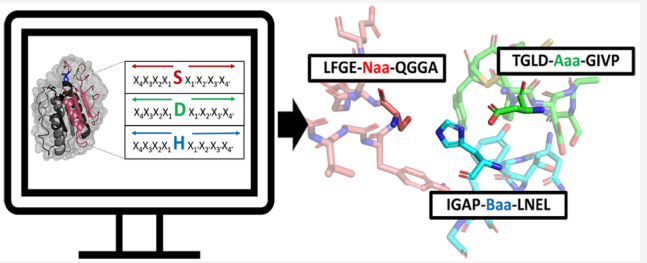
Esterase Sequence Composition Patterns for the Identification of Catalytic Triad Microenvironment Motifs

Marko Babić, Patrizia Janković, Silvia Marchesan, Goran Mauša,* and Daniela Kalafatovic*

ABSTRACT: Ester hydrolysis is of wide biomedical interest, spanning from the green synthesis of pharmaceuticals to biomaterials' development. Existing peptide-based catalysts exhibit low catalytic efficiency compared to natural enzymes, due to the conformational heterogeneity of peptides. Moreover, there is lack of understanding of the correlation between the primary sequence and catalytic function. For this purpose, we statistically analyzed 22 EC 3.1 hydrolases with known catalytic triads, characterized by unique and well-defined mechanisms. The aim was to identify patterns at the sequence level that will better inform the creation of short peptides containing important information for catalysis, based on the catalytic triad, oxyanion holes and the triad residues microenvironments. Moreover, fragmentation schemes of the primary sequence of selected enzymes alongside the study of their amino acid frequencies, composition, and physicochemical properties are proposed. The results showed highly conserved catalytic sites with distinct positional patterns and chemical microenvironments that favor catalysis and revealed variations in catalytic site composition that could be useful for the design of minimalistic catalysts.



INTRODUCTION

Enzymes are proteins that exploit their specific three-dimensional (3D) structure and the amino acid (aa) chemistry of their active site to catalyze chemical reactions by lowering their activation energy.[1,2] The active site is a small portion of the enzyme that consists of the catalytic and binding sites.[3] The former contains residues that are directly involved in the process of catalysis, whereas the latter consists of residues that form temporary bonds with the substrate. Enzymes ensure that all the metabolic processes in cells occur at rates that sustain life. Depending on the reaction they catalyze, enzymes are grouped in seven EC (Enzyme Commission) classes, namely oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, and translocases.[4−6]

Hydrolases (EC 3) are biomolecules that use water to cleave chemical bonds, they are ubiquitous in nature, and carry out degradative reactions in the human body.[7] Often, hydrolases contain a catalytic triad in their active site, that is a set of three amino acids including the nucleophile (for example, Ser/Cys), one basic aa (for example, His) and one acidic aa (for example, Asp/Glu).[8] In the case of ester hydrolysis, the acidic residue modifies the pK_a of the basic one, which deprotonates the nucleophile that performs a *nucleophilic attack* on an acyl carbon, initiating the reaction (Figure S8 of the Supporting Information, SI). During ester hydrolysis, two tetrahedral intermediate states emerge that are stabilized by a set of backbone amides collectively called the oxyanion hole, having binding and catalytic functions.[9−11] The use of chromogenic substrates, such as 4-nitrophenyl acetate, is a convenient strategy to monitor the reaction, enabling high-throughput screening and its application to directed evolution studies.[12]

With the rise of non-natural industrial processes, there is an increasing demand for new and improved enzymes.[13,14] However, they can be difficult and expensive to produce and are unstable in organic solvents or under harsh conditions of temperature and pH.[15] Peptides as short as two-to-three amino acids have emerged as an alternative offering simple and tunable catalysts,[16−18] but with lower catalytic efficiency when compared to enzymes. This drawback is attributed to the conformational heterogeneity of peptides and the lack of well-defined 3D structures characteristic for enzymes. However, peptide self-assembly offers the possibility to obtain nano-structures with a higher degree of order and improved catalytic efficiency over molecular peptides.[19]

With the growing demand of sustainable chemical synthesis, precise theoretical approaches leading to improved peptide designs are needed.[20] The catalytic activity of short peptides is influenced by the residues that make up the sequence and by the order in which they appear.[21,22] The catalytic triad is the most studied part of the enzymatic sequence in the design of

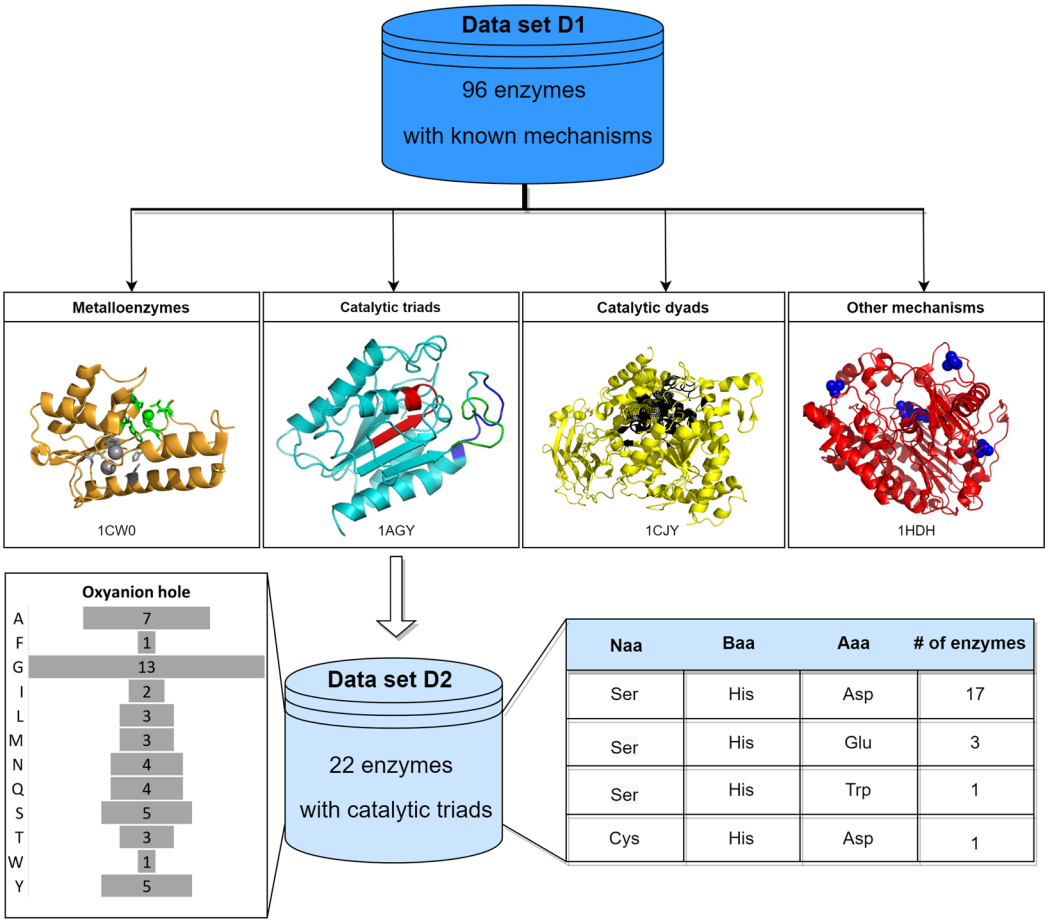


Figure 1. Schematic representation of the data set formation by filtering the enzymes from the M-CSA (EMBL-EBI) database following the EC 3.1 classification. *D1* contains 96 enzymes active on ester bonds, with known mechanisms. Representative examples of each mechanism are shown: VSR endonuclease (PDB ID: 1CW0) for metalloenzymes, cutinase (PDB ID: 1AGY) for triads, cytosolic phospholipases A2 (PDB ID: 1CJY) for dyads, and arylsulfatase (PDB ID: 1HDH) for other mechanisms. *D2* composed of 22 enzymes was obtained (Table S1) by reducing the data set size to EC 3.1 hydrolases with the catalytic triad. (Naa = nucleophilic aa, Baa = basic aa, and Aaa = acidic aa of the catalytic triad).

catalytic peptides.[23−25] Often, the triad residues are singled out and incorporated into scaffolds or sequences able to form nanostructures to maximize the number of active residues in close proximity and promote catalytic activity,[26,27] although not always successfully.[28] Distinct structure packing and catalytic efficiency can be obtained when comparing peptide analogues that differ only in the position of residues forming the catalytic triad (for example, Glu:Ser:His).[24] The catalytic activity of Cys:His:Asp improves with the proximity of thiol (Cys) and imidazole (His) groups within the sequence, especially when promoted by steric hindrance in the presence of bulky, aromatic amino acids (Phe).[25] Other factors, such as peptide termini modification, can affect catalytic performance,[29] and catalyst susceptibility to undergo side-reactions.[30]

Little attention has been drawn to the microenvironment surrounding the triad residues and their position within the enzyme sequence. Moreover, existing designs do not take into consideration oxyanion holes, steric locators, or electrostatic stabilizers naturally occurring in esterases.[23−25] Therefore, better understanding of the chemistry and composition of enzymes and their active sites would expedite the search for peptide sequences that could act as catalysts. It would ultimately reduce the cost and impact on the environment by diminishing the amount of solvents and chemicals required for the synthesis of active peptides.

In this paper, we focused on EC 3.1 ester hydrolases containing catalytic triads, unique at the mechanism, evolution, and reactive center level (Figure 1), to identify representative sequence motifs active on a variety of substrates. The aim was to identify patterns at the primary sequence level, in enzymes that share little or no significant sequence identity, have unique roles for residues performing catalysis and that vary in third and fourth EC numbers to account for substrate variability. The analysis was minimally affected by adaptation to environmental stress, nonester functional groups in substrates, or by different species- and cell-specific requirements, such as turnover rates or specificity required for biological processes. Emphasis was put on residues impacting catalysis (that is, triad residues, oxyanion holes, and microenvironments) with the intention to find universal catalysts, able to promote enzymatic promiscuity.[31,32]

With the objective to abstract the key biochemical information from enzymes regarding their catalytic function, we compared the full protein sequence with the proposed sequence fragments in terms of chemical composition and aa frequencies alongside the positioning of *important* residues at the primary sequence level. In addition, the triad microenvironments of different lengths were assessed to search for positional and composition patterns and to identify dominant motifs that could be used for the informed design of catalytic

peptides. Short sequence fragments based on 22 triad-containing ester hydrolases (Table S2) are provided as well as the microenvironment-based consensus motifs (Table 2). These sequences can be further modeled into self-assembled or cyclic peptides with improved order and/or rigidity or into more complex scaffolds.

METHODS

Data Sets. The data on 96 naturally occurring EC 3.1 enzymes including their catalytic mechanisms, active sites, and oxyanion hole residues were collected via the publicly available Mechanism and Catalytic Site Atlas (M-CSA) database from EMBL-EBI.[33] The enzymes were filtered by selecting the EC 3.1 subclass in the "Browse" interface on the 25th of June 2021 (96 enzymes were listed). The application of a resolution cutoff was not necessary because the X-ray crystallography of all the selected enzymes had a 3.0 Å or greater resolution.

The data set *D1* contains 96 EC 3.1. entries with M-CSA ID, enzyme name, Uniprot AC, EC number, PDB ID, enzyme sequence length, the *full* sequence followed by *long*, *medium*, and *short* fragments, position of the active site in the primary sequence, catalytic triad/dyad residues, oxyanion hole residues, other stabilizers, metal binding sites, bound metal ions, species, and CATH (Class (C), Architecture (A), Topology (T), and Homologous superfamily (H)) numbers. A smaller data set *D2* (Table S1), composed of EC 3.1 enzymes containing catalytic triads and oxyanion holes, was derived from *D1*. The full protein sequences were collected from the Uniprot database. The collected PDB files were processed in PyMol to mark the active residues. In this work we set three main rules:

1. enzymes have unique mechanisms,
2. only enzymes with catalytic triad are used for *D2*,
3. residues taking part in the catalytic triad and/or in the oxyanion hole are defined as *important residues*.

A unique mechanism was defined as having at least one different catalytic residue or using the same catalytic residue in a different way.

The sequence logos were generated for $X_i - Naa - X_{i'}$, $X_i - Aaa - X_{i'}$ and $X_i - Baa - X_{i'}$ (where Naa = nucleophilic aa, Aaa = acidic aa, and Baa = basic aa of the catalytic triad), for three fragment lengths $i/i' = \{4, 8, 16\}$ via the WebLogo service.[34,35] The overall aa frequency was determined using the Protein Calculator.[36] The aa frequency in specific positions in the microenvironment was calculated by counting the aa found on each of their respective positions and then calculating the percentage dividing by the number of sequences (22).

Properties Calculation. The R package *Peptides*[37] was used to calculate the aa composition of peptides, expressed in mole percentage (mol%). The aa are categorized based on their intrinsic properties into 11 subcategories:[38] Tiny (Ala, Cys, Gly, Ser, Thr), Small (Ala, Cys, Asp, Gly, Asn, Pro, Ser, Thr, Val), Aliphatic (Ala, Ile, Leu, Val), Aromatic (Phe, His, Trp, Tyr), Nonpolar (Ala, Cys, Phe, Gly, Ile, Leu, Met, Pro, Val, Trp, Tyr), Polar (Asp, Glu, His, Lys, Asn, Gln, Arg, Ser, Thr), Charged (Asp, Glu, His, Lys, Arg), Basic (His, Lys, Arg), Acidic (Asp, Glu), Sulfur (Cys, Met), and Hydroxylic (Ser, Thr).

All the generated fragment types (*long*, *medium*, *short*) were created by marking important residues according to M-CSA and analyzed alongside the *full* protein sequences in terms of aa composition. In addition, the composition properties were calculated for all the microenvironments ($X_i - Naa - X_{i'}$, $X_i - Aaa - X_{i'}$ and $X_i - Baa - X_{i'}$, for $i/i' = \{4, 8, 16\}$).

Furthermore, the physicochemical properties, including Cruciani properties,[39] instability index,[40] hydrophobicity on Eisenberg scale,[41] hydrophobic moment[41] with a rotational angle of 100° and a sequence fraction length of 11, Boman index,[42] net charge at $pH = 7.4$ on Lehninger scale,[43] and isoelectric point using Lehninger scale, were computed for *D1* and *D2*.

Consensus Microenvironment Determination. To determine the consensus sequences for $X_4 - Naa - X_{4'}$, $X_4 - Aaa - X_{4'}$ and $X_i - Baa - X_{4'}$ microenvironments, 22 sequences from *D2* were overlapped by centering each microenvironment sequence around the corresponding triad residue. The aa frequency was counted for each position of the microenvironment ($i = 4$) and consequently a "consensus" microenvironment was built. When two residues exhibited the highest (and equal) frequency for a single position, we selected the one which is scarcer in the entire microenvironment. For example, if both Ala and Glu appeared 3 times in a single position, and Glu was scarcer than Ala in the whole microenvironment, then Glu would be annotated in the consensus sequence. This was based on the assumption that the appearance of a less frequent aa is a more significant pattern than the appearance of an aa that commonly occurs in the microenvironment. However, if three or more residues shared the highest frequency, then "Xaa" annotation was used to represent that no consensus aa was found for that position.

Homology Analysis. To confirm that data set *D2* was representative, a homology analysis was performed. Each of the 22 sequences were accessed through Uniprot and ran through Uniprots BLASTp to find their respective homologues. The BLASTp program was configured to find only reviewed sequences (UniprotKB Swiss-Prot) with an E value of 0.01 or higher and was limited to a maximum of 100 results. If the enzyme had over 100 homologues the E threshold was increased to 0.0001 and resubmitted. The Clustal-Omega (v 1.2.4, from Uniprot) multiple sequence alignment tool was used to align the homologues using BLASTp and configured to 5 iterations of the alignment. The homologues were then used to create consensus sequences for $X_4 - Xaa - X_{4'}$ microenvironments for each of the *D2* ($n = 22$) enzymes separately. Homology-based consensus sequences were obtained by performing the alignments, cutting out of sections of the alignments containing the microenvironments and then counting the most frequent aa in each respective position. For simplicity, a column with more than 90% gaps was removed, and the remaining sequences were merged. If the positions had three or more aa sharing the highest frequency, then they were marked as "Xaa", to indicate that no consensus was found. If the sequence had two most frequent residues, then both were included in the consensus sequence. The resulting consensus sequences were then analyzed by overlaying them and calculating the total positional frequencies.

RESULTS AND DISCUSSION

We set out to determine whether patterns in composition, physicochemical properties, or important aa positions exist in enzymes of the EC 3.1. subclass, catalyzing ester hydrolysis. Patterns in the primary sequence were searched by statistically analyzing two data sets (*D1* and *D2*) based on the selected entries from the M-CSA database. *D1* consists of 96 identified enzymes having unique mechanisms for ester hydrolysis,

Table 1. Positional Patterns of *Important* Residues at the Sequence Level for *D2*.

Positional pattern						CATH	Oxyanion holes	Triad	Number of enzymes
1	2	3	4	5	6				
Naa/Oxy	Oxy	Oxy	Aaa	Baa		3.40.50.1110	3	Naa, Aaa, Baa	4
Oxy	Naa	Oxy	Aaa	Baa		3.40.50.1820	2	Naa, Aaa, Baa	13
Oxy	Oxy	Naa	Oxy	Aaa	Baa		3	Naa, Aaa, Baa	4
Naa	Oxy	Baa	Oxy	Aaa		3.40.50.180	2	Naa, Baa, Aaa	1

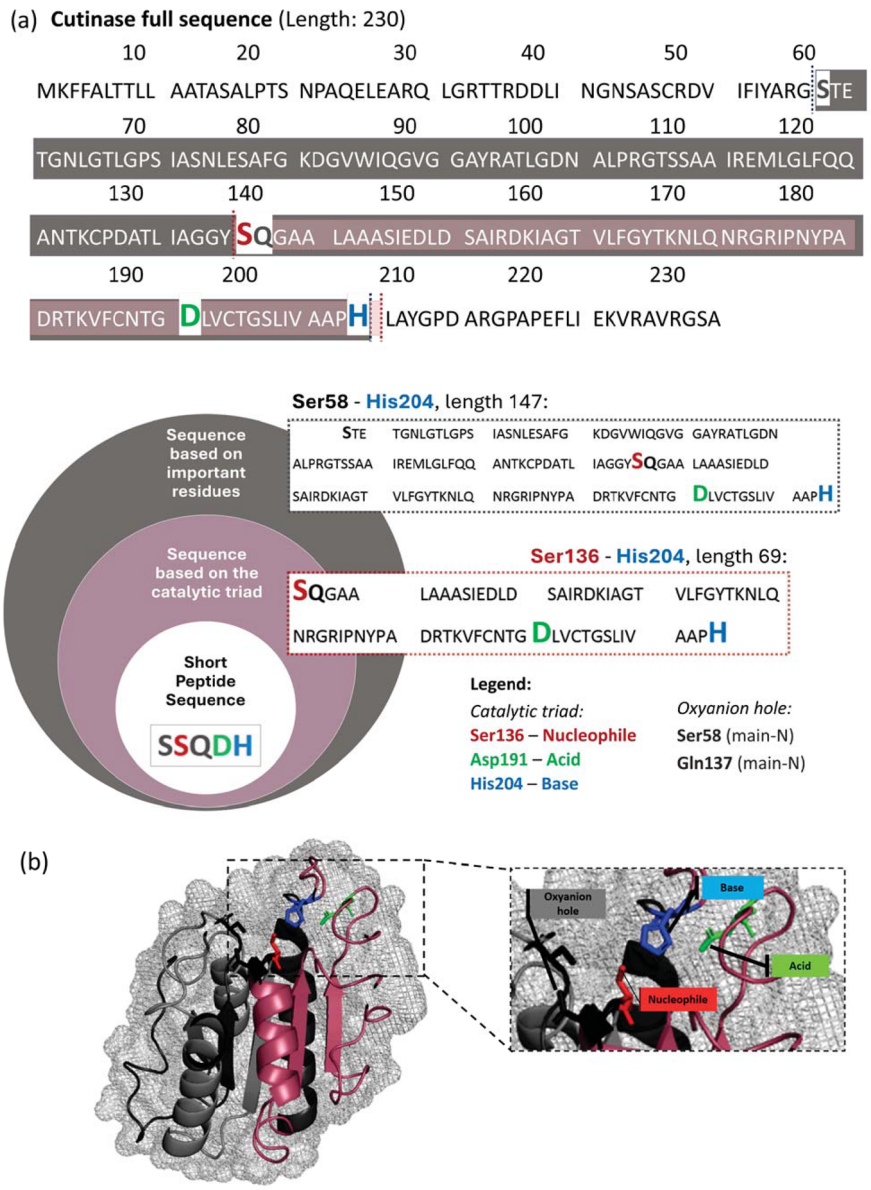


Figure 2. (a) Schematic representation of the *full* (Met1-Ala230) sequence (taken from Uniprot (P00590)), *long* (Ser58-His204), *medium* (Ser136-His204), and *short* (Ser58, Ser136, Gln137, Asp191, His204) fragments of cutinase (1AGY). (b) 3D structure of 1AGY with highlighted triad members Ser (Naa), Asp (Aaa), and His (Baa), obtained in PyMol.

grouped into metalloenzymes, catalytic triads, dyads, and others (Figure 1). The subset of 22 esterases with known catalytic triads constituted the *D2* data set (Table S1). The most common catalytic triad, shared by 17 enzymes, was composed of serine, histidine, and aspartic acid (Ser:His:Asp) standing for *nucleophile* (Naa), *basic aa* (Baa) and *acidic aa* (Aaa), respectively (Figure 1). Other possible triads were Ser:His:Glu, Ser:His:Trp, and Cys:His:Asp indicating that His is always present, the nucleophile additionally allows for Cys, whereas the Aaa can accommodate also Glu and Trp (1ESC). Even though Trp is not acidic in chemical property, it modifies the pK_a of the catalytic base His through its main chain oxygen. Moreover, 12 out of 20 aa were found in the oxyanion holes, with Gly and Ala being the most frequent ones (Figure 1). The physicochemical properties including net charge, isoelectric point, instability index, hydrophobicity, hydrophobic moment, Boman index, and Cruciani properties were computed for *D1* and *D2* (Figure S1). Similar results were obtained for all the

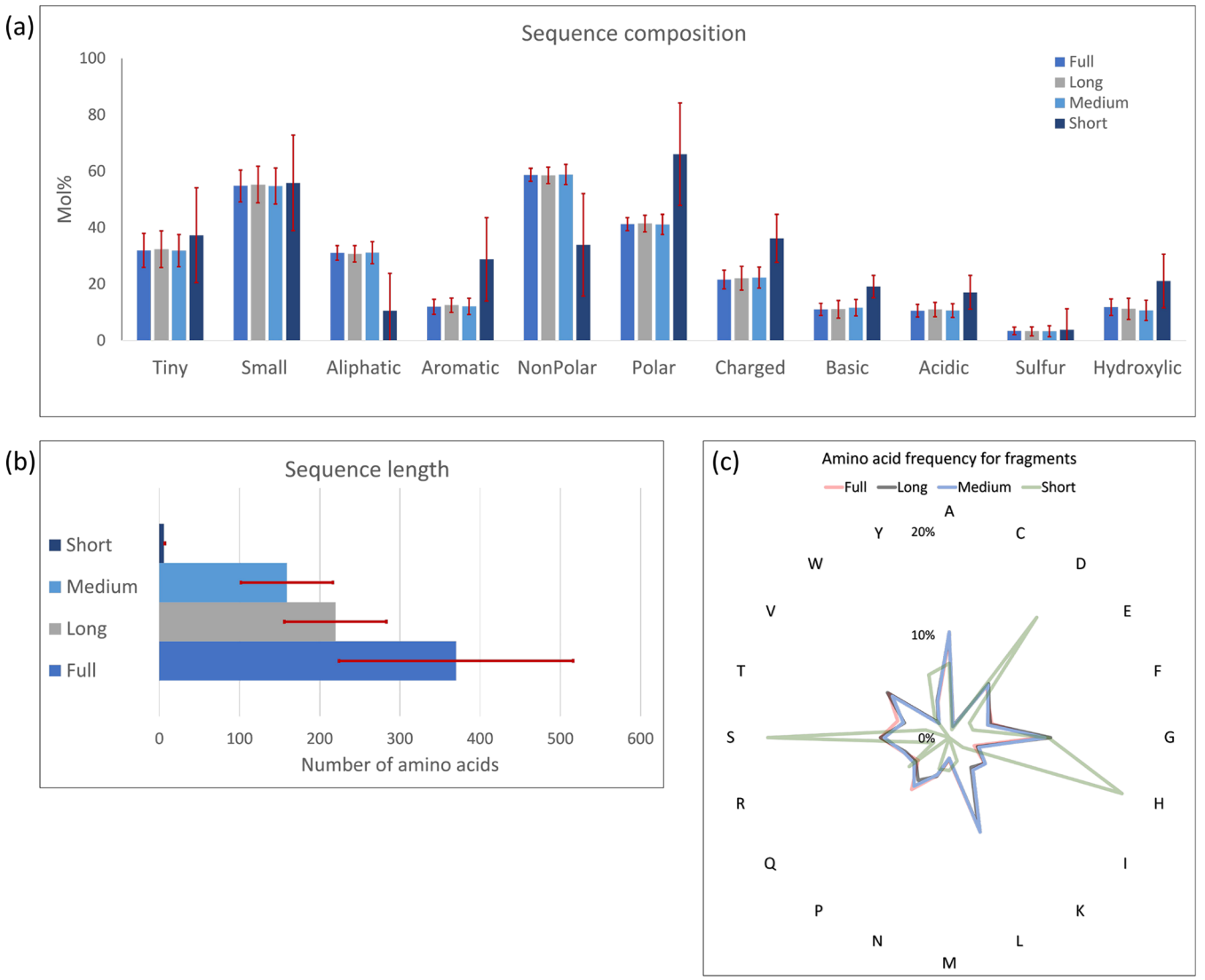


Figure 3. (a) Distribution of fragments composition; (b) the comparison of sequence lengths for each fragment type where the histogram is representing the average value and the red line the standard deviation; and (c) a radar chart of their aa frequencies, where the aa are presented using the one letter code.

properties assessed, except hydrophobicity that was slightly higher in *D2*.

Rather than focusing on homology analysis, that could provide insight into aa conservation trends for each *D2* entry, we focused on searching for patterns among enzymes that do not share significant identity or close evolutionary pathways. This can be compared to divergently evolved enzymes showing conserved 3D structural or active site geometries despite being distant in phylogeny. The motivation was to search for universal sequence motifs in esterases and to identify sequence-level patterns that were overlooked and underutilized to aid in the discovery of minimalistic catalysts.

Positional Patterns. The positioning of active residues in enzymes varies depending on their evolutionary background represented through CATH domains. The colon symbol (:) is used to indicate aa that are not adjacent in the sequence as opposed to the dash (-) that indicates aa connected through peptide bonds. In triad-containing hydrolases, the three-dimensional order in a fully folded protein corresponds to the Naa:Baa:Aaa spatial distribution.[44−46] In data set *D2*, this was observed for all 22 enzymes, with three different CATH numbers (Table 1), namely the α/β-hydrolase fold (3.40.50.1820), SGNH domain (3.40.50.1110), and the methylesterase (3.40.50.180) superfamilies. However, at the sequence level, their positional distribution from the N- to the C-terminus corresponds to Naa:Aaa:Baa. Only the methylesterase superfamily has a Naa:Baa:Aaa distribution maintained also at the sequence level. When oxyanion holes (Oxy) are added to this analysis, four groups can be discerned, namely Naa/Oxy:Oxy:Oxy:Aaa:Baa, Oxy:Naa:Oxy:Aaa:Baa, Oxy:Oxy:Naa:Oxy:Aaa:Baa, and Naa:Oxy:Baa:Oxy:Aaa (Table 1). In SGNH domains, characterized by three-point oxyanion holes, the roles of nucleophile and of the first oxyanion hole are played by the same aa, indicating the double role of this residue. α/β hydrolase folds contain examples with two- and three-point oxyanion holes, having one oxyanion hole adjacent to the nucleophile. 1JKM is the only example in *D2* where the oxyanion hole is not adjacent to the nucleophile. In the 3D structure, all the groups have the (Oxy:)Oxy:Oxy:Naa:Baa:Aaa spatial configuration confirming a highly conserved spatial distribution of the *important* residues throughout the superfamilies.

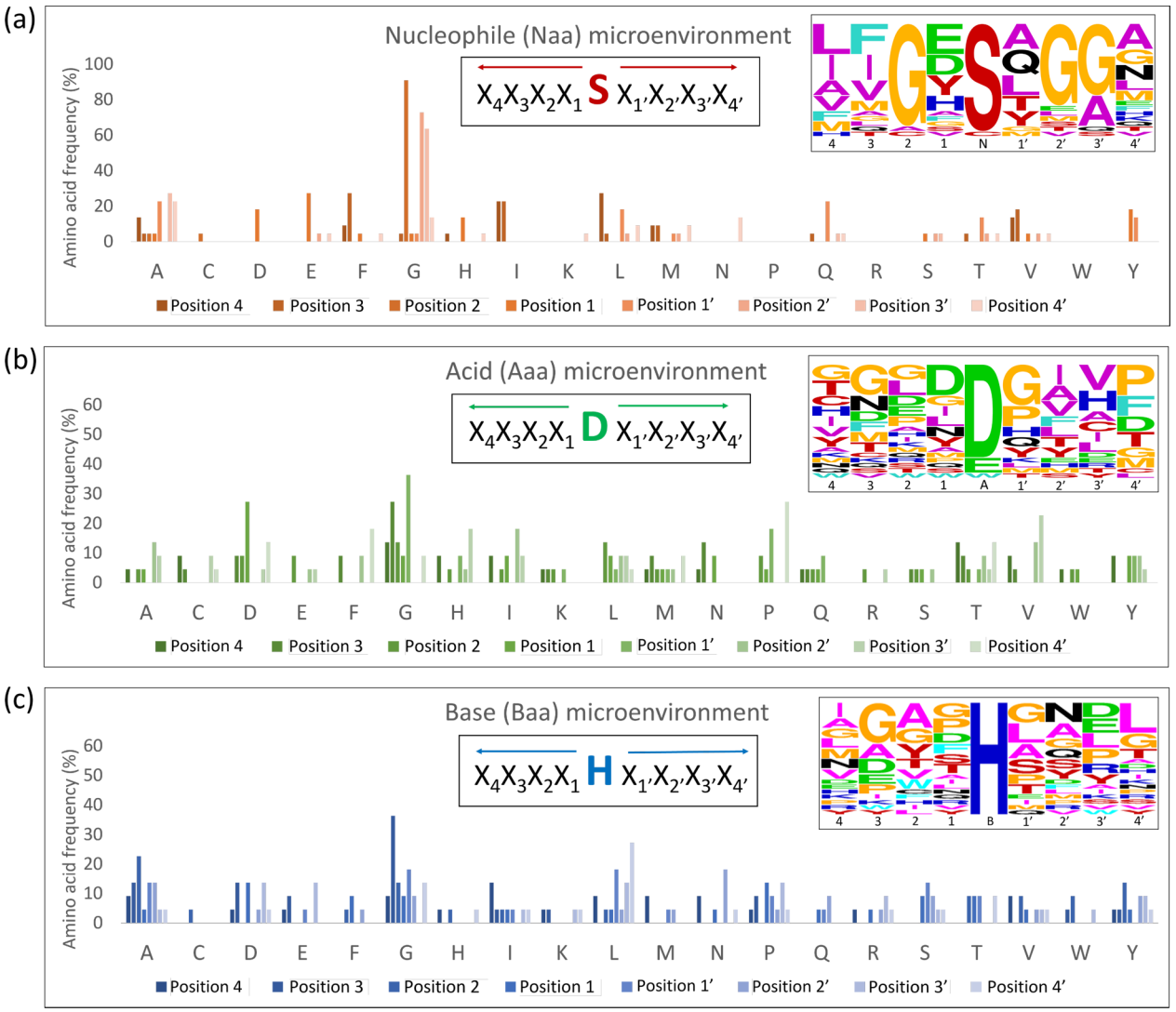


Figure 4. Positional aa frequency for (a) the nucleophilic elbow, (b) the acid loop, and (c) the base loop represented through the most common triad in *D2*: Ser (Naa), Asp (Aaa), and His (Baa) alongside the sequence logos showing the frequency of each aa in positions X_i, where X can be any aa.

Fragmentation Scheme and Composition Patterns. With the intention of investigating how trimming of full protein sequences would affect the overall aa composition, we created a fragmentation scheme consisting of identifying shorter sequences based on *important* residues, that is, the aa taking part in the catalytic triad and the oxyanion hole (Figure 2). On the basis of the primary sequence of selected enzymes with known catalytic triads (*D2*), the following fragmentation scheme is proposed (Figure 2): (1) the complete enzyme sequence (*full*), (2) the sequence based on *important* residues including the segment between the first and the last *important* aa (*long*), (3) the sequence based on the catalytic triad containing the segment between the first and last catalytic aa (*medium*), and (4) the sequence composed only of *important* residues in the order in which they appear in the primary sequence (*short*).

By creating the fragments we observed that the catalytic triad residues of enzymes in data set *D2* were spaced out in the sequence and that the obtained fragments contained large portions of the *full* sequence, resulting in overlapping composition profiles and aa frequencies (Figure 3). In *D2*, the *long* fragment covered on average 62.8% while the *medium* one covered up to 45% of the *full* sequence (Figure 3b). Additionally, the creation of *short* fragments (Table S2) containing only the *important* residues involved in catalysis implied cutting off a considerable part of the sequence. The *short* peptides amount to only 1.54% of the total sequence length of their respective primary *full* sequences, and are enriched in Ser, His, and Asp residues (Figure 3c), being the most common aa in triads.

The composition profiles were created for the selected fragments reflecting on their properties for the following categories: tiny, small, aliphatic, aromatic, nonpolar, polar, charged, basic, acidic, sulfur, and hydroxylic. The average length of the fragments was 370 amino acids for the *full*, 220 amino acids for the *long*, 159 amino acids for the *medium*, and 6 amino acids for the *short* sequences. Additionally, the aa frequencies of each fragment were compared (Figure 3) showing an overlap between *full*, *long*, and *medium* sequences. The calculated composition properties expressed in mole percentage (mol%) and the aa frequencies appear to be similar for all the fragments except for the *short* peptides. As expected,

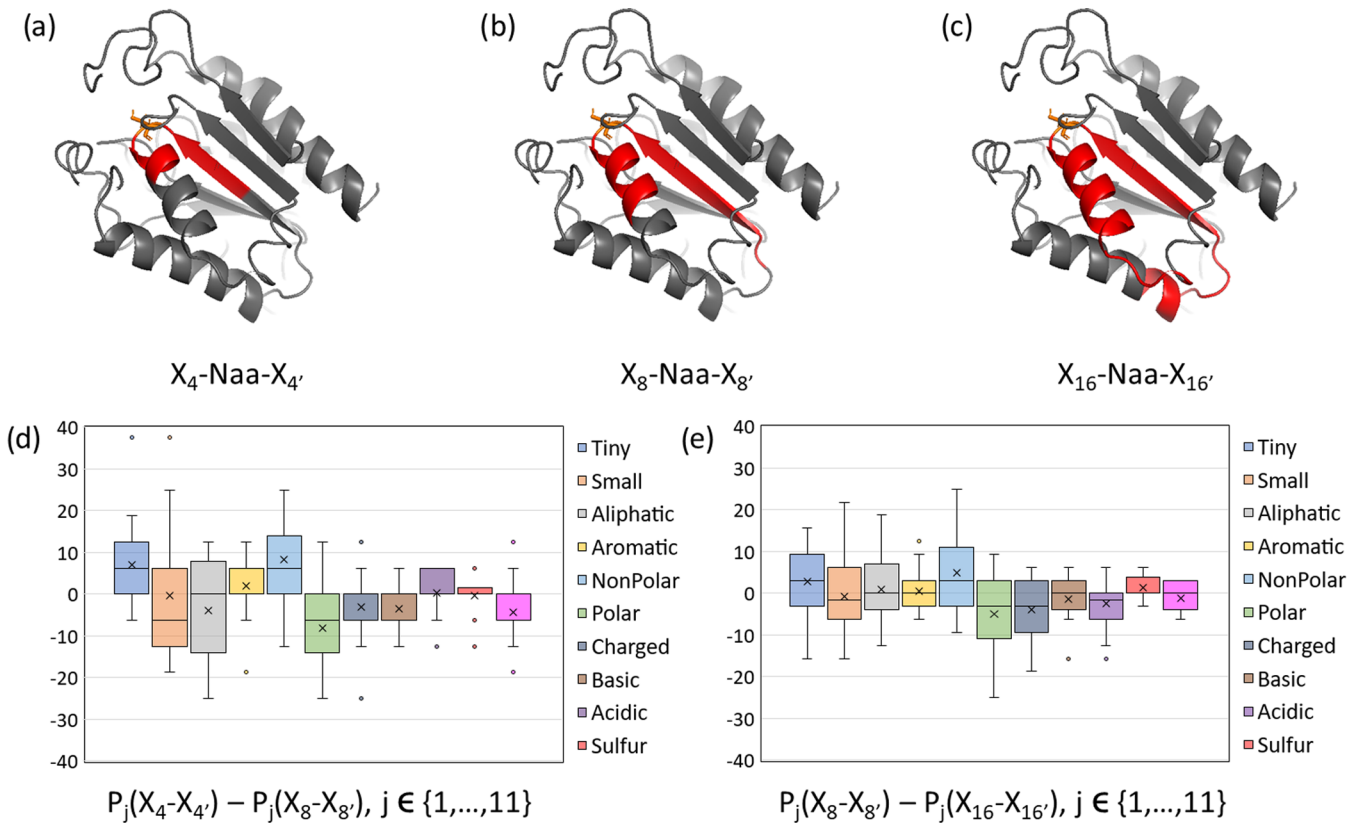


Figure 5. Nucleophile microenvironments: (a) $X_4 - Naa - X_{4'}$, (b) $X_8 - Naa - X_{8'}$, and (c) $X_{16} - Naa - X_{16'}$ highlighted in red. Box plots for each property P_j (mol%) obtained by subtracting $X_4 - X_{4'}$ from $X_8 - X_{8'}$ in (d) and $X_8 - X_{8'}$ from $X_{16} - X_{16'}$ in (e).

the *short* sequences show different composition and aa frequency profiles because they are predominantly made of hydrophilic and charged residues, which are characteristic of the triad.

The difference between the composition properties of *full*, *long*, *medium*, and *short* sequences were estimated by the Friedman statistical test of significance, as the groups of properties did not follow the normal distribution according to the one-way Kolmogorov–Smirnov test. With the statistical level of confidence set to 5%, the results revealed that all composition properties, with the exception of tiny and small, contain at least one pair of significantly different groups. The post hoc tests with Holm-Bonferroni correction were applied to reveal that only the *short* sequences are significantly different from all the other groups, as presented in Figure S2. Although such results are expected, as the *short* sequences contain only a small fraction of the whole protein, it is worth noting that the *long* and *medium* fragmentation schemes conserved the ratio of aa categories.

Physico-chemical properties including hydrophobicity, hydrophobic moment, net charge, and isoelectric point were compared for all the fragments (Figure S7). In accordance with the composition profiles, the mentioned properties were similar for *full*, *long*, and *medium* fragments. As expected, hydrophobic moment and hydrophobicity decreased in *short* sequences, while the net charge and isoelectric point showed a similar but narrow distribution compared to longer fragments. Interestingly, the isoelectric point of all fragments shows that they are neutral at pH values between 5 and 6, and that, at physiological pH, all the fragments are negatively charged, except for the *short* sequences, which are neutral.

Active Site Microenvironments. The lack of research on catalytic triad microenvironments prompted us to investigate residues in close proximity of the catalytic triad members. For this purpose, three types of environments were selected that included portions of the primary sequence centered on each aa of the catalytic triad (that is, Naa, Aaa, or Baa). We analyzed sequences that extended equidistantly on each side of the catalytically active aa by 4, 8, or 16 residues to the left, toward the N-terminus and to the right, toward the C-terminus (Figure 4). Consequently, three sequences having lengths of 8, 16, and 32 residues were created for each microenvironment (Naa, Aaa, or Baa). The catalytic aa, positioned at the center, was not included in the sequence property analysis. The aa positions were denoted X_i, where i indicates the positions from the active aa toward the N-terminus and, similarly, i' toward the C-terminus. Therefore, the microenvironments are denoted as $X_i - Naa - X_{i'}$, $X_i - Aaa - X_{i'}$, and $X_i - Baa - X_{i'}$ for the nucleophile, acidic, and basic residues, respectively.

We hypothesized that if a microenvironment important for ester hydrolysis exists, patterns of residues or chemical properties should emerge by comparing sequences of different lengths (i = 4, 8, 16). The aa frequency analysis and the corresponding sequence logos show a dominant presence of glycine across the considered microenvironments (Figure 4). It is possible that the presence of Gly, being the smallest aa, allows for the required flexibility and space near the nucleophile for the catalysis to occur, as it involves formation of an acyl-enzyme intermediate. In addition, the shortest microenvironment sequences (i = 4) are predominantly made of nonpolar residues, and each shows characteristic patterns. In comparison to the *full* sequence, the $X_4 - Naa - X_{4'}$ is rich in

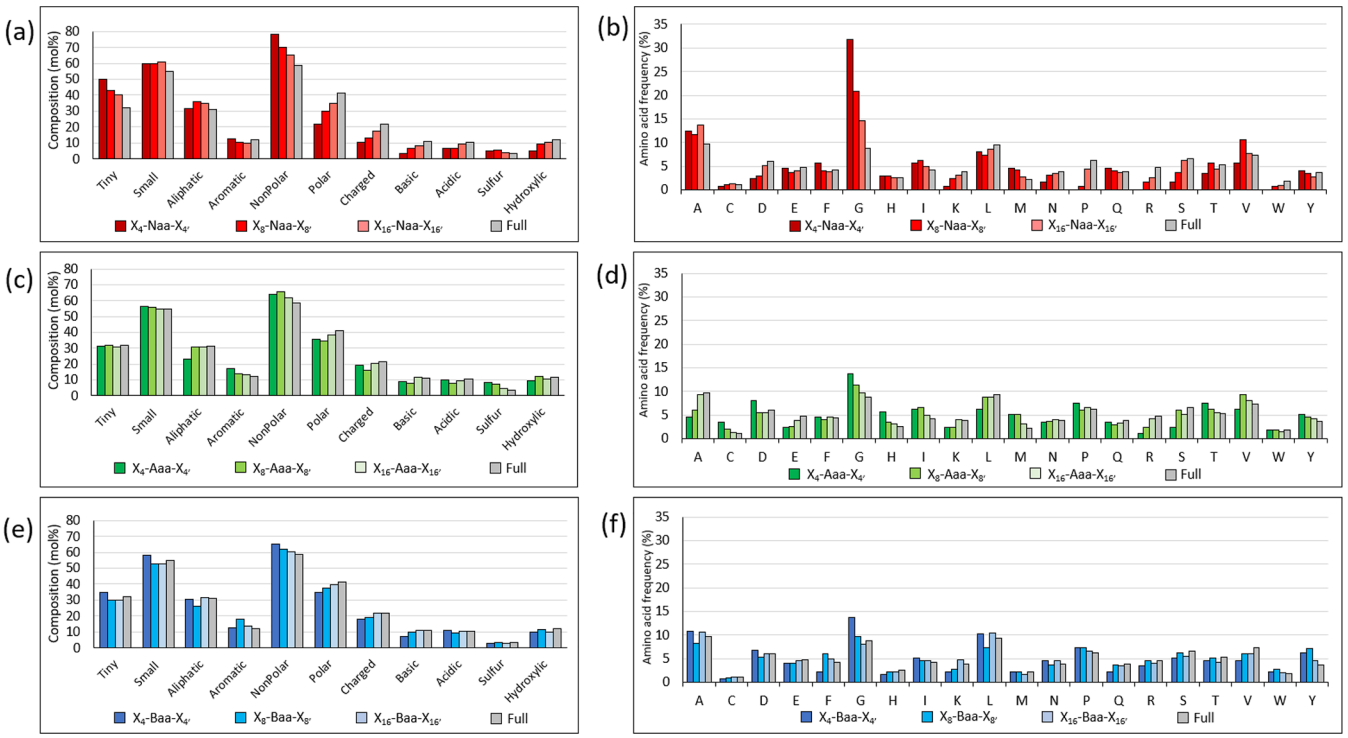


Figure 6. Comparison of composition and aa frequencies for the Naa (a, b), Aaa (c, d), and Baa (e, f) microenvironments, and the *full* sequence showing their similarity and a dominantly nonpolar character.

Ala and Gly; the $X_4 - Aaa - X_{4'}$ in Gly, His, and Cys; and the $X_4 - Baa - X_{4'}$ in Gly and Leu. While the nucleophile (Naa) microenvironment features the least sterically hindering Gly and Ala, it is somewhat surprising that the base (Baa) microenvironment features a bulky Leu. One possible explanation is that, when the folding process leads to His positioning strategically adjacent to the nucleophile that it activates, the presence of Leu provides additional steric hindrance and rigidity to maintain it in place. Since His is not directly involved in the acyl enzyme intermediate, it can be assumed that it requires less space and conformational flexibility nearby, relative to the Naa microenvironment. Moreover, several amino acids, for example, Pro, Arg, and Trp, were absent in $X_4 - Naa - X_{4'}$ or occurred rarely, such as Cys, underrepresented in $X_4 - Baa - X_{4'}$, or Arg in $X_4 - Aaa - X_{4'}$.

When analyzing the positional frequencies, we noticed that specific positions in the proximity of the active residues were favoring certain aa. For example, the $X_4 - Naa - X_{4'}$ residue positions X_2, $X_{2'}$, and $X_{3'}$ have a high frequency of Gly being 90.9%, 72.7%, and 63.6%, respectively (Figure 4a and Table S5). These observations are in agreement with enzyme motifs such as the "Gly-Xaa-Ser-Xaa-Gly", characteristic of $\alpha\beta$ hydrolase fold esterases,[47,48] where Ser is the nucleophile, Gly residues are dominant in X_2 and $X_{2'}$ and *Xaa* can accommodate any aa. Moreover, X_4 is rich in aliphatic Leu (27.3%), Ile (22.7%), and X_3 in aromatic Phe (27.3%), which are all nonpolar and large compared to Gly. Although the $X_4 - Baa - X_{4'}$ and $X_4 - Aaa - X_{4'}$ microenvironments show Gly dominance in various positions, it is less pronounced than in $X_4 - Naa - X_{4'}$. Interestingly, the $X_4 - Aaa - X_{4'}$ microenvironment shows the polar, negatively charged Asp (27.3%) in X_1. It is possible that Asp, through noncovalent polar interactions such as H-bonding, favors the proton exchanges that are required for the catalytic process to occur. Overall, in 53.2% of all $i = 4$ microenvironments, the positions near the active site residues are populated by Gly, Ala, Leu, Ile, Pro, and Val, which are nonpolar and relatively small residues. These residues are important for both the rigidity of the active site and for the correct protein folding, elucidating in which positions an aa can be placed to favor hydrogen bonds, disulfide bridges, and hydrophobic or other intramolecular interactions. For example, Leu, Ala, and Ile favor the formation of α-helices, due to their low helical penalty.[49] Ile and Val are often preferred β-sheet components in which Gly is known to produce bends.[50]

Box plots summarizing the main differences between environments of increasing lengths, obtained by subtracting $i = 4$ from $i = 8$ and $i = 8$ from $i = 16$ for each active aa, are shown in Figures 5, S5, and S6. This analysis allowed for the identification of residues that are favorable for the correct active site configuration and, consequently, for enzymatic activity. An aa is considered less favorable near the catalytic residue when its frequency decreases by increasing the microenvironment length and vice versa.

Resemblances between the microenvironment and the *full* sequences (Figure 6) were found for both aa composition and frequency, regardless of the microenvironment length ($i = \{4, 8, 16\}$). This finding reveals that even short sequences ($i = 4$), such as octapeptides, can match the composition profiles of the *full* sequences. Conversely, the differences between the *full* and *short* fragments (that is, hexapeptides) are statistically significant.

However, small changes in sequence composition between microenvironments should be interpreted with caution. Some aa are placed in more than one property category, and sometimes the change in frequency of one aa reflects on changes in several properties. In the case of charged residues in

$X_4 - Baa - X_{4'}$, the increase of charged aa is related only to acidic (negatively charged) residues.

Nucleophile (Naa) Microenvironment. A detailed description of the main patterns observed for the $X_4 - Naa - X_{4'}$, $X_8 - Naa - X_{8'}$ and $X_{16} - Naa - X_{16'}$ microenvironments is provided in Figures 5, S3a, and S4a. $X_4 - Naa - X_{4'}$ has a preference toward nonpolar, small, and tiny residues, mostly Gly (31.8% > 8.8%) and Ala (12.5% > 9.7%), followed by Leu (8% < 9.4%) and Val (5.7% < 7.3%) that show a slightly lower frequency compared to the *full* sequence (Figure 6a,b). X_2, $X_{2'}$, and $X_{3'}$ are mostly populated by Gly and Ala to avoid steric clashes with the substrate, while in X_4, X_3, and $X_{4'}$, that are further from the active nucleophile, the frequency of tiny amino acids diminishes. The frequency shifts to larger residues such as Phe (for example, 27.3% in X_3), implying that residues four positions away from the nucleophile are too far to cause steric interference with catalysis (Table S5). It is worth mentioning the unusual presence of the polar Asn (13.6%) in $X_{4'}$, since it is rare in *D2* triad enzymes. Moreover, a prevalence of Val (54.6%) and, generally, nonpolar aa (90.9%) was observed in X_6, contrasted by the neighboring X_7 and X_8 positions, which are dominated by polar residues. X_7 has a 36.4% prevalence of charged residues, of which 27.3% is Lys.

When comparing $X_4 - Naa - X_{4'}$ to $X_{16} - Naa - X_{16'}$, there is an increase in polar, charged, basic, and hydroxyl-containing residues, with a decrease in tiny and nonpolar properties, meaning that tiny and small residues are favored only in the closer proximity of the active site (Figures 5d,e, S3a, and S4a). All the microenvironments minimally varied in Glu, His, Leu, and Gln in comparison to *full* sequence frequencies indicating a "*chemical indifference*" toward the catalytic microenvironment. Interestingly, Trp, Pro, and Arg are absent, while Cys, Asp, Lys, Asn, Ser, and Thr are underrepresented (below 5%) in $X_4 - Naa - X_{4'}$ (Figure 6b).

We expected to find only small and nonpolar residues close to the nucleophile, due to the potential interference of charged and polar residues with the catalytic process. However, X_1 and $X_{1'}$ positions, adjacent to the nucleophile, allowed charged and polar residues. In contrast to the dominant nonpolar and aliphatic character of the microenvironment, X_1 is rich in polar, negatively charged Glu (27.3%) or Asp (18.2%), basic His (13.6%) or aromatic, nonpolar Tyr (18.2%). In *D2* enzymes, the X_1 position allows for 9 amino acids with varying properties and frequencies, with the most frequent being the negatively charged Glu (27.3%). Alternatively, $X_{1'}$ contains an equal frequency of nonpolar Ala and polar Gln (22.7%), followed by nonpolar Leu (18.2%), polar Thr (13.6%), and aromatic Tyr (13.6%) (Figure 4a). This suggests that X_1 and $X_{1'}$ residues may have an important role in catalysis. Additionally, the type of aa in X_1 is related to the oxyanion hole configuration. This is supported by the following lines of evidence:

- Asp is found in X_1 of all SGNH hydrolases, which have a three-point oxyanion hole;
- Glu is found in X_1 of all α/β hydrolase folds with three-point oxyanion holes;
- α/β hydrolase folds with two-point oxyanion holes have predominantly Tyr or His in X_1, which can be also populated by Phe, Gly, Ser, and Val.

SGNH hydrolases that contain Asp in X_1 are specific because their nucleophile has a double role. It uses the main-chain nitrogen atom as part of the oxyanion hole, while the side-chain oxygen/sulfur atom covers the nucleophile role. Throughout evolution, Asp is highly conserved in X_1 of SGNH hydrolases.[51−53] Moreover, all *D2* enzymes from the α/β hydrolase fold group with three-point oxyanion holes (one of them in $X_{1'}$) have a Glu in X_1 and show a dominant Phe-Gly-Glu-(Naa)-Ala-Gly-Gly/Ala motif. It should be noted that four out of six enzymes from this group have a sequence identity of more than 25% (Table S3). This implies that the motif could be the result of genetic factors rather than reflecting a catalytic pattern. α/β hydrolase domains that prefer His or Tyr, are those having two-point oxyanion holes with the nucleophile adjacent to the second oxyanion residue.

Position $X_{1'}$ was found to be the third oxyanion hole in 17 out of 22 enzymes in *D2*. In this case, the side chains of aa in $X_{1'}$ are rotated away from the nucleophile to allow their main-chain nitrogen to form part of the oxyanion hole. This could explain the presence of polar residues in $X_{1'}$, due to their ability to form hydrogen bonds, which would stabilize the active site configuration. It is worth noting that $X_{1'}$ oxyanion holes were mostly nonpolar, but also allowed aromatic and polar residues. Interestingly, $X_{1'}$ never accommodated charged residues in any of the *D2* enzymes (Table S5) or their homologues (Table S4).

Acid (Aaa) Microenvironment. Compared to the *full* sequence, the $X_4 - Aaa - X_{4'}$ microenvironment favors nonpolar aa (64.2% > 58.7%) and is rich in Gly (13.6% > 8.8%). In addition, it shows similar frequency of acidic (10.6% ∼ 10.2%) and basic (11% ∼ 9.1%) residues (Table S8). Ile (6.3% > 4.2%), Pro (7.4% > 6.2%), and Tyr (5.1% > 3.6%) are more frequent than in the *full* sequence (Figure 6c,d and Table S9), indicating that they may interfere with the catalytic process. Other chemical property trends are absent (Figure S5), except for the increase in sulfur containing aa (3.4% < 8.5%) Met and Cys, due to several enzymes having disulfide bonds near the catalytic acid.[54,55] $X_4 - Aaa - X_{4'}$ shows the dominance of Gly in $X_{1'}$ (36.4%) and X_3 (27.3%) but allows a wider range of aa in all positions, compared to $X_4 - Naa - X_{4'}$ (Figure 4b, and Table S9).

Interestingly, X_1 and $X_{4'}$ contain 27.3% Asp and Pro, respectively. The presence of Pro usually indicates the appearance of turns in the protein structure, while the frequent occurrence of Asp was not expected in X_1, adjacent to the active Aaa. Asp is one of the seven most common residues found in catalytic sites,[56] and its presence in X_1 may suggest a role in the catalytic mechanism or in improving the stability of the active site because of its potential to form additional hydrogen bonds with surrounding residues. All SGNH hydrolases in *D2* presented the basic His in $X_{3'}$. Moreover, the three enzymes with Glu in the Aaa role had Asp in X_1, Phe in $X_{4'}$, and Gly in $X_{1'}$, X_5, and X_{15}.

In $X_4 - Aaa - X_{4'}$, the C-terminal side is less polar than the N-terminal one, which is evident from positional frequencies of nonpolar residues in $X_{4'}$ (72.7%), $X_{3'}$ (63.6%), $X_{2'}$ (77.3%), and $X_{1'}$ (72.7%) compared to 54.6% in X_1, X_2, and X_3, and 63.6% in X_4 (Table S6). In larger acid microenvironments (Figures S3b and S4b), positions X_6 through X_9 have a high frequency of nonpolar amino acids corresponding to 77%, whereas the average nonpolar aa frequency for the $X_{16} - Aaa - X_{16'}$ microenvironment is 60.8%. Moreover, when compared to the $X_{16} - Aaa - X_{16'}$ average of 20.6%, a high frequency of charged residues was found in X_{16} (45.5%), X_{11} (36.4%), $X_{14'}$ (50%), and $X_{15'}$ (36.4%).

Table 2. Consensus Sequences Obtained from *D2* Data Set and Their Homologues (*D3*)[a]

	X4	X3	X2	X1	Naa	X1'	X2'	X3'	X4'	Chi-square test	p-value
***D2* consensus**	Leu	Phe	Gly	Glu	Ser	Gln/Ala	Gly	Gly	Ala	**113.926**	1.41×10^{-21}
Frequency (%)	27.3	27.3	90.9	27.3	95.5	22.7	72.7	63.6	22.7		
***D3* consensus**	Leu	Val/Phe	Gly	Asp	Ser	Gln/Ala	Gly	Gly	Ala	/	/
Frequency (%)	31.8	18.2	100.0	27.3	95.5	27.3	72.7	50.0	27.3		
	++	+	++	+		++	+++	++	++		

	X4	X3	X2	X1	Aaa	X1'	X2'	X3'	X4'	Chi-square test	p-value
***D2* consensus**	Thr/Gly	Gly	Leu/Gly	Asp	Asp	Gly	Ile	Val	Pro	**51.085**	8.84×10^{-9}
Frequency (%)	13.6	27.3	13.6	27.3	81.8	36.4	18.2	22.7	27.3		
***D3* consensus**	Val	Gly	Pro/Gly	Asp	Asp	Gly	Val/Gly	Trp/His	Pro	/	/
Frequency (%)	27.3	22.7	18.2	31.8	72.7	36.4	18.2	18.2	27.3		
	+	++	+	++		+++	+		+++		

	X4	X3	X2	X1	Baa	X1'	X2'	X3'	X4'	Chi-square test	p-value
***D2* consensus**	Ile	Gly	Ala	Pro/Gly	His	Leu/Gly	Asn	Xaa*	Leu	**37.295**	4.12×10^{-6}
Frequency (%)	13.6	36.4	22.7	13.6	100.0	18.2	18.2	13.6	27.3		
***D3* consensus**	Val	Gly	Ala	Asp	His	Gly	Asp	Glu	Ile	/	/
Frequency (%)	27.3	40.9	31.8	22.7	90.9	27.3	18.2	22.7	18.2		
	+	++	++	+		+	+	+	+		

[a]The χ^2 test results and *p*-values are provided for each microenvironment-based consensus sequence. Additional observations: + shared chemical property; + + identical in aa, different frequency; + ++ identical in aa and frequency; * Xaa corresponds to Glu, Asp, Leu and Pro, sharing the highest frequency.

A comparative analysis has shown that increasing the microenvironment size from $X_4 - Aaa - X_{4'}$ to $X_{16} - Aaa - X_{16'}$ results in the increase of polar (38.4% > 35.8%) and aliphatic (31% > 23.3%) residues. In addition, a decrease in small (54.7% < 56.3%), aromatic (13.2% < 17.1%), nonpolar (61.7% < 64.2%), and sulfur (4.4% < 8.5%) properties (Figure 6) confirms the existence of the Aaa microenvironment.

Base (Baa) Microenvironment. The $X_i - Baa - X_i'$ microenvironment shows less aa-specific positional patterns compared to the nucleophile and acid ones (Figures S3c, S4c, and S6). In X_3 and $X_{1'}$, there is a prevalence of Gly with frequencies of 36.4% and 18.2%, respectively (Table S7). $X_4 - Baa - X_{4'}$ exhibits low charged (13.6% or less for 4 out of 8 positions) property frequencies near the active His (Baa). Interestingly, a lower basic residue frequency (5.3%) is observed as opposed to the 2.2 times higher acidic frequency (11.4%) in the same positional range. Moreover, the frequency of charged residues (Table S7) in positions X_2 (His), $X_{1'}$ (Glu) and X_5 (Asp) is low (4.6% each) compared to both the $X_{16} - Baa - X_{16'}$ (21.6%) and *full* sequence (21.6%) averages (Table S8). The frequency of acidic residues in $X_i - Baa - X_i'$ microenvironments of different lengths is similar, implying that basic residues are not favorable in close proximity of the active base. Consequently, the reduced frequency of charged residues in $X_4 - Baa - X_{4'}$, is a result of the reduced basic frequency alone.

Frequency differences are more pronounced when comparing the average values of microenvironments of different lengths (Figures 4c, 6e,f) which confirms the existence of the Baa microenvironment. In $X_4 - Baa - X_{4'}$, tiny (34.7% > 32%), small (58% > 54.8%) and nonpolar (65.3% > 58.7%) aa are more frequent, while charged (18.2% < 21.6%), polar (34.7% < 41.3%) and basic aa (7.4% < 11%) are less frequent than in the *full* sequence. The N-terminal side of $X_4 - Baa - X_{4'}$ contains more nonpolar residues (71.6%) than the C-terminal side (60.2%). Similarly to other $X_i = 4$ microenvironments, Gly (13.6%), Ala (10.8%), and Leu (10.2%) are the most frequent (Figure 6).

Consensus Motifs Based on Identified Patterns. The identified patterns near the catalytic triad residues confirmed the existence of nucleophile, acid, and base microenvironments in *D2* enzymes. Gly is favored in all the triad microenvironments, especially around the Naa, in agreement with the dominant Gly-Xaa-Ser-Xaa-Gly enzyme motif in α/β hydrolase folds.[47,48] The X_1 position, adjacent to the active residues, stands out in all microenvironments as it often contains charged aa (Glu, Asp, His) when adjacent to the Naa, the acidic Asp when adjacent to the Aaa, and an oscillating polar/nonpolar character in proximity of Baa. This suggests that the triad hydrolytic mechanism might have additional layers of complexity in the evolutionary or microenvironment sense, prompting new questions for further research.[57] Furthermore, positions directly adjacent to the Naa residue are likely used for substrate stabilization and selection based on their properties and codependencies on oxyanion hole and substrate types.

On the basis of the microenvironment analysis, we propose consensus motifs (Table 2) that could be useful for their insertion into scaffolds or further modeled *in silico*. Additionally, they could be modified at their N-terminus by introducing known self-assembly promoting motifs to allow the formation of nanostructures. According to the selection criteria described above, the *D2*-based consensus sequences are as follows:

- Leu:Phe:Gly:Glu-**Naa**-Gln/Ala:Gly:Gly:Ala
- Thr/Gly:Gly:Leu/Gly:Asp-**Aaa**-Gly:Ile:Val:Pro
- Ile:Gly:Ala:Pro/Gly-**Baa**-Leu/Gly:Asn:Xaa:Leu

The proposed consensus sequences were analyzed with the appropriate test of statistical significance to estimate whether the observations are due to chance. The goodness-of-fit χ^2 test was used to evaluate the null hypothesis that the observed distribution of dominant amino acids per position fits their random distribution. The observed distributions were computed as relative frequencies of every aa at each position within the microenvironment $(O(aa)_i = freq(aa)_i/n$, where i represents positions from X_4 to $X_{4'}$ and $n = 22$ is the number of

triad-containing ester hydrolases). The aa with the highest rate of appearance ($\max(O(aa)_i)$) was then compared with its expected random distribution ($E(aa) = freq(aa)/N$, where $N = 176$ represents the total number of observations), which is equal to the relative frequency in the whole microenvironment. The χ^2 value was computed using the expression 1 and compared to the critical value at 1% significance level for $n - 1 = 7$ degrees of freedom, which is equal to 18.475.

$$\chi^2 = \sum_{i=1}^{n} \frac{(\max(O(aa)_i) - E(aa))^2}{E(aa)} \quad (1)$$

The χ^2 test confirmed that the frequencies of dominant amino acids are significantly different (with p-value < 0.01) from their random appearance in the whole microenvironment, indicating that the consensus sequences are representative for the 22 selected examples of triad-containing ester hydrolases.

Moreover, to reinforce the confidence of our results, we used 974 homologues (*D3*) of *D2* enzymes (Table 2) to assess whether the variation of the aa surrounding the triad will be affected (Figure S9). A consensus sequence for each microenvironment ($X_4 - Xaa - X_{4'}$) was obtained, based on the most frequent residues in each position. It is evident that the Naa consensus sequences overlap, except in X_1, where Glu is replaced by Asp, however maintaining the physicochemical property (polar, acidic aa). For Aaa and Baa microenvironments 60% and 50% consensus sequence overlap can be noticed, respectively (Table 2). Despite slight variations in aa patterns (similarity in 16 out of 24 positions), the physicochemical properties of these sequences fully overlapped in 23 out of 24 positions.

These dissimilarities could be due to the heterogeneous number of homologues for each *D2* entry, ranging from 3 to 100 (Table S4), which could create an unintentional bias toward a particular mechanism. For example, 1ESC included homologues with different mechanisms, resulting in a consensus sequence that inaccurately represented the initial *D2* enzyme (Table S4). Furthermore, some homologues were noncatalytic, had altered catalytic function, or had the catalytic site relocated to another part of the sequence, resulting in inaccurate alignment or misalignment of active sites. This was evident from catalytic triad residues being replaced by aa that are unlikely to have a catalytic function,[56] such as Leu residues replacing His (Table S4). Some of these issues could be avoided by selecting only the homologues that have the same EC number as the reference sequence. However, this approach would solve only partly the issue as it would reduce the sample size. As a possible future direction, identified mutations that lead to misalignment could be manually aligned and filtered by EC number.

CONCLUSIONS

In this paper, we provide a thorough analysis of the aa sequence pertaining the catalytic site and oxyanion hole of 22 ester hydrolases (EC 3.1) that contain catalytic triads. We focused on the primary aa sequence and chemical character, and analyzed fragments of different lengths based on *important* residues, which take part in the triad and/or in the oxyanion hole. In all examples, the triad follows the sequential Naa:Aaa:Baa order from the N- to the C-terminus, except for the enzyme with 3.40.50.180 CATH, which showed the Naa:Baa:Aaa disposition at the primary sequence level.

The provided *short* fragments (Table S2) can be easily synthesized and experimentally validated for ester hydrolysis. Alternatively, they can be used for further development of catalytic peptides for insertion into scaffolds or for self-assembly into supramolecular and biocatalytic nanostructures. A drawback of the fragment approach that intends to maintain fidelity to the enzymatic aa sequence, is the large in-sequence distance, spanning almost half of the full enzyme (45%) for fragments based on triads (*medium*), and over half (62.5%) of it, for fragments including triads and oxyanion holes (*long*). As a result, such fragments could be rather costly to produce. Additionally, *short* sequences are biased toward polar triad components, and may not reproduce the chemical microenvironment found near important residues on enzymes, with potential loss in catalytic performance.

The existence of a microenvironment in chemical property and aa frequency trends led to the identification of consensus motifs (Table 2). These motifs inform us about which chemical environments, which aa and their corresponding positions favor catalysis, and what are the acceptable variations in the proximity of triad residues. The obtained motifs can be further modeled in silico to search for optimal geometries.

Overall, this study was performed on a subset of esterases with catalytic triads having unique mechanisms. A wider study, involving a larger number of enzymes and an automatized search approach seems beneficial, and it will be part of future studies. Moreover, it would be of interest to extend such an analysis to other EC 3 subclasses of enzymes, such as proteases.

DATA AND SOFTWARE AVAILABILITY

Publicly available M-CSA www.ebi.ac.uk/thornton-srv/m-csa is the source of data about the active sites, while sequences were taken from Uniprot www.uniprot.org/. The data was manually curated as described in section data sets. Crystallographic data was downloaded in PDB format from RCSB-PDB www.rcsb.org/.

PyMol software www.pymol.org/2/ was used to mark the active residues. The R-package "Peptides" was used to compute aa composition and physicochemical properties of peptides, as described in section Properties calculation.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00977.

Physico-chemical properties of the enzymes (*full* sequence) for data sets *D1* and *D2* (Figure S1), results of the Friedman statistical test of significance followed by the post hoc tests with Holm-Bonferroni correction (Figure S2), data set *D2* including information on catalytic triads and oxyanion holes for each enzyme (Table S1), list of *short* fragments obtained from *full* sequences in *D2*, taking into account important residues (Table S2), sequence logos for the three microenvironments (Figures S3-S4), further description of Aaa (Figure S5) and Baa (Figure S6) microenvironments, box plots showing physicochemical properties for *full*, *long*, *medium* and *short* fragments (Figure S7), identity matrix of *D2* enzymes (Table S3), schematic representation of the consensus sequences determination based on homologues (Figure S9), homologues-based con-

sensus sequences for each microenvironment (Table S4), positional frequency tables for *Xi* = 4 Naa (Table S5), Aaa (Table S6) and Baa (Table S7) microenvironments, ester hydrolysis reaction mechanism (Figure S8), chemical properties of $X_4 - Xaa - X_{4'}$ microenvironments and *full* sequence (Table S8), and aa frequency of $X_4 - Xaa - X_{4'}$ microenvironments and *full* sequence (Table S9) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Goran Mauša − *Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia;* orcid.org/0000-0002-0643-4577; Email: gmausa@riteh.hr

Daniela Kalafatovic − *Department of Biotechnology, University of Rijeka, 51000 Rijeka, Croatia; Center for Advanced Computing and Modeling, University of Rijeka, 51000 Rijeka, Croatia;* orcid.org/0000-0002-9685-1162; Email: daniela.kalafatovic@uniri.hr

Authors

Marko Babić − *Department of Biotechnology, University of Rijeka, 51000 Rijeka, Croatia;* orcid.org/0000-0003-2300-8825

Patrizia Janković − *Department of Biotechnology, University of Rijeka, 51000 Rijeka, Croatia;* orcid.org/0000-0002-8904-4004

Silvia Marchesan − *Chemical and Pharmaceutical Sciences Department, University of Trieste, 34127 Trieste, Italy;* orcid.org/0000-0001-6089-3873

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

With this work the authors give a contribution and support the advancing of women in chemistry. This research was funded by the Croatian Science Foundation/Hrvatska zaklada za znanost, grant numbers UIP-2019-04-7999 (D.K., P.J.) and DOK-2021-02-3496 (D.K., M.B.), and by the University of Rijeka, grant number uniri-pr-tehnic-19-10 (G.M.). The authors acknowledge the Centre for Artificial intelligence and cyber security (AIRI) at the University of Rijeka.

REFERENCES

(1) Kraut, J. How Do Enzymes Work? *Science* **1988**, *242*, 533−540.

(2) Drienovska, I.; Roelfes, G. Expanding the Enzyme Universe with Genetically Encoded Unnatural Amino Acids. *Nat. Catal.* **2020**, *3*, 193−202.

(3) Bhatia, S.*Introduction to Pharmaceutical Biotechnology*, Volume *2*; IOP Publishing: Bristol, U.K., 2018; pp 2053-2563.

(4) McDonald, A. G.; Tipton, K. F. Enzyme Nomenclature and Classification: the State of the Art. *FEBS J.* **2021**, 1−18.

(5) ExplorEnz website. Available online: https://www.enzyme-database.org/, (last accessed on 18−01−22).

(6) McDonald, A. G.; Boyce, S.; Tipton, K. F. ExplorEnz: the Primary Source of the IUBMB Enzyme List. *Nucleic Acids Res.* **2009**, *37*, D593−D597.

(7) Alcántara, A. R.; Hernaiz, M.-J.; Sinisterra, J.-V.*Comprehensive Biotechnology* (Second Edition); Academic Press: Burlington, 2011; pp 309−331.

(8) Polgár, L. The Catalytic Triad of Serine Peptidases. *Cell. Mol. Life Sci.* **2005**, *62*, 2161−2172.

(9) Dimitriou, P. S.; Denesyuk, A. I.; Nakayama, T.; Johnson, M. S.; Denessiouk, K. Distinctive structural motifs co-ordinate the catalytic nucleophile and the residues of the oxyanion hole in the alpha/beta-hydrolase fold enzymes. *Protein Sci.* **2019**, *28*, 344−364.

(10) Simón, L.; Goodman, J. M. Enzyme Catalysis by Hydrogen Bonds: The Balance between Transition State Binding and Substrate Binding in Oxyanion Holes. *J. Org. Chem.* **2010**, *75*, 1831−1840.

(11) Zhang, Y.; Kua, J.; McCammon, J. A. Role of the Catalytic Triad and Oxyanion Hole in Acetylcholinesterase Catalysis: An ab initio QM/MM Study. *J. Am. Chem. Soc.* **2002**, *124*, 10572−10577.

(12) Lin, J.-L.; Wagner, J. M.; Alper, H. S. Enabling tools for high-throughput detection of metabolites: Metabolic engineering and directed evolution applications. *Biotechnol. Adv.* **2017**, *35*, 950−970.

(13) Bunzel, H. A.; Anderson, J. R.; Mulholland, A. J. Designing better enzymes: Insights from directed evolution. *Curr. Opin. Struct. Biol.* **2021**, *67*, 212−218.

(14) Richter, F.; Blomberg, R.; Khare, S. D.; Kiss, G.; Kuzin, A. P.; Smith, A. J. T.; Gallaher, J.; Pianowski, Z.; Helgeson, R. C.; Grjasnow, A.; Xiao, R.; Seetharaman, J.; Su, M.; Vorobiev, S.; Lew, S.; Forouhar, F.; Kornhaber, G. J.; Hunt, J. F.; Montelione, G. T.; Tong, L.; Houk, K. N.; Hilvert, D.; Baker, D. Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **2012**, *134*, 16197−16206.

(15) Zozulia, O.; Dolan, M.; Korendovych, I. Catalytic Peptide Assemblies. *Chem Soc Rev* **2018**, *47*, 3621−3639.

(16) Garcia, A.; Kurbasic, M.; Kralj, S.; Melchionna, M.; Marchesan, S. A biocatalytic and thermoreversible hydrogel from a histidine-containing tripeptide. *Chem. Commun.* **2017**, *53*, 8110−8113.

(17) Kleinsmann, A. J.; Nachtsheim, B. J. A minimalistic hydrolase based on co-assembled cyclic dipeptides. *Org Biomol Chem* **2020**, *18*, 102−107.

(18) Huang, Z.; Guan, S.; Wang, Y.; Shi, G.; Cao, L.; Gao, Y.; Dong, Z.; Xu, J.; Luo, Q.; Liu, J. Self-assembly of amphiphilic peptides into bio-functionalized nanotubes: a novel hydrolase model. *J. Mater. Chem. B* **2013**, *1*, 2297−2304.

(19) Han, J.; Gong, H.; Ren, X.; Yan, X. Supramolecular nanozymes based on peptide self-assembly for biomimetic catalysis. *Nano Today* **2021**, *41*, 101295.

(20) Poliakoff, M.; Licence, P. Green chemistry. *Nature* **2007**, *450*, 810−812.

(21) Li, Y.; Zhao, Y.; Hatfield, S.; Wan, R.; Zhu, Q.; Li, X.; McMills, M.; Ma, Y.; Li, J.; Brown, K. L Dipeptide seryl-histidine and related oligopeptides cleave DNA, protein, and a carboxyl ester. *Bioorgan. Med. Chem.* **2000**, *8*, 2675−2680.

(22) Duncan, K. L.; Ulijn, R. V. Short Peptides in Minimalistic Biocatalyst Design. *Biocatalysis* **2015**, *1*, 67−81.

(23) Zhang, C.; Shafi, R.; Lampel, A.; MacPherson, D.; Pappas, C. G.; Narang, V.; Wang, T.; Maldarelli, C.; Ulijn, R. V. Switchable Hydrolase Based On Reversible Formation Of Supramolecular Catalytic Site Using A Self-Assembling Peptide. *Angew. Chem., Int. Ed.* **2017**, *129*, 14703−14707.

(24) Baruch-Leshem, A.; Chevallard, C.; Gobeaux, F.; Guenoun, P.; Daillant, J.; Fontaine, P.; Goldmann, M.; Kushmaro, A.; Rapaport, H. Catalytically active peptides affected by self-assembly and residues order. *Colloids Surf., B* **2021**, *203*, 111751.

(25) Takahashi, T.; Cheung, M.; Butterweck, T.; Schankweiler, S.; Heller, M. J. Quest for a turnover mechanism in peptide-based enzyme mimics. *Catal Commun* **2015**, *59*, 206−210.

(26) Maeda, Y.; Javid, N.; Duncan, K.; Birchall, L.; Gibson, K. F.; Cannon, D.; Kanetsuki, Y.; Knapp, C.; Tuttle, T.; Ulijn, R. V.; Matsui, H. Discovery of Catalytic Phages by Biocatalytic Self-Assembly. *J. Am. Chem. Soc.* **2014**, *136*, 15893−15896.

(27) Gulseren, G.; Khalily, M. A.; Tekinay, A. B.; Guler, M. O. Catalytic supramolecular self-assembled peptide nanostructures for ester hydrolysis. *J. Mater. Chem. B* **2016**, *4*, 4605−4611.

(28) Kurbasic, M.; Garcia, A. M.; Viada, S.; Marchesan, S. Heterochiral tetrapeptide self-assembly into hydrogel biomaterials for hydrolase mimicry. *J. Pept. Sci.* **2022**, *28*, No. e3304.
(29) Kurbasic, M.; Garcia, A. M.; Viada, S.; Marchesan, S. Tripeptide Self-Assembly into Bioactive Hydrogels: Effects of Terminus Modification on Biocatalysis. *Molecules* **2021**, *26*, 173.
(30) Carlomagno, T.; Cringoli, M. C.; Kralj, S.; Kurbasic, M.; Fornasiero, P.; Pengo, P.; Marchesan, S. Biocatalysis of D, L-peptide nanofibrillar hydrogel. *Molecules* **2020**, *25*, 2995.
(31) Leščić Ašler, I.; Ivić, N.; Kovačić, F.; Schell, S.; Knorr, J.; Krauss, U.; Wilhelm, S.; Kojić-Prodić, B.; Jaeger, K.-E. Probing Enzyme Promiscuity of SGNH Hydrolases. *ChemBioChem* **2010**, *11*, 2158−2167.
(32) Leveson-Gower, R. B.; Mayer, C.; Roelfes, G. The importance of catalytic promiscuity for enzyme design and evolution. *Nat. Rev. Chem.* **2019**, *3*, 687−705.
(33) Ribeiro, A. J. M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* **2018**, *46*, D618−D623.
(34) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **2004**, *14*, 1188−1190.
(35) Schneider, T. D.; Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **1990**, *18*, 6097−6100.
(36) Anthis, N. J.; Clore, G. M. Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm. *Protein Sci.* **2013**, *22*, 851−858.
(37) Osorio, D.; Rondón-Villarreal, P.; Torres, R. Peptides: A package for data mining of antimicrobial peptides. *Small* **2015**, *12*, 44−444.
(38) Kalafatovic, D.; Mauša, G.; Rešetar Maslov, D.; Giralt, E.Bottom-Up Design Approach for OBOC Peptide Libraries. *Molecules***2020**, *25*.3316
(39) Cruciani, G.; Baroni, M.; Carosati, E.; Clementi, M.; Valigi, R.; Clementi, S. Peptide studies by means of principal properties of amino acids derived from MIF descriptors. *J Chemometr* **2004**, *18*, 146−155.
(40) Guruprasad, K.; Reddy, B. B.; Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng. Des. Sel.* **1990**, *4*, 155−161.
(41) Eisenberg, D.; Schwarz, E.; Komaromy, M.; Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **1984**, *179*, 125−142.
(42) Boman, H. G. Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.* **2003**, *254*, 197−215.
(43) Nelson, D.; Cox, M.*Lehninger Principles of Biochemistry*; Macmillan: London, 2005.
(44) Denesyuk, A.; Dimitriou, P. S.; Johnson, M. S.; Nakayama, T.; Denessiouk, K. The acid-base-nucleophile catalytic triad in ABH-fold enzymes is coordinated by a set of structural elements. *PLOS One* **2020**, *15*, No. e0229376.
(45) Dodson, G.; Wlodawer, A. Catalytic triads and their relatives. *Trends Biochem. Sci.* **1998**, *23*, 347−352.
(46) Blow, D. M.; Birktoft, J. J.; Hartley, B. S. Role of a Buried Acid Group in the Mechanism of Action of Chymotrypsin. *Nature* **1969**, *221*, 337−340.
(47) Oh, C.; Kim, T. D.; Kim, K. K. Carboxylic Ester Hydrolases in Bacteria: Active Site, Structure, Function and Application. *Crystals* **2019**, *9*, 597.
(48) Ollis, D. L.; Cheah, E.; Cygler, M.; Dijkstra, B.; Frolow, F.; Franken, S. M.; Harel, M.; Remington, S. J.; Silman, I.; Schrag, J.; Sussman, J. L.; Verschueren, K. H.; Goldman, A. The α/β hydrolase fold. *Protein Eng. Des. Sel.* **1992**, *5*, 197−211.
(49) Nick Pace, C.; Martin Scholtz, J. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophys. J.* **1998**, *75*, 422−427.
(50) Richardson, J. S.; Richardson, D. C. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl. Acad. Sci.* **2002**, *99*, 2754−2759.
(51) Yang, J.; Zhang, Y.; Xu, J.; Geng, Y.; Chen, X.; Yang, H.; Wang, S.; Wang, H.; Jiang, X.; Guo, X.; Zhao, G. Serum Activity of Platelet-Activating Factor Acetylhydrolase Is a Potential Clinical Marker for Leptospirosis Pulmonary Hemorrhage. *PLOS One* **2009**, *4* (1), 1−11.
(52) Lo, Y.-C.; Lin, S.-C.; Shaw, J.-F.; Liaw, Y.-C. Crystal Structure of Escherichia coli Thioesterase I/Protease I/Lysophospholipase L1: Consensus Sequence Blocks Constitute the Catalytic Center of SGNH-hydrolases through a Conserved Hydrogen Bond Network. *J. Mol. Biol.* **2003**, *330*, 539−551.
(53) Mølgaard, A.; Kauppinen, S.; Larsen, S. Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases. *Structure* **2000**, *8*, 373−383.
(54) Matak, M. Y.; Moghaddam, M. E. The role of short-range Cys171−Cys178 disulfide bond in maintaining cutinase active site integrity: A molecular dynamics simulation. *Biochem. Biophys. Res. Commun.* **2009**, *390*, 201−204.
(55) Komiya, D.; Hori, A.; Ishida, T.; Igarashi, K.; Samejima, M.; Koseki, T.; Fushinobu, S. Crystal Structure and Substrate Specificity Modification of Acetyl Xylan Esterase from Aspergillus luchuensis. *Appl. Environ. Microbiol.* **2017**, *83*, e01251−17.
(56) Holliday, G. L.; Mitchell, J. B.; Thornton, J. M. Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis. *J. Mol. Biol.* **2009**, *390*, 560−577.
(57) Gagler, D. C.; Karas, B.; Kempes, C. P.; Malloy, J.; Mierzejewski, V.; Goldman, A. D.; Kim, H.; Walker, S. I. Scaling laws in enzyme function reveal a new kind of biochemical universality. *Proc. Natl. Acad. Sci. U.S.A.* **2022**, *119*, No. e2106655119.