LAS sequences have differences mostly under 1 mol% while the majority of D1-LAS differences are under 3 mol%. This difference is similar to the difference between D1 and D2 full sequences.

When cross-comparing the D2-LAS to D1-LAS an increase is seen towards small, tiny and aliphatic with a decrease in charged and polar parameters across all sequence percentage groups with the highest differences being in comparing small sequence percentage (+7,14% for non-polar, +5,63% aliphatic and +4,21% small mol%).

SAS sequences contain only active residues. These residues will differ in D1 enzymes mostly based on catalytic mechanism being a metal enzyme, dyad, triad or other, while in D2 they differ mostly based on the type and number of oxyanion hole members which can be two or three. A factor is also the size of the enzymes which categorize SAS into their respective groups. The largest differences across all SAS groups were in tiny mol% in D1 (6,96%), while SAS differences in D2 were the largest in charged, tiny and aromatic mol% with large differences in hydroxylic and sulfur mol%, showing cysteine active residues non-existent in small sequence percentage SAS and replaced with hydroxylic ones. D1-SAS have smaller differences between larger sequence percentage gaps while having larger control to SAS differences than D2 enzymes.

### 3.2.6 Other data analyses

Tested parameters include Cruciani polarity, Cruciani hidrophobicity, Cruciani H-bonding, hydrophobicity, hydrophobic moment, aliphatic index, isoelectic point, net charge, instability index and Boman index. These parameters help put the protein composition in context.

The resulting Whisker-Box plots show the Cruciani properties detailing polarity, hydrophobicity and H-bonding as scored by Cruciani et. al. 2004 (Figure 8). In addition, Boxman index (Figure 11), instability index (Figure 10), isoelectric point and net charge (Figure 9) are presented.

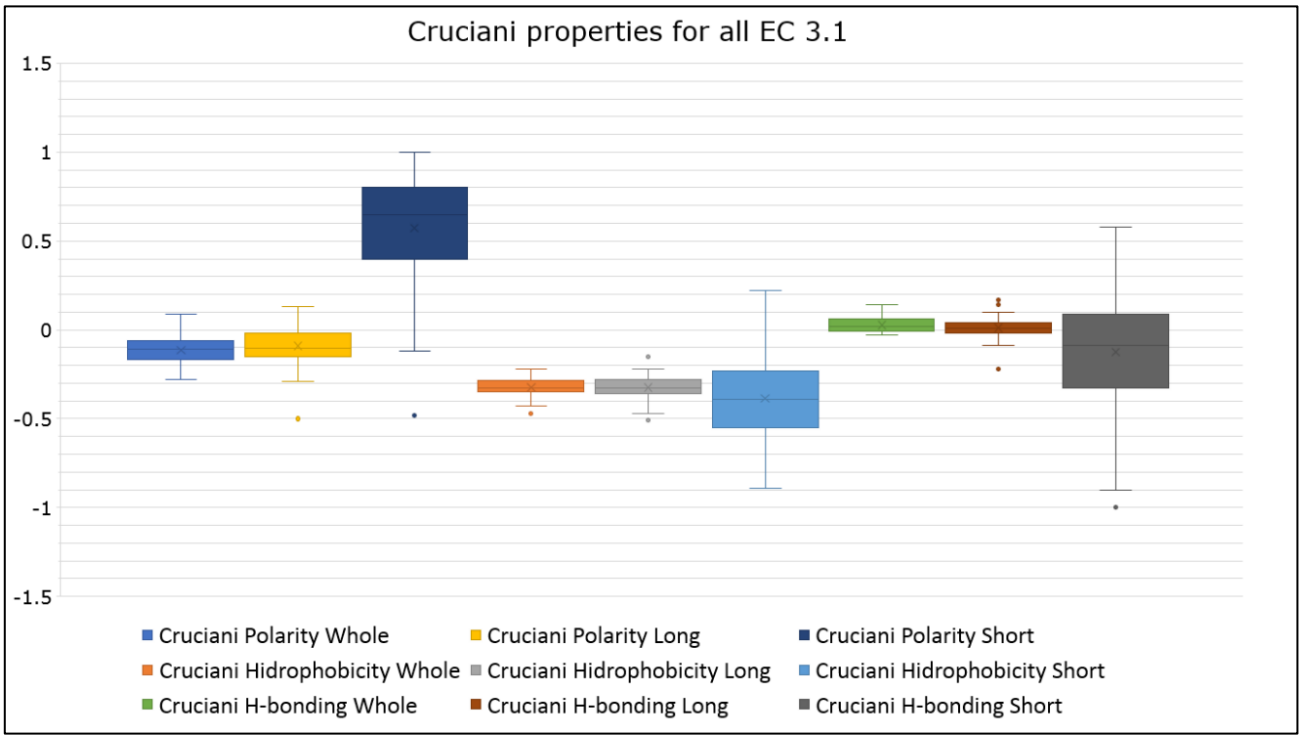


**Figure 8 – Cruciani properties of EC 3.1: polarity (left), hydrophobicity (middle) and H-bonding (right) of dataset 1. The full sequence is represented as "whole" , LAS as "long" and SAS as "short".**

In the data presented in the Whisker-Box plots a clear deviation is seen in Cruciani polarity of the SAS sequence with large IQRs for other SAS sequence Cruciani properties.The isoelectric point decreases in pH from full to LAS and towards the SAS sequence both in median and mean, while retaining similar values for both maximum and minimum ranges.

Net charge IQRs are lowest in SAS, from -4 to 2, and highest in the full sequence with a range of -28,6 to 20,5 while the median and mean remain between -5 and 0 net charge across all three groups. The plot shows that most proteins are in the negative net charge region.

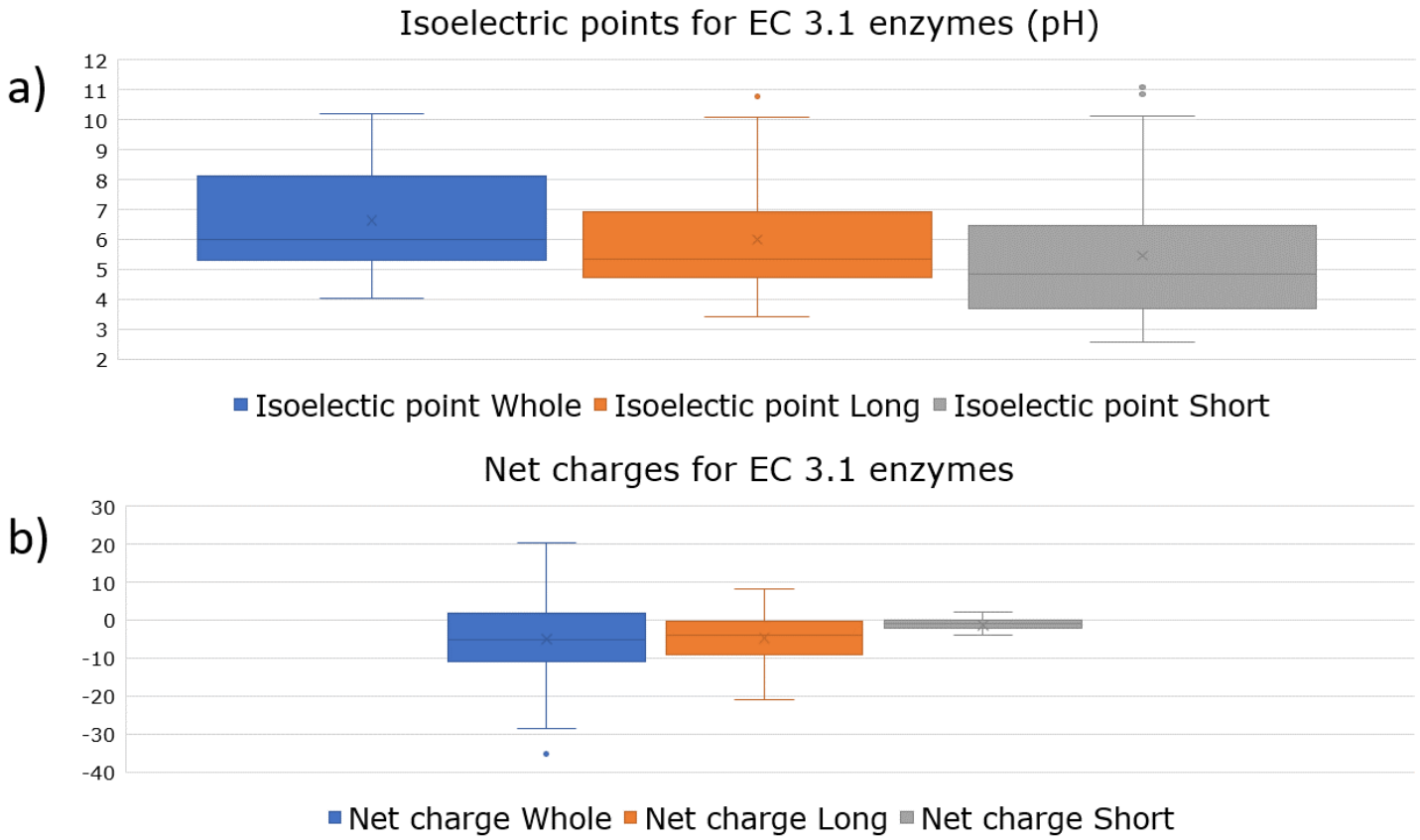


**Figure 9 – Isoelectric point and net charge graphs –showing the pI (A) and net charge (B) of dataset 1. The sequences used were the full protein sequence ("Whole", left), LAS ("Long", middle) and SAS ("Short", right) of dataset 1 enzymes.**

Instability index shows most proteins in the range are under the 40 score value except for the SAS sequence group indicating that the stability decreases with decreasing the sequence length.

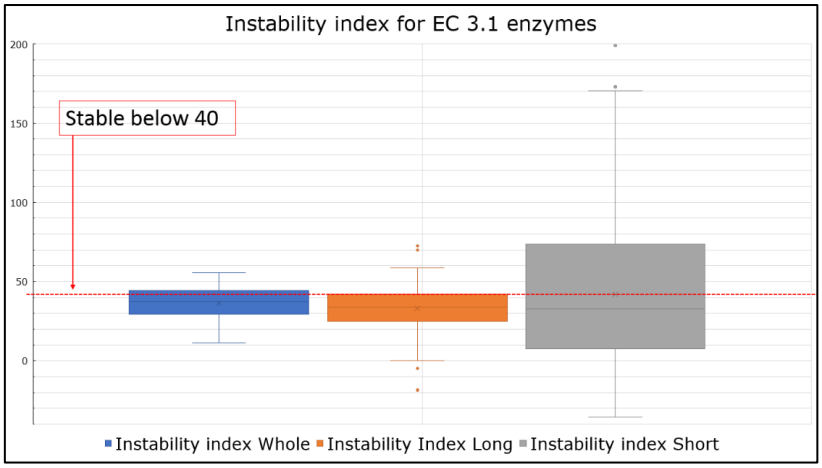


**Figure 10 – Instability index of EC 3.1 enzymes – This graph shows if the protein sequence is stable using an arbitrary score based on the amino acid composition. The graph represents the full sequence ("Whole", left), LAS ("Long", middle) and SAS ("Short", right) of dataset 1 EC 3.1 enzymes.**

Bowman index shows all full protein and LAS sequences are under the 2,48 value while many SAS sequences are above said value.

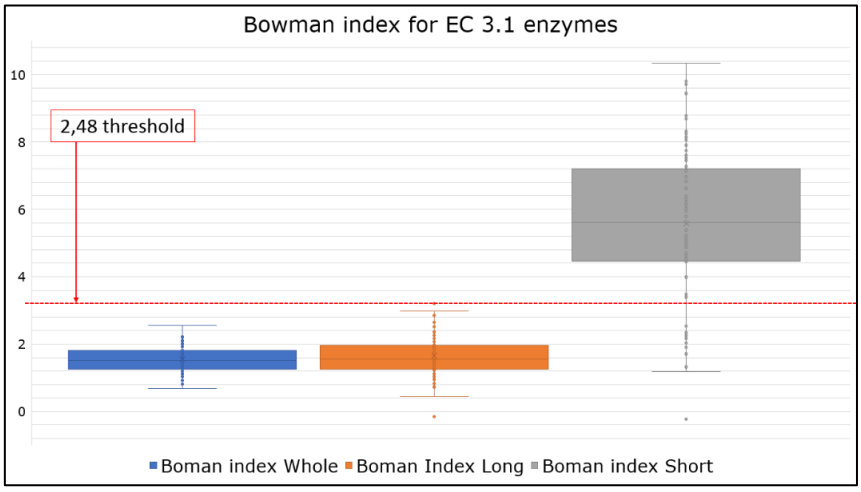


**Figure 11 – Bowman index of EC 3.1 enzymes showing solubility values of residues from the sequence that grade if the sequence has a chance to bind to other proteins. The table contains sequences of the full enzyme ("Whole", left), LAS ("Long", middle) and SAS ("Short", right) from dataset 1.**

### 3.2.7 Catalytic loop mol% data

The catalytic loop of the acid and base residues as well as the nucleophilic elbow shown distinct mol% profiles when analyzed via the 4x4 method. The 4x4 loops have their active residues excluded from the analysis as the aim was not to confirm the properties of catalytic residues, but to discover the

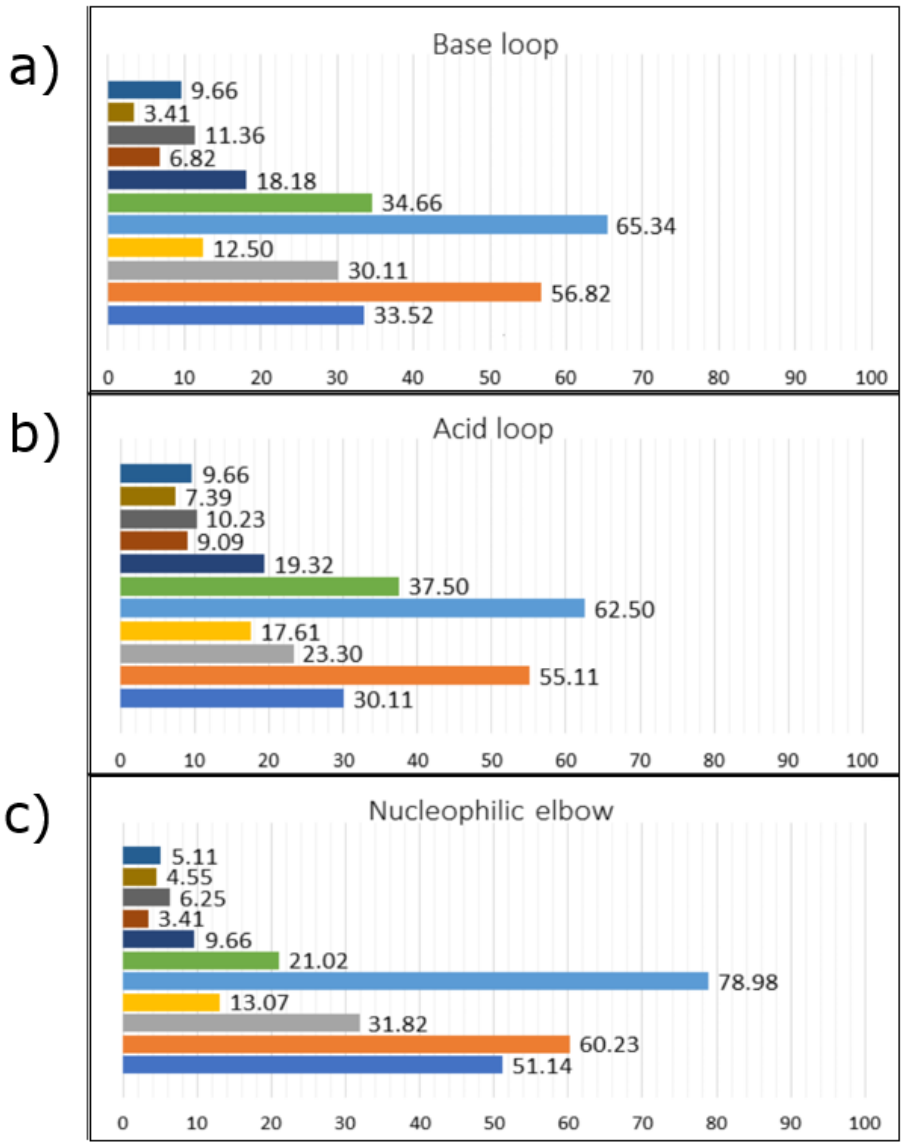


**Figure 12 - Profiles of catalytic loops tested for EC 3.1 triad enzymes: a) Base loop, b) Acid loop and c) Nucleophilic elbow. The graphs highlight the differences in mol% near catalytically active residues.**

properties of their surrounding amino acids. The main differences between mol% are shown in Figure 12 and highlighted in Table 6 in yellow.

The base loops contain higher nonpolar (+6,8%), lower basic (-4,2%) and charged (-3,5%) mol% than D2 data has.

The acid loops have an increase of aromatic (5,62%) and nonpolar (+4%) residues. The sulfur mol% in the acid loop is on average double that of a standard sequence, 7,39% compared to the control which is 3,49%.

The nucleophilic elbow has the most pronounced difference when compared to the average composition full D2 sequences having increased tiny (+19,3%) and nonpolar (+20,5%) mol% and decreased charged (-12%), basic (-7,6%), hydroxylic (-6,8%) and acidic (-4,3%) mol%.

**Table 6 - Catalytic loop mole percentages - The highlighted yellow cells show mol% that are significantly different from the control.**

| Mol% | Control | Base seq | Nuc seq | Acid seq |
|---|---|---|---|---|
| Tiny | 31.84 | 33.52 | 51.14 | 30.11 |
| Small | 54.71 | 56.82 | 60.23 | 55.11 |
| Aliphatic | 30.91 | 30.11 | 31.82 | 23.30 |
| Aromatic | 11.99 | 12.50 | 13.07 | 17.61 |
| NonPolar | 58.51 | 65.34 | 78.98 | 62.50 |
| Polar | 41.49 | 34.66 | 21.02 | 37.50 |
| Charged | 21.63 | 18.18 | 9.66 | 19.32 |
| Basic | 11.05 | 6.82 | 3.41 | 9.09 |
| Acidic | 10.59 | 11.36 | 6.25 | 10.23 |
| Sulfur | 3.49 | 3.41 | 4.55 | 7.39 |
| Hydroxylic | 11.93 | 9.66 | 5.11 | 9.66 |

## 3.3 Catalytic residue sequence analysis

The amino acid distance between two active residues was considered as an important parameter. The average distances of the nucleophile to base distance is the largest averaging at 151,18 amino acids. While the nucleophile-acid distance is 114,86 and base-acid distance 45,05. On average, the oxyanion hole members are closest to the catalytic nucleophile (45,02 amino acids) then to the catalytic acid (132,75 amino acids) and lastly to the catalytic base (173,30 amino acids).

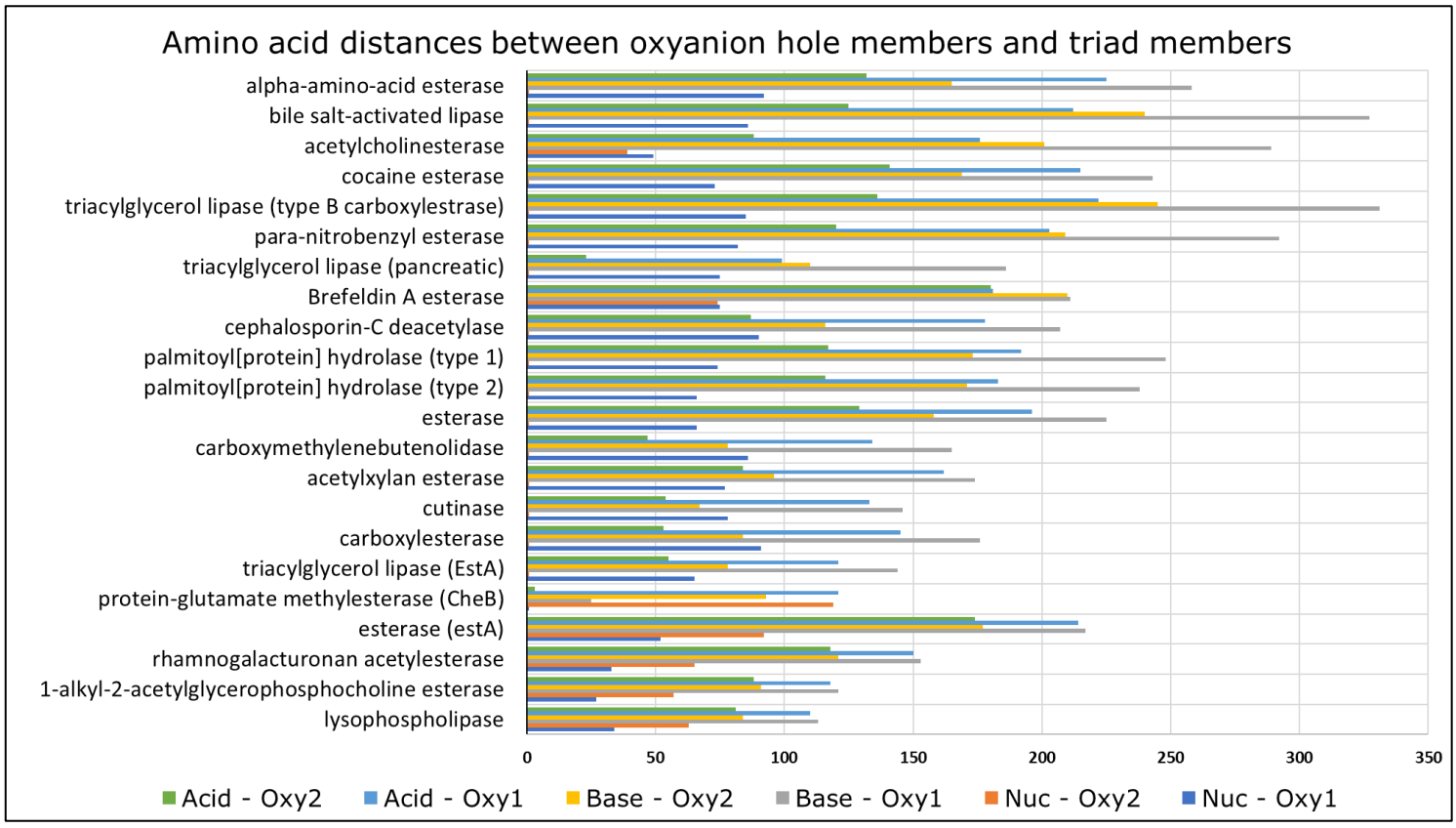


**Figure 13 – Amino acid distances between catalytically active residues – Y-axis shows protein name and X-axis shows residue distance in number of amino acids.**

The results of this analysis also show that 16 out of the 22 proteins had one of the oxyanion hole members directly adjacent to the catalytic nucleophile. When the residue number, which represents its placement in the amino acid sequence, is examined, on average 63,05% of the full protein sequence falls between two most distant catalytically active

residues. The amino acid distances are represented by a bar graph in Figure 13.

## 3.4 Geometry of active sites

### 3.4.1 Catalytically active residue distances

The distances between catalytic triad members were measured to see if they consistently contain similar lengths. The distance between the nucleophiles γO and acids δO is on average 7,2 Å and the interquartile range (IQR) is 0,7 Å, which is 10% of the overall distance average.

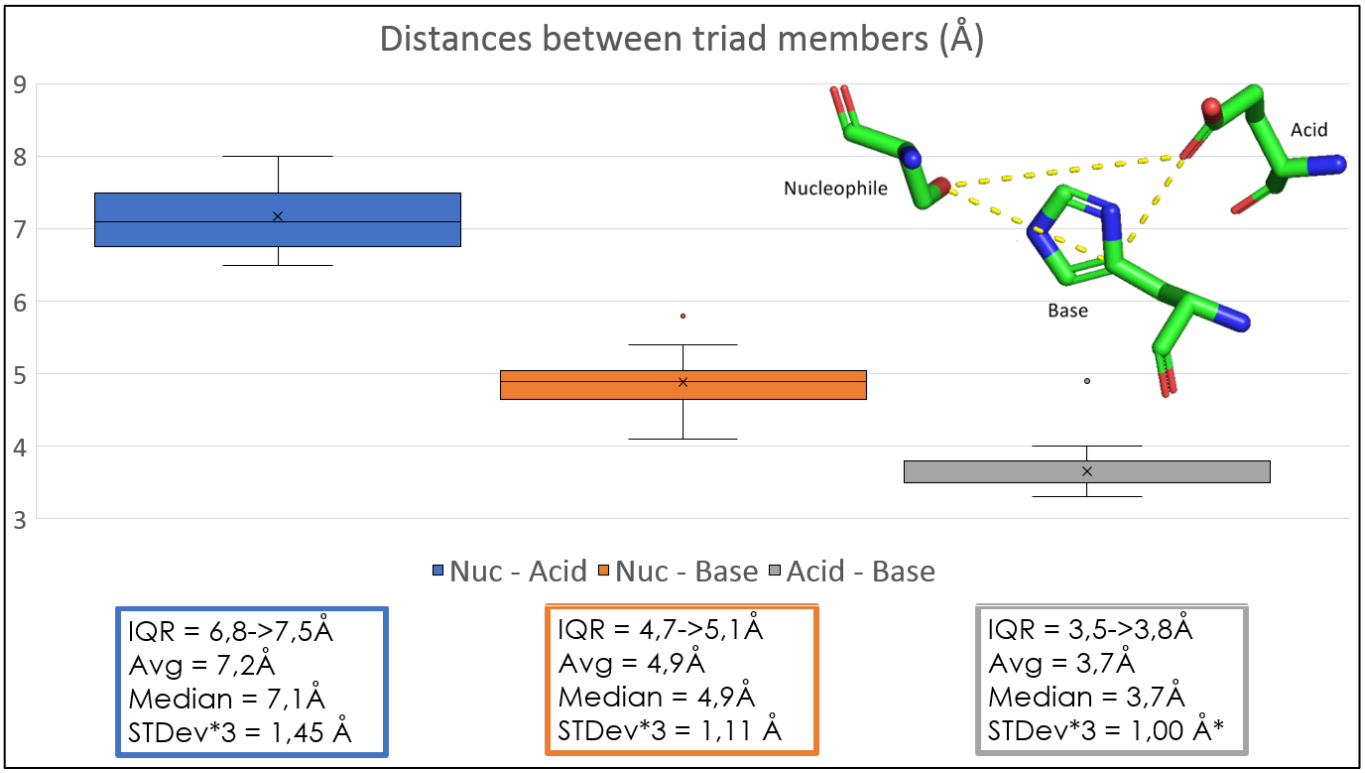


**Figure 14 - Distances between triad members - All measurements are in angstroms. The picture to the top right corner depicts the measuring points, the Whisker-Box plot graphs represent all of the data measured and below it the statistical analysis of the data. Blue (left) represents the nucleophile to acid distances, orange (middle) the nucleophile to base distances and grey (right) the acid to base distances.**

Furthermore, 1,45 Å is the calculated 3 times standard deviation which contains 99,7% of data (3σ further in the study).

From the nucleophiles γO to the bases γC distances are on average 4,9 Å and the IQR is within 0,4 Å with a 1,11 Å 3σ. For the acid δO to base γC distance the average is 3,7 Å, IQR is 0,3 Å and the 3σ is 1,0 Å (Figure 15)

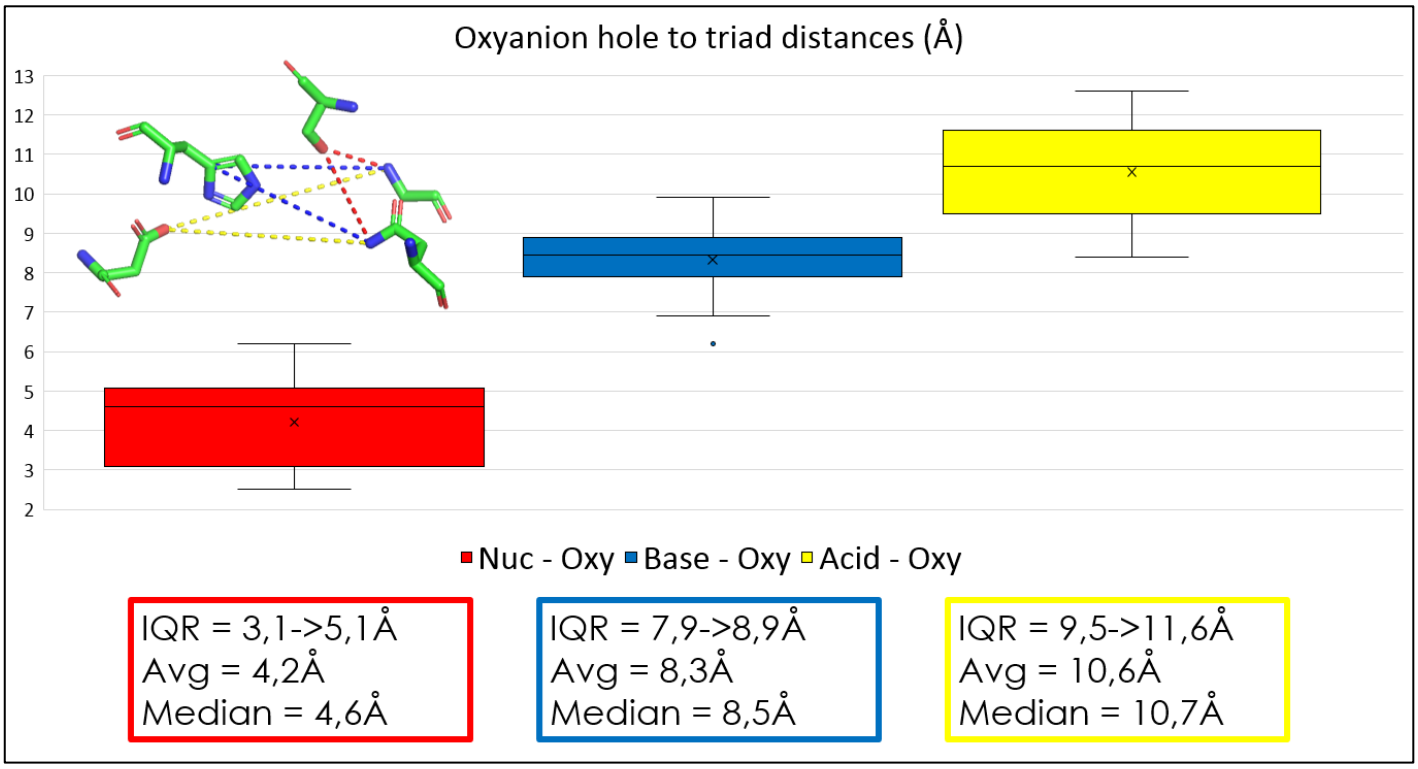


**Figure 15 – Oxyanion hole member distances to triad members - The amino acid residues in top left corner show reference measurement points on the catalytic triad members. The measurements are in angstroms (Å) and the statistical data is presented directly below respective distance Whisker box. The distances from the oxyanion hole members were measured against the catalytic nucleophile (red, left), base (blue, middle) and acid (yellow, right).**

Distances between oxyanion hole members and triad members were also analyzed to see if they would show consistent results. The distance between the nucleophile and oxyanion holes are usually the smallest in distance because many 3.1 esterases have a side chain amine directly neighboring the catalytically active nucleophile. The average distance from nucleophiles is 4,2 Å, while bases have an 8,3 Å and acids a 10,6 Å average distance. The base to oxyanion hole is, in contrast, the most consistent with an IQR

of 1 Å while nucleophile to oxyanion is 2 Å and acid to oxyanion is 2.1 Å in IQR.

### 3.4.2 Angles between catalytically active residues

Angles in catalytically active residues were measured in D2. In the nucleophiles γO the average angle is of 28° with an IQR of 5,2°, which is 18,6% of the angle average. The angles in the nucleophile have a 3σ of +/- 9,19° from the 28 ° mean. See red graph in Figure 16.

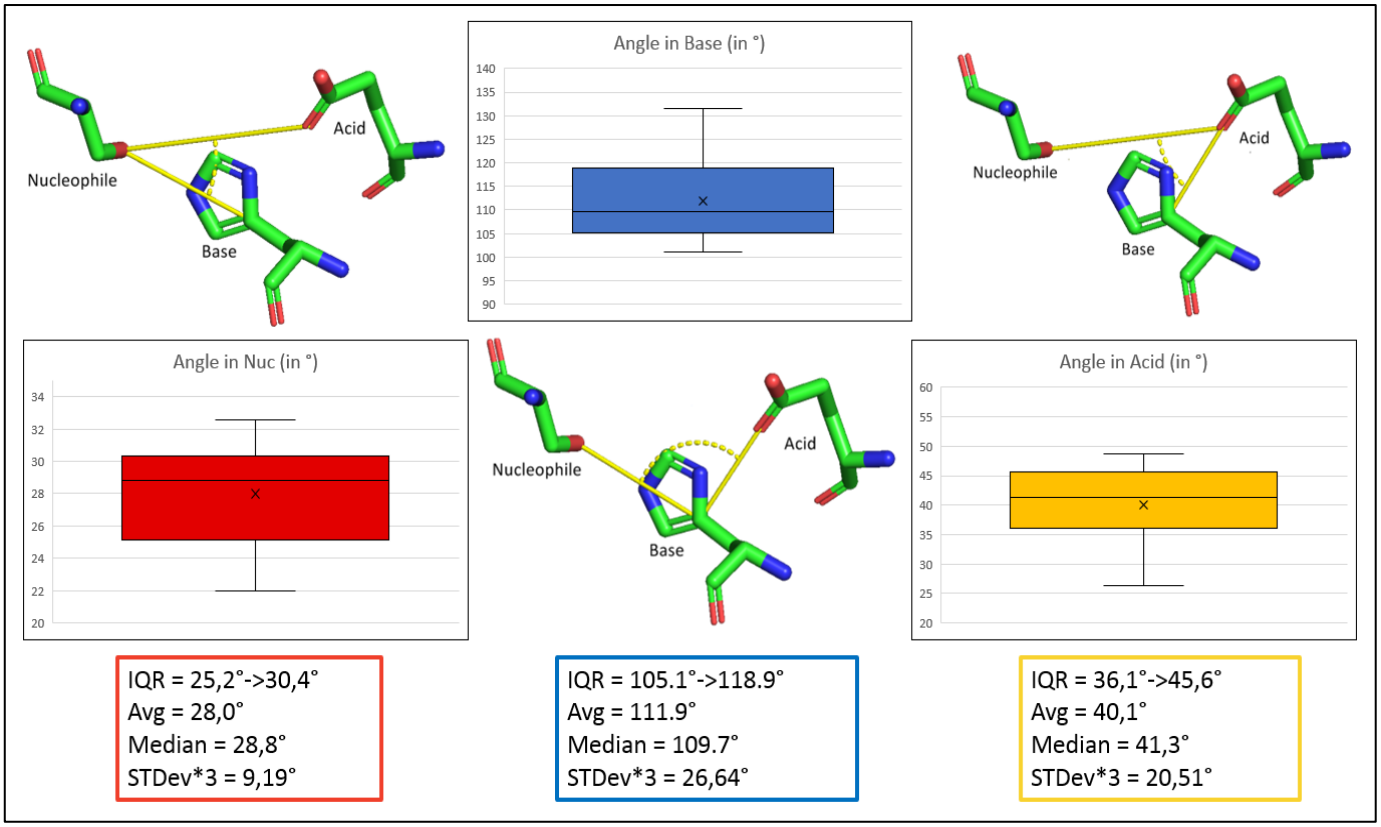


**Figure 16- Angles of catalytic triad members – figure shows which atoms were the angles measured from, the statistics of these measurements and a Whisker Box plot graph of the data. Red (left) is representative for the nucleophile; blue (middle) is the base; and yellow (right) is the acid. All y-axes are in degrees. The stick representations show from which atom the measurements were performed.**

Angles in base are within a 13,8° IQR with an average angle of 111,9°, which means the IQR is 12,3% of the angle average. The angles in the base have a 3σ of +/-26,64° from the 111,9° mean. See blue graph in Figure 16.

Angles in the acid are on average 40,1° with an IQR of 9,5° which is a 23,7% of the angle average. The angles in the acid have a 3σ of +/-20,51° from the 40,1° mean. The acid members are the least geometrically conserved of the catalytic triad members. See yellow graph in Figure 16- Angles of catalytic triad members – figure shows which atoms were the angles measured from, the statistics of these measurements and a Whisker Box plot graph of the data. Red (left) is representative for the nucleophile; blue (middle) is the base; and yellow (right) is the acid. All y-axes are in degrees. The stick representations show from which atom the measurements were performed.

Angles in the nucleophile and base in relation to the oxyanion holes was also measured. The nucleophile  angles are more dispersed than the angles in the base when measuring them in between oxyanion holes. The average angle for the nucleophiles is 65,4°, the IQR is from 71,9° to 52,7° and the 3σ is 58°. The average angle for the base, however, is 31,9°, the IQR is from 29,3°to 34,1° and the 3σ is 18,5°.

## 3.5 Synthetic peptide comparison

The representative triad containing EC 3.1 protein was chosen to be cutinase, a 230 amino acid enzyme from the EC 3.1.1.74 with a Ser-His-Asp triad distanced 7,1 Å, 4,9 Å and 3,7 Å for it nucleophile-acid, nucleophile-base and acid-base distances, respectively. The average being 7,2 Å, 4,9 Å and 3,7 Å.

Using the secondary measurement references the distances from aspartate to histidine for cutinase are 1,8 Å and 2,5 Å, while the distance from serine to histidine is 1,7 Å (Figure 17, b).

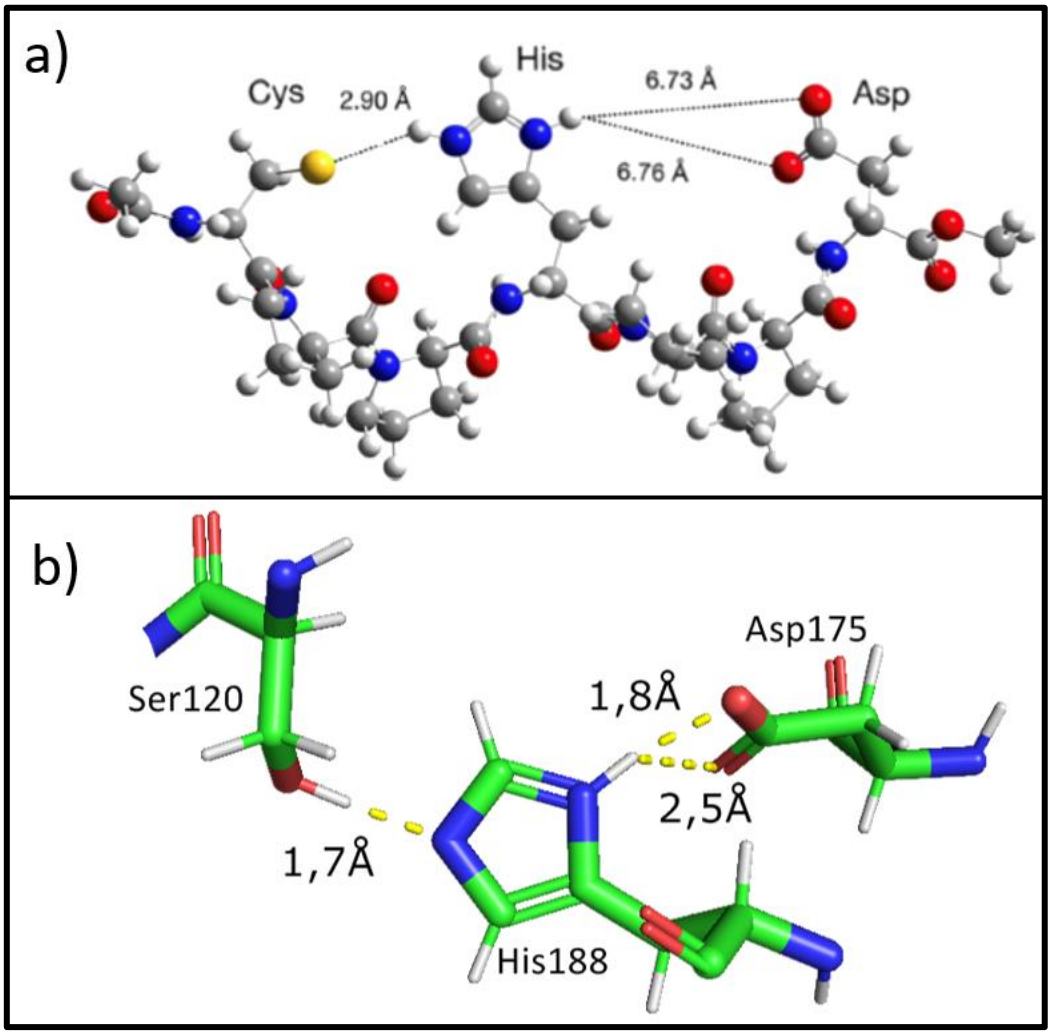


**Figure 17 - Synthetic peptide (a) to natural protein (b) comparison – The top picture is a rotated image copied from Hung et. al. 2017. The bottom is a PyMol generated picture of cutinase triad members and their respective distances.**

The C3H6D9 from the literature has these same distances 6.73 and 6.76 Å for aspartate-histidine distance and 2.9 Å for the cysteine-histidine distance, for reference (Figure 17, a).

# 4 Discussion

This study set out to identify common distances and angles between catalytically active residues and to characterize the properties of amino acids near these catalytic residues.

The catalytic triad containing EC 3.1 enzymes have shown strict adherence to a specific geometric pattern in between the three catalytic residues and the two oxyanion holes used as stabilizers. The average angles are given as a 28°/111,9°/40,1° triangle with said angles in the nucleophile, base and acid residues, respectively. These residues are distanced between each other 3.7 Å, 4.9 Å and 7.1 Å from acid to base, base to nucleophile and nucleophile to acid, respectively. The angles within the triad fluctuate within 20% of the total average angle value in their interquartile range and only 10% within the average distance in the interquartile range for distance measurements. These parameters are consistent with previous research on triad distances measuring similar distances.[55]

Oxyanion hole geometries were less conserved than the triad residue parameters. The oxyanion member distances are on average 4,2 Å/ 8,3 Å/ 10,6 Å from the nucleophile, base and acid respectively. Angles are an average of 65,4° and 31,9° for the nucleophile and base respectively. Interestingly, the nucleophile geometry varied twice as the geometry in base residues, despite their crucial function being stabilizing the nucleophilic attack which should in theory happen at a specific 90°angle.

In addition, it was noticed that a specific composition is prevalent in the close proximity of the catalytically active residues of the triad. All catalytic roles favor a non-polar environment with lower-than-average number of basic and hydroxylic compounds. This is probably because such amino acids could impede catalytic function by redistributing local charges, changing the triad geometry or causing steric clashes with the substrate and active residues.

The specific changes in composition (mol%) are seen even from the analysis of LAS and SAS in some parameters. For example, the nonpolar category has the highest mol% difference when comparing all EC 3.1 enzymes and only triad containing ones, with an increase of nonpolar mol% in the triad containing enzymes. This can be rationalized as the full EC 3.1 has mechanisms involving metal ions instead of residue reactants in over half of the dataset. Because cation mechanisms would not have the same constraints as a triad mechanism or would even favor polar molecules for substrate stabilization.

Additionally, a trend was found in the sequence percentage LAS analysis regarding nonpolar mol%. There was an increase in the nonpolar mol% when the LAS sequence percentage was lowered. When the nonpolar mol% of the whole protein was subtracted from the LAS of the same sequence percentage category the following occurred: LAS (35-50% seq) had an increase of +0.40%, while the LAS (75-85% seq) had a decrease of 1.87%.

Nucleophiles were found to additionally prefer amino acids with a 'tiny' character, but this was expected with their characteristic GXSXG motif. Nucleophiles, in fact, tend to have a significant percentage less of any sort of non-neutral amino acid near them compared to the base and acid role residues. Acid role residues had an increase in sulfur containing and aromatic residue prevalence which is unusual given the DIALL sequence motifs often found in human serine proteases.[56]

The catalytic peptides found in the literature that were selected for comparison also had a high degree of nonpolar and small mol% for charged, acidic and basic properties for residues given that 85% of the sequence were prolines. In the case of Hyp-C3H6D9 these prolines were replaced by hydroxyprolines which would possibly give the entire molecule a more hydroxylic and polar mol%. However, the prevalence of such characteristics may be completely irrelevant because of the PPII helix formed by these structures that would negate the steric or charge disrupting effects that

amino acids of such character might cause in a naturally occurring enzyme.[40]

Furthermore, the literature provided minimized structures for a catalytic peptide with a Cys-His-Asp triad. Their respective distances are 6,76/6,73 Å for a base-acid distance, and 2,90 Å for a nucleophile-base distance as measured in the study, with a linear planar placement between these residues.[40] Measuring from the same reference points, a cutinase (which is the closest in terms of all average geometric parameters found in this study) has an equivalent nucleophile-base distance of 1,7 Å and a 1,6/2,5 Å base-acid distance. These distances could be different if examined in a crystal or using a solution simulation using molecular dynamics software, however the distances could explain the relatively lower activity compared to natural enzymes.

Defining an active site as 'long' and 'short' might work when analyzing enzymes with catalytic loops in which all of the catalytically active residues are contained within a single amino acid string within the loop, or on the same secondary structure. However, in the case of Rossman fold containing EC 3.1 catalytic triads this is not possible because of the interspaced nature of active residues spanning nearly to the entire protein sequence length. The average distance of the last and first active residue spanning 63% of the entire protein. In addition to this, the catalytic process would most likely be compromised if the catalytically active residues were on the very end of the enzyme. A spacer or secondary structure would be needed to stabilize the structure, which would then contain even larger percentages of the native protein. This makes measuring active site composition properties difficult or inaccurate. It also creates a problem when trying to create a shorter peptide with the same function, making cutting out the active site portion redundant.

Finally, this work had limitations. The measurements were done on a small sub-subclass of enzymes with a common evolutionary background. The

active sites often contain many other residues that orient the triad members which could be important for catalytic efficiency and the active residues we did analyze are unique in their rigidity through the substrate binding process. Other sites that are less rigid, from other families or have different mechanisms may provide additional problems or even completely different approaches than done here. The geometry analysis could be improved by adding more measurement points and by simulating the PDB enzyme crystals in solution using molecular dynamics software. Another improvement could be achieved by comparing different PDB structures of the same proteins with their activity, trying to correlate the activity with geometric change. On the issue of composition analysis there are further plans to analyze the conservation of amino acids near active residues, to look at amino acid prevalence based on how near it is from the active site and to look at the neighboring residues through a more advanced spatial model which also takes into consideration 3D physical distance from active residue, not only sequence distance.

In the future, this can be expanded by adding molecular dynamics to the analysis and by testing these concepts in enzyme assays. A catalytic peptidomimetic candidate can by created *de novo* with varying degrees of triad geometry flexibility and with different residue compositions around the active residues. Molecular dynamics (MD) were performed while writing this thesis and are an avenue to expand this research.

# 5 Conclusions

It was found that catalytic triads within the EC 3.1 subclass have strict geometric parameters in between the three residues and with the oxyanion hole stabilizers. The triad members consistently had a configuration of a 28°/111,9°/40,1° triangle with said angles in the nucleophile, base and acid residues, respectively. The triangle has sides with distances of 3.7 Å

/4.9 Å /7.1 Å from acid to base, base to nucleophile and nucleophile to acid, respectively.

These parameters are conserved among triad containing enzymes measuring interquartile ranges within 10% of the average for distances and around 20% for angles.

It was also discovered that a certain composition is prevalent near the catalytically active residues of the triad. Namely, catalytic residues favor a non-polar environment with low number of basic, acidic and charged residues.

The synthetic peptides in the literature do not closely resemble these geometries, nor do they purposefully attempt to copy natural amino acid surroundings around the catalytic residues. This could be the source of their inefficiency for catalysis, in tandem with the lack of an oxyanion hole.

In summary, this thesis shows us how future catalytic peptides used for ester hydrolysis should be designed by having in mind the polarity and charge of residues near the catalytic triad and attempt to incorporate a scaffold that can accommodate for a certain geometry among the triad members as well as the presence of an oxyanion hole.

# 6 Literature


1 Berg JM, Tymoczko JL, Stryer L. Proteases: Facilitating a Difficult Reaction. *Biochem 5th Ed* 2002.https://www.ncbi.nlm.nih.gov/books/NBK22526/ (accessed 17 May2021).

2 Cooper GM. The Central Role of Enzymes as Biological Catalysts. *Cell Mol Approach 2nd Ed* 2000.https://www.ncbi.nlm.nih.gov/books/NBK9921/ (accessed 28 Jun2021).

3 EC 3. Hydrolases. https://www.qmul.ac.uk/sbcs/iubmb/enzyme/EC3/ (accessed 24 Aug2021).

4 Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res* 2018; **46**: D618–D623.

5 Pravda L, Berka K, Svobodová Vařeková R, Sehnal D, Banáš P, Laskowski RA *et al.* Anatomy of enzyme channels. *BMC Bioinformatics* 2014; **15**: 379.

6 Structure and Inhibition of Tuberculosinol Synthase and Decaprenyl Diphosphate Synthase from Mycobacterium tuberculosis | Journal of the American Chemical Society. https://pubs.acs.org/doi/10.1021/ja413127v (accessed 24 Aug2021).

7 Structure and evolution of the serum paraoxonase family of detoxifying and anti-atherosclerotic enzymes | Nature Structural & Molecular Biology. https://www.nature.com/articles/nsmb767 (accessed 24 Aug2021).

8 Catalytic Metal Ion Rearrangements Underline Promiscuity and Evolvability of a Metalloenzyme - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0022283613000120?via%3Dihub (accessed 24 Aug2021).

9 Patel S, Marťínez-Ripoll M, Blundell TL, Albert A. Structural Enzymology of Li+-sensitive/Mg2+-dependent Phosphatases. *J Mol Biol* 2002; **320**: 1087–1094.

10 Spyridaki A, Matzen C, Lanio T, Jeltsch A, Simoncsits A, Athanasiadis A *et al.* Structural and biochemical characterization of a new Mg(2+) binding site near Tyr94 in the restriction endonuclease PvuII. *J Mol Biol* 2003; **331**: 395–406.

11 York JD, Ponder JW, Chen Z, Mathews FS, Majerus PW. Crystal Structure of Inositol Polyphosphate 1-Phosphatase at 2.3-.ANG. Resolution. *Biochemistry* 1994; **33**: 13164–13171.

12 A novel endonuclease mechanism directly visualized for I-PpoI | Nature Structural & Molecular Biology. https://www.nature.com/articles/nsb1299_1096 (accessed 24 Aug2021).

13 Lyu Y, Ye L, Xu J, Yang X, Chen W, Yu H. Recent research progress with phospholipase C from Bacillus cereus. *Biotechnol Lett* 2016; **38**: 23–31.

14 Dessen A. Structure and mechanism of human cytosolic phospholipase A2. *Biochim Biophys Acta BBA - Mol Cell Biol Lipids* 2000; **1488**: 40–47.

15 Mutational analysis of the Streptomyces scabies esterase signal peptide | SpringerLink. https://link.springer.com/article/10.1007%2Fs002530050669 (accessed 24 Aug2021).

16 The Reaction Mechanism of Phospholipase D from Streptomyces sp. Strain PMF. Snapshots along the Reaction Pathway Reveal a Pentacoordinate Reaction Intermediate and an Unexpected Final Product - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0022283604004073?via%3Dihub (accessed 24 Aug2021).

17 Yuen MH, Mizuguchi H, Lee Y-H, Cook PF, Uyeda K, Hasemann CA. Crystal Structure of the H256A Mutant of Rat Testis Fructose-6-phosphate,2-kinase/Fructose-2,6-bisphosphatase: FRUCTOSE 6-PHOSPHATE IN THE ACTIVE SITE LEADS TO MECHANISMS FOR BOTH MUTANT AND WILD TYPE BISPHOSPHATASE ACTIVITIES *. *J Biol Chem* 1999; **274**: 2176–2184.

18 Crystal structures of rat acid phosphatase complexed with the transition-state analogs vanadate and molybdate - LINDQVIST - 1994 - European Journal of Biochemistry - Wiley Online Library. https://febs.onlinelibrary.wiley.com/doi/10.1111/j.1432-1033.1994.tb18722.x (accessed 24 Aug2021).

19 PTEN catalysis of phospholipid dephosphorylation reaction follows a two-step mechanism in which the conserved aspartate-92 does not

function as the general acid — Mechanistic analysis of a familial Cowden disease-associated PTEN mutation - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0898656807000356?via%3Dihub (accessed 24 Aug2021).

20 Does Positive Charge at the Active Sites of Phosphatases Cause a Change in Mechanism? The Effect of the Conserved Arginine on the Transition State for Phosphoryl Transfer in the Protein-Tyrosine Phosphatase from Yersinia | Journal of the American Chemical Society. https://pubs.acs.org/doi/10.1021/ja992361o (accessed 24 Aug2021).

21 Wong BJ, Gerlt JA. Divergent Function in the Crotonase Superfamily: An Anhydride Intermediate in the Reaction Catalyzed by 3-Hydroxyisobutyryl-CoA Hydrolase. *J Am Chem Soc* 2003; **125**: 12076–12077.

22 Crystal structure of plant pectin methylesterase - Johansson - 2002 - FEBS Letters - Wiley Online Library. https://febs.onlinelibrary.wiley.com/doi/full/10.1016/S0014-5793%2802%2902372-4 (accessed 24 Aug2021).

23 1.3 Å Structure of Arylsulfatase from Pseudomonas aeruginosa Establishes the Catalytic Mechanism of Sulfate Ester Cleavage in the Sulfatase Family: Structure. https://www.cell.com/structure/fulltext/S0969-2126(01)00609-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0969212601006098%3Fshowall%3Dtrue (accessed 24 Aug2021).

24 Crystal Structure of an Enzyme-Substrate Complex Provides Insight into the Interaction between Human Arylsulfatase A and its Substrates During Catalysis - ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0022283600942979?via%3Dihub (accessed 24 Aug2021).

25 Chen L, Kong X, Liang Z, Ye F, Yu K, Dai W *et al.* Theoretical Study of the Mechanism of Proton Transfer in the Esterase Estb from Burkholderia Gladioli. *J Phys Chem B* 2011; **115**: 13019–13025.

26 Lim K, Tempczyk A, Bonander N, Toedt J, Howard A, Eisenstein E *et al.* A Catalytic Mechanism for d-Tyr-tRNATyrDeacylase Based on the Crystal Structure of Hemophilus influenzae HI0670 *. *J Biol Chem* 2003; **278**: 13496–13502.

27 Dodson G, Wlodawer A. Catalytic triads and their relatives. *Trends Biochem Sci* 1998; **23**: 347–352.

28 Buller AR, Townsend CA. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc Natl Acad Sci* 2013; **110**: E653–E661.

29 Ekici ÖD, Paetzel M, Dalbey RE. Unconventional serine proteases: Variations on the catalytic Ser/His/Asp triad configuration. *Protein Sci Publ Protein Soc* 2008; **17**: 2023–2037.

30 Wells JA, Cunningham BC, Graycar TP, Estell DA, Blow DM, Fersht AR *et al.* Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin. *Philos Trans R Soc Lond Ser Math Phys Sci* 1986; **317**: 415–423.

31 Pantoliano MW, Ladner RC, Bryan PN, Rollence ML, Wood JF, Poulos TL. Protein engineering of subtilisin BPN': enhanced stabilization through the introduction of two cysteines to form a disulfide bond. *Biochemistry* 1987; **26**: 2077–2082.

32 Carlos JL, Klenotic PA, Paetzel M, Strynadka NC, Dalbey RE. Mutational evidence of transition state stabilization by serine 88 in Escherichia coli type I signal peptidase. *Biochemistry* 2000; **39**: 7276–7283.

33 Ordentlich A, Barak D, Kronman C, Ariel N, Segall Y, Velan B *et al.* Functional Characteristics of the Oxyanion Hole in Human Acetylcholinesterase. *J Biol Chem* 1998; **273**: 19509–17.

34 Der BS, Edwards DR, Kuhlman B. Catalysis by a De Novo Zinc-Mediated Protein Interface: Implications for Natural Enzyme Evolution and Rational Enzyme Engineering. *Biochemistry* 2012; **51**: 3933–3940.

35 Al-Garawi ZS, McIntosh BA, Neill-Hall D, Hatimy AA, Sweet SM, Bagley MC *et al.* The amyloid architecture provides a scaffold for enzyme-like catalysts. *Nanoscale* 2017; **9**: 10773–10783.

36 Conjugating Catalytic Polyproline Fragments with a Self-Assembling Peptide Produces Efficient Artificial Hydrolases | Biomacromolecules. https://pubs.acs.org/doi/10.1021/acs.biomac.9b01620 (accessed 24 Jun2021).

37 Short peptides self-assemble to produce catalytic amyloids | Nature Chemistry. https://www.nature.com/articles/nchem.1894 (accessed 24 Jun2021).

38 Zastrow ML, Peacock AFA, Stuckey JA, Pecoraro VL. Hydrolytic catalysis and structural stabilization in a designed metalloprotein. *Nat Chem* 2012; **4**: 118–123.

39 Friedmann MP, Torbeev V, Zelenay V, Sobol A, Greenwald J, Riek R. Towards Prebiotic Catalytic Amyloids Using High Throughput Screening. *PLOS ONE* 2015; **10**: e0143948.

40 Design of Polyproline-Based Catalysts for Ester Hydrolysis | ACS Omega. https://pubs.acs.org/doi/10.1021/acsomega.7b00928 (accessed 24 Jun2021).

41 Zhang C, Xue X, Luo Q, Li Y, Yang K, Zhuang X *et al.* Self-Assembled Peptide Nanofibers Designed as Biological Enzymes for Catalyzing Ester Hydrolysis. *ACS Nano* 2014; **8**: 11715–11723.

42 Switchable Hydrolase Based on Reversible Formation of Supramolecular Catalytic Site Using a Self-Assembling Peptide - Zhang - 2017 - Angewandte Chemie International Edition - Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201708036 (accessed 24 Jun2021).

43 Rational design of metalloenzymes: From single to multiple active sites. *Coord Chem Rev* 2017; **336**: 1–27.

44 Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A *et al.* De novo computational design of retro-aldol enzymes. *Science* 2008; **319**: 1387–1391.

45 Weng Y-Z, Chang DT-H, Huang Y-F, Lin C-W. A study on the flexibility of enzyme active sites. *BMC Bioinformatics* 2011; **12**: S32.

46 Janković P, Šantek I, Pina AS, Kalafatovic D. Exploiting Peptide Self-Assembly for the Development of Minimalistic Viral Mimetics. *Front Chem* 2021; **9**: 594.

47 Zhang S. Discovery and design of self-assembling peptides. *Interface Focus* 2017; **7**: 20170028.

48 Lee S, Trinh THT, Yoo M, Shin J, Lee H, Kim J *et al.* Self-Assembling Peptides and Their Application in the Treatment of Diseases. *Int J Mol Sci* 2019; **20**: 5850.

49 Gupta S, Singh I, Sharma AK, Kumar P. Ultrashort Peptide Self-Assembly: Front-Runners to Transport Drug and Gene Cargos. *Front Bioeng Biotechnol* 2020; **8**: 504.

50 Maeda Y, Makhlynets OV, Matsui H, Korendovych IV. Design of Catalytic Peptides and Proteins Through Rational and Combinatorial Approaches. *Annu Rev Biomed Eng* 2016; **18**: 311–328.

51 Smith AJT, Müller R, Toscano MD, Kast P, Hellinga HW, Hilvert D *et al.* Structural Reorganization and Preorganization in Enzyme Active

Sites: Comparisons of Experimental and Theoretically Ideal Active Site Geometries in the Multistep Serine Esterase Reaction Cycle. 2008. doi:10.1021/ja803213p.

52 Peptide studies by means of principal properties of amino acids derived from MIF descriptors - Cruciani - 2004 - Journal of Chemometrics - Wiley Online Library. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/cem.856 (accessed 8 Sep2021).

53 Osorio D, Rondón-Villarreal P, Torres Sáez R. *'Peptides' Calculate indices and theoretical physicochemical properties of peptides and protein sequences*. 2014 doi:10.13140/2.1.1755.8407.

54 Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet TIG* 2000; **16**: 276–277.

55 Iengar P, Ramakrishnan C. Knowledge-based modeling of the serine protease triad into non-proteases. *Protein Eng Des Sel* 1999; **12**: 649–656.

56 Yousef GM, Elliott MB, Kopolovic AD, Serry E, Diamandis EP. Sequence and evolutionary analysis of the human trypsin subfamily of serine peptidases. *Biochim Biophys Acta BBA - Proteins Proteomics* 2004; **1698**: 77–86.

# 7 Životopis

## Marko Babić


- **Državljanstvo:** hrvatsko
- (+385) 981732592
- markobabic333@outlook.com
- Whatsapp Messenger: +385981732592
- Nikole Cara 11, 51000, Rijeka, Hrvatska


## 7.1 *RADNO ISKUSTVO*

- 08/2014 – 12/2014 – Rijeka, Hrvatska

### 7.1.1 TEHNIČARI/TEHNIČARKE ZA TELEKOMUNIKACIJE – **T-COM**

Radio na T-com terminalima za DSLAM i pomagao tehničarima na terenu.

- 08/2016 – 12/2017 – Rijeka, Hrvatska

### 7.1.2 DJELATNIK SLUŽBE ZA KORISNIKE – **T-COM**

Riješavanje prigovora za račune i prodaja.

- 01/2020 – 11/2020 – Rijeka, Hrvatska

### 7.1.3 TEHNIČAR ZA ODRŽAVANJE – **BELVEDER D.O.O.**

Održavanje i sigurnost stadiona Kantrida.

- 28/06/2021 – 03/09/2021 – Rijeka, Hrvatska

### 7.1.4 KURATOR BAZE PODATAKA – **EMBL-EBI, THORNTON INSTITUT**

Unos u M-CSA bazu podataka kuriranjem znanstvenih radova o enzimatskim mehanizmima.

## 7.2 OBRAZOVANJE I OSPOSOBLJAVANJE

- 10/2013 – 07/2017 – Radmile Matejčić 2, Rijeka, Hrvatska

### 7.2.1 SVEUČILIŠNI PRVOSTUPNIK BIOTEHNOLOGIJE I ISTRAŽIVANJA LIJEKOVA – **ODJEL ZA BIOTEHNOLOGIJU, SVEUČILIŠTE U RIJECI**

- http://www.biotech.uniri.hr/hr/

- 10/2017 – TRENUTAČNO – Radmile Matejčića 2, Rijeka, Hrvatska

### 7.2.2 MAGISTAR ISTRAŽIVANJA I RAZVOJA LIJEKOVA – **ODJEL ZA BIOTEHNOLOGIJU, SVEUČILIŠTE U RIJECI**

- http://www.biotech.uniri.hr/hr/

## 7.3 JEZIČNE VJEŠTINE

**Materinski jezik/jezici:** HRVATSKI

**Drugi jezici:**

| Razine: A1 i A2: temeljni korisnik; B1 i B2: samostalni korisnik; C1 i C2: iskusni korisnik | | | | | |
|---|---|---|---|---|---|
| | RAZUMIJEVANJE | | GOVOR | | PISANJE |
| | Slušanje | Čitanje | Govorna produkcija | Govorna interakcija | |
| **ENGLESKI** | C2 | C2 | C2 | C2 | C2 |

## 7.4 DIGITALNE VJEŠTINE

Avogadro, UCSF Chimera, PyMol, Marvin, Vina, VMD, NAMD, Cresset Spark, Microsoft Office program, GROMACS molekularno modeliranje, rad u programu SnapGene, Linux (osnove), Zotero, izvrsna sposobnost pretraivanja interneta i strukturiranih baza podataka (PubMed Ensembl)

## 7.5 VOZAČKA DOZVOLA

**Vozačka dozvola:** B kategorije

## 7.6 KONFERENCIJE I SEMINARI

- 21/08/2017 – 25/08/2017 – Rijeka

### 7.6.1 LJETNA ŠKOLA KEMIJIE

## 7.7 PROJEKTI

- 12/2013 – 10/2017

### 7.7.1 PUTUJUĆI ZNANSTVENICI

Stvoritelj projekta za popularizaciju znanosti i edukaciju osnovnoškolske djece "Putujući znanstvenici". Projekt je postao tradicija na Odjelu za biotehnologiju čije se manifestacije još uvijek organiziraju.

- 20/04/2015 – 24/04/2015

### 7.7.2 13. FESTIVAL ZNANOSTI

Predstavljanje Odjela za biotehnologiju na Riječkom Korzu i održavanje pokusa na Otvorenim danima Odjela.

- 21/08/2017 – 25/08/2017

### 7.7.3 LJETNA ŠKOLA KEMIJE

Održao predavanje "Utjecaj fizike na biologiju živih bića".

## 7.8 POČASTI I NAGRADE

- 14/05/2009

### 7.8.1 POHVALA ZA RAD NA PROJEKTU I SUDJELOVANJE NA DRŽAVNOJ SMOTRI IZ DEMOKRATSKOG GRAĐANSTVA – AGENCIJA ZA ODGOJ I OBRAZOVANJE

- 28/02/2011

### 7.8.2 POHVALA ZA SUDJELOVANJE NA ŽUPANIJSKOM NATJECANJU IZ MATEMATIKE – PGŽ

- 05/05/2015

### 7.8.3 POHVALNICA ZA SUDJELOVANJE U 13. FESTIVALU ZNANOSTI – SVEUČILIŠTE U RIJECI

## 7.9 HOBIJI I INTERESI

HOBIJI I INTERESI

- Crtanje olovkom i ugljenom
- Karate i fitness
- Popularizacija znanosti i znanstvene pismenosti
- U slobodno vrijeme piše kratke SF priče