Gregor Paul Wirnsberger, BSc

Classification of alpha/beta-hydrolase domain containing proteins using 3D point clouds

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme:

Biochemistry and Molecular Biomedicine

submitted to

Graz University of Technology

Supervisor

Univ.- Prof. Dr. Karl Gruber

Institute of Molecular Bioscience

Dr. Christian Gruber

Institute of Molecular Bioscience

Graz, July 2021

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

16.07.2021, Gregor Wirnsberger E.h.

Date, Signature

Acknowledgment

First of all, I would like to express my gratitude towards both of my supervisors Univ.-Prof. Dr. Karl Gruber and Dr. Christian Gruber for their constant support and supervision whenever I had questions regarding my thesis or troubles I had to face. I would like to thank Priv.-Doz. Dr. Andreas Krassnigg, Dr. Georg Steinkellner and Michael Hetmann, MSc as well, who did their best to help me. Special thanks goes to the whole team of Innophore for the computational resources, their expertise and the interesting discussions. Furthermore, I want to thank Assoz. Prof. Dr. Zimmermann for providing the assay data and the following discussions as well as Univ.-Prof. Dr. Dagmar Kratky for the discussion regarding assays and general background information about the proteins involved in this thesis. I am grateful for the fact that that I was able to incorporate my own ideas on the approach and on problem solving while gathering the data I was interested in.

On the private side, I would like to thank my parents Andrea and Karlheinz as well as my girlfriend Kristin, who always supported and motivated me, even in times when I was not so companionable due to some persistent bugs in my code.

Abstract

Comparing or annotating proteins based on their sequence identity can work if proteins feature a high sequence identity. Another way of comparing proteins is possible on the structural level. One can compare the active site of the proteins of interest and make suggestions on their activity and substrate spectrum. But what happens if the sequence identity is not very high and there are only few to no crystal structures available? To battle this circumstance, we used different (bio)informatic tools on a data set of 210 possible α-β hydrolase domain (ABHD) containing proteins. On the one hand database scraping of the Uniprot database (5) and clustering proteins on the obtained data was done. On the other hand, homology models were built, the properties of the active site were calculated and the proteins were compared based on these data. In this process, algorithms to search for active site motifs in proteins were created. Clustering based on the database entries resulted in 13 different clusters. For 185 of the 210 protein sequences, homology models could be generated, and for 119 of them, an active site cavity could be determined. The clustering based on these 119 proteins resulted in 17 clusters. All 210 proteins were clustered based on their activity in different assays which lead to the formation of 22 clusters. A comparison of all results from all three approaches showed 8 different clusters, of which, based on their annotated catalytic activity in the Uniprot, two clusters present heterogenous results, four show homogenous results and two could not be assessed. For three proteins, which lack in an annotation of their catalytic activity, their catalytic activity can be predicted based on the match in clustering by the three approaches mentioned above.

Content

Introduction

The α/β-hydrolase fold is a fold that is often found in different hydrolytic enzymes which can have a widely differing phylogenetic origin and catalytic function. It consists of six α-helices and eight β-strands (Figure 1). A core feature of this fold is the nucleophilic elbow. This is a beat strand followed by a turn and an α helix. The nucleophile of the catalytic triad is located on this turn. The arrangement in the catalytic triad is preserved in this fold. It is commonly found in the loop region whereas the binding site is not preserved. What the proteins featuring this specific fold have in common is the catalytic triad. Here only the histidine in the nucleophile-histidine-acidic catalytic triad is completely conserved, whereas the other two residues can be formed by different types of amino acids. This motif resembles through its geometry and its sequence arrangement a mirror-image of the serin protease catalytic triad. (1).

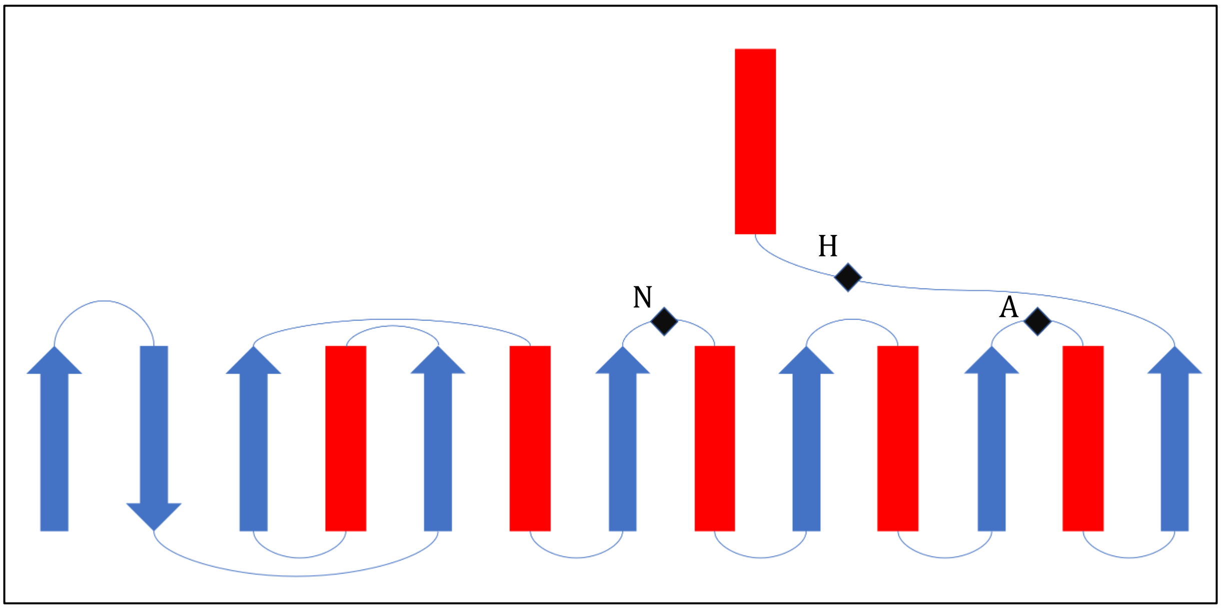

Figure 1: Modified topology diagram of the prototypic α/β-hydrolase fold as published in (1) (Page 200 Fig. 3). Blue arrows indicate β-strands, red rectangle represent α-helices and black squares indicate the triad position with its residues labelled as N for nucleophile, H for Histidine and A for the acidic residue.

The presence of serine hydrolases is very frequent in mammals and feature a wide variety of catalytic activities. They are involved in almost all physiological processes and are drug targets for the treatment of various diseases such as diabetes, obesity and neurodegenerative disorders. (2) All proteins used in this thesis can be found in the mouse (Mus musculus).

The catalytic mechanism in serine hydrolases involves the activation of a conserved serine nucleophile in order to attack the substrate which can be an ester, thioester or amide bond to form an acyl-enzyme intermediate. This proceeds into hydrolysis of this specific intermediate followed by the release of the product (3).

For this thesis, a set of 210 proteins were considered. These proteins are members of the serine hydrolase superfamily and are involved in energy and lipid metabolism as well as membrane physiology and signalling.

One of the biggest challenges in computational biology, is to determine the 3D structure of proteins based on their amino-acid sequence. Programs and different approaches have been invented to solve this problem. The latest and greatest program called DeepMind's AlphaFold 2 showed its capability in the CASP14 event where it outperformed every other method through the use of an artificial intelligence network to predict structures (4). Since this approach is not yet available for most scientists, the program Yasara was used to build the homology models needed for further exploration of the proteins in this thesis. Here a PSI-BLAST search against the Uniprot database (5) is performed to generate a position specific scoring matrix, which in turn is then used to search the protein data bank (6) for suitable templates. These then get ranked according to their alignment score and their structural quality. In the process of choosing the right template, different factors are considered. If the template is an oligomer, the target will be built as an oligomer to conserve the side-chain interactions. To find the initial rotamer solution regarding the simple repulsive energy function, dead-end elimination is used after the building process of the graph of the side chain rotamer network. Loop optimization is done by trying out different conformations and reoptimizing the side chain for each of them. The side chain rotamers are enhanced regarding electrostatics, packing interactions and solvation effect. The model is then refined by using an unrestrained high-resolution refinement with explicit solvent molecules. This method applies to all combinations of templates and alignments. To receive the best model possible unsuitable regions are eliminated and a hybrid model is built with better fitting fragments from other models. (7)

Another option for predicting the catalytic activity of a protein is to compute the active site and its properties. To find the active site cavity, an algorithm designed for predicting ligand binding site can be used, which in this case is the LIGSITE algorithm. (8) Here the protein gets mapped onto a 3D grid. Points that are closer or as close as 3 Å to the coordinates of atoms of the protein of interest are counted as protein points, otherwise as solvent. Solvent points, which are accounted as cavities, need to be in a sequence of grid points that starts and ends with protein points and encompass the solvent grid points in between. This is tested for the x-, y-, z- axis as well as for the cubic axis. If a solvent point exceeds the threshold of valid axes, it will be used in the final cavity prediction. In the end, points that are tagged as cavity points get clustered based on their spatial distance. If a cavity point is in distance of 3 Å or closer to another cavity point, both of them get clustered, otherwise a new cluster is created. (8) For all the validated cavity points, their physicochemical properties are calculated based on their surrounding amino acids to obtain point clouds that represent the active site.

Almost all analysis were done by creating scripts and programs in python. Python is a powerful, fast, and open programming language that runs on every operating system. Its usage is versatile and ranges from Web and Internet Development, Scientific and Numeric computing to networking and more (9).

Aims and Hypothesis

The aim of this thesis is to cluster and characterise proteins potentially containing an ABHD fold on a structural level, even though there are only few crystal structures available. In order to accomplish this task, a set of bioinformatical tools was used. The idea is to compare proteins on their structural level, more specifically using the properties of their active sites. Homology modelling can be used to create models that represent the structures of proteins, thus to enable the calculation of properties regarding the active site of proteins where no experimental structure has been solved. The characteristics mentioned are then compared to find the most similar proteins. Moreover, if it is possible to find a cluster with matching proteins regarding their catalytic activity and feature a protein without an annotated catalytic activity, the prediction of the catalytic activity for this specific protein is feasible.

Results

Sequence identity

Based on the fact that the amino acid sequences of all proteins are known, a multiple sequence alignment was done. Of these 210 sequences, 13 pairs of 20 unique proteins feature a sequence identity greater than 80%, which makes 9.5% of all sequences. The mean sequence identity of all proteins against each other is 6.04%, with the lowest overlap in sequence (beside 0% and therefore not overlapping at all) is 1.12% and the highest overlap (beside being 100% and therefore being the same) is 97.93%

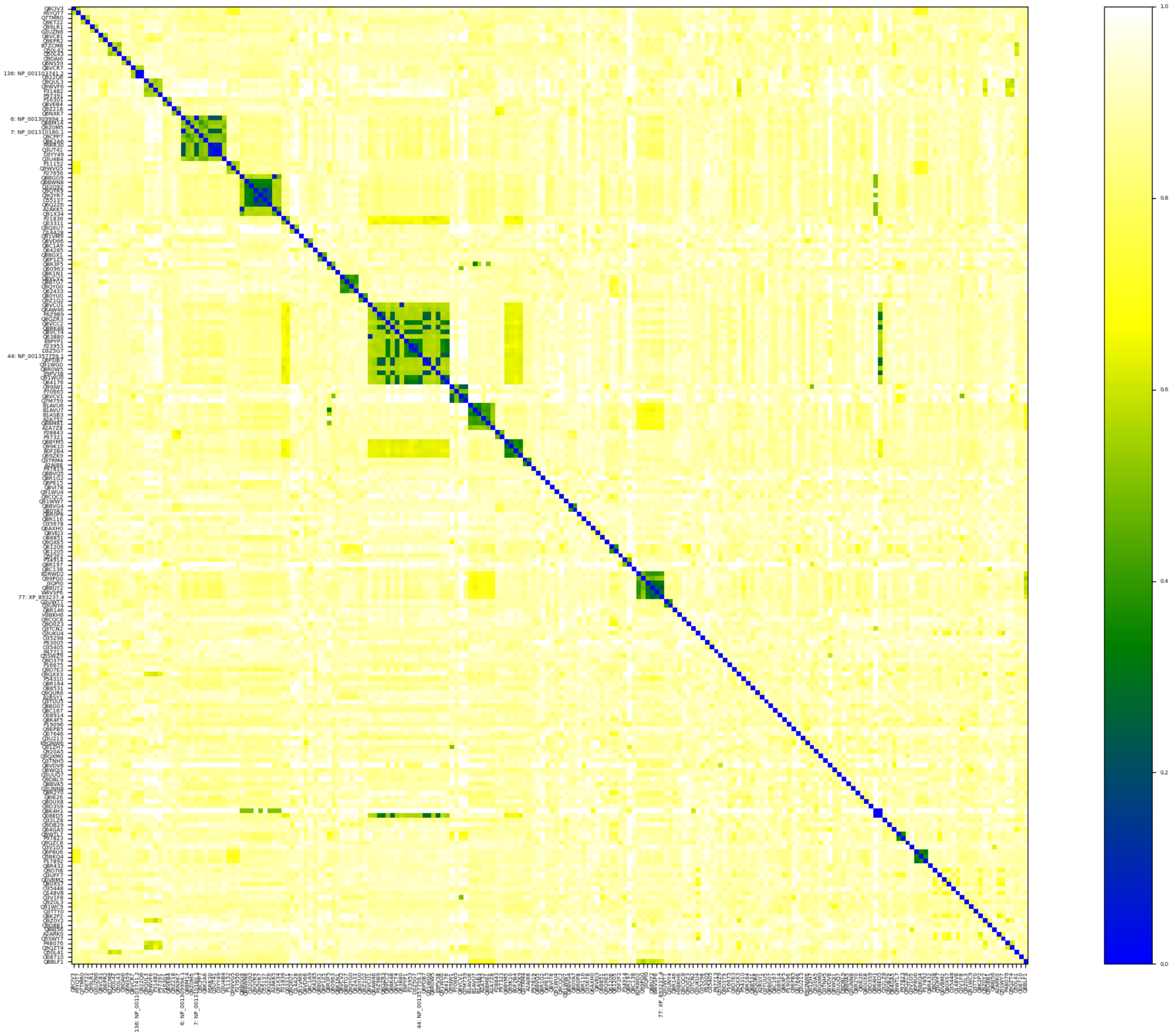

Figure 2: **Heatmap of the sequence identity of all proteins** where 0 represents identical sequences and 1 features no overlaps in the multiple sequence alignment from Clustal w. This shows a small overall sequence identity and few proteins that feature a sequence identity greater than 60%.

This data was extracted from the multiple sequence alignment file produced by Clustal w.

Since the superfamily of α-β hydrolases are a quite diverse family (2), they do not feature an overall high sequence similarity (Figure 2). Some are very similar, but the majority do not fulfil the requirements for clustering.

Cross References

Most proteins in this data set are annotated in the Uniprot data base and feature a set of cross references to different data bases which annotate features in proteins based on their own algorithms and data. One protein can feature many annotations as well as more than one of one data base. After the search for different cross references, the result was diverse and features many different entries for the proteins of interest (Figure 3, Figure 4). Despite the heterogenous annotations, the entries with the highest appearance matched regarding the catalytic activity and are the following: Prosite PS00941 -Carboxylesterases type-B signature 2, PFAM PF00135 -Carboxylesterase type B, Gen Ontology GO:0052689 -carboxylic ester hydrolase activity and ExPasy Enzyme 3.1.1.-Hydrolases.

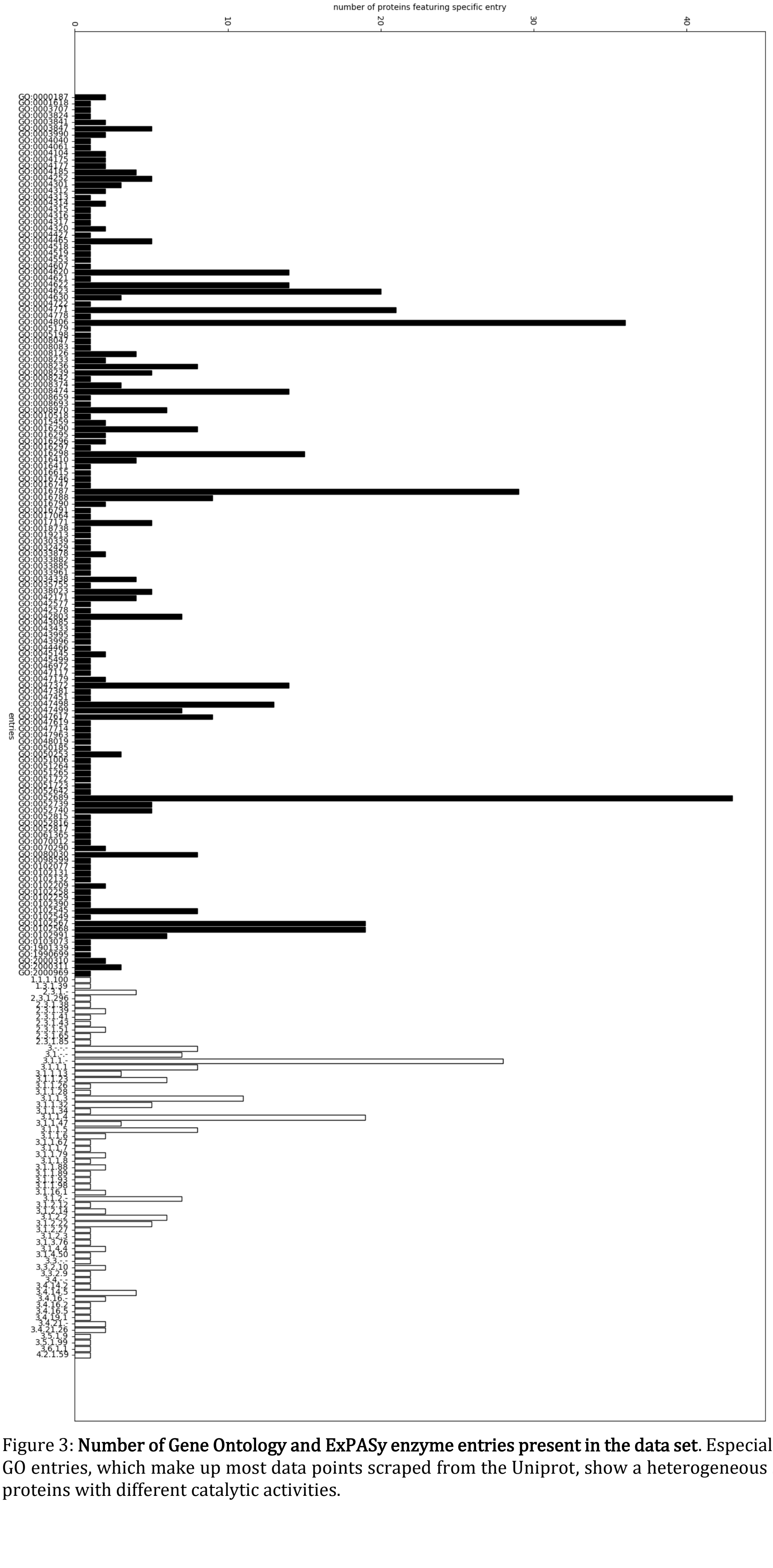

Figure 3: **Number of Gene Ontology and ExPASy enzyme entries present in the data set**. Especially the GO entries, which make up most data points scraped from the Uniprot, show a heterogeneous set of proteins with different catalytic activities.

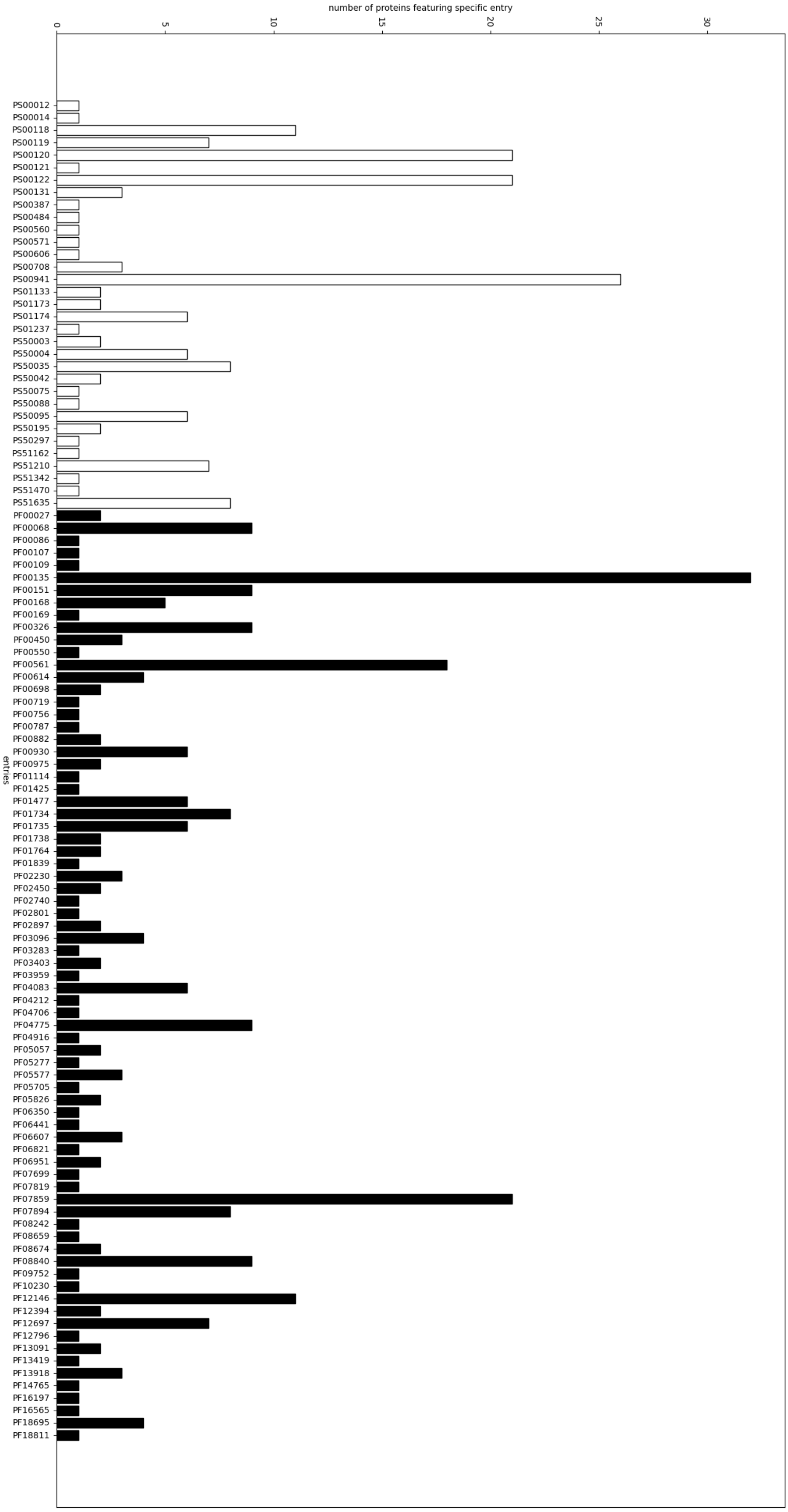

Figure 4:**Number of Prosite and PFAM entries present in the data set.** Here, the same principle applies as in Figure 3. Both PFAM and Prosite entries, of which not every protein features an annotation, show a diverse pattern among the proteins.

The cross references in the Uniprot to different motif, patterns and pathway data bases can be used to cluster the proteins based on the presence and absence of these. For this purpose, the cross references to Prosite, PFAM and GO were selected. Clustering is done by using categorical clustering (Figure 19).

Homology models

It was possible to build homology models for 185 of the 210 protein sequences. In this process a total of 113 different templates were used. No sufficient template could be found for the remaining protein sequences. The mean sequence similarity is 66.70% +/- 25.49% whereas the mean sequence identity is 53.75% +/- 27.73% (Figure 5) and the mean ratio of aligned to total residues is 75.31%.

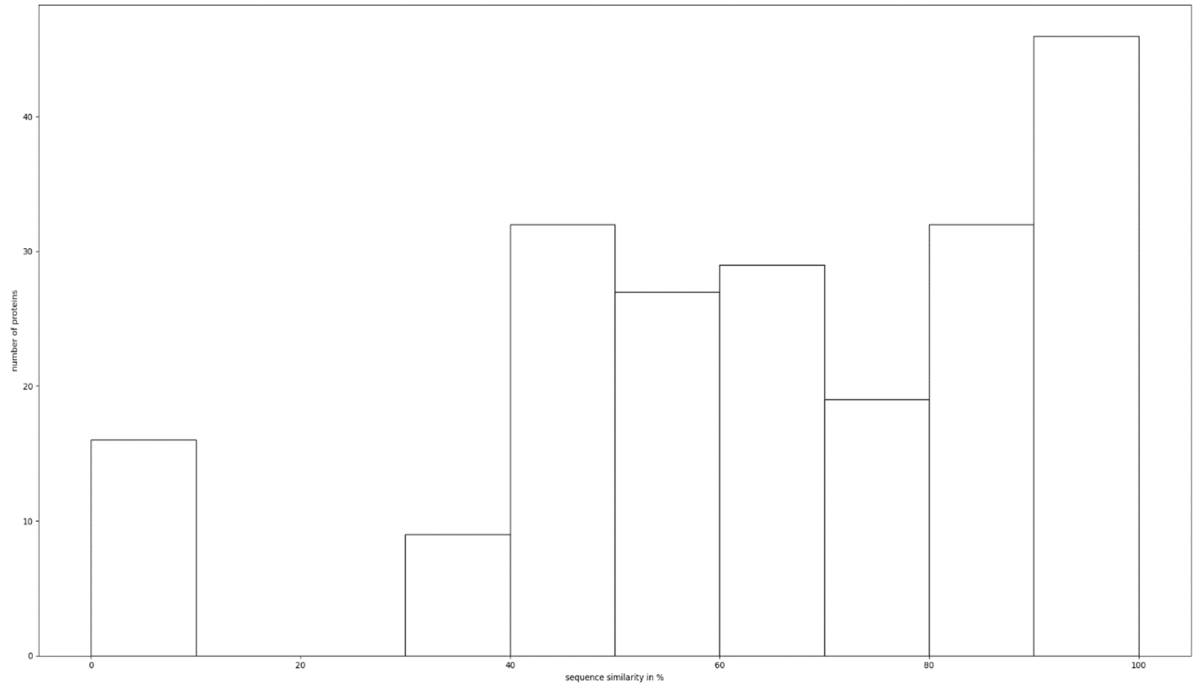

Figure 5: **Sequence similarity distribution** above all proteins with an overall mean sequence similarity of 66.705 +/- 27.73%. Most proteins have a sequence similarity that make them useable as models to predict the actual protein.

The quality of a homology model is not only dependent on sequence similarity but also on general availability and selection of templates, the program used to build the models, the result of the energy minimization during the building process and other factors. Due to varying model quality not all homology models can be used. The more loop regions are present in the models, the more challenging it is for the program to build a reliable model. In the following, examples of different outcomes are shown. All protein representations were created using Pymol.

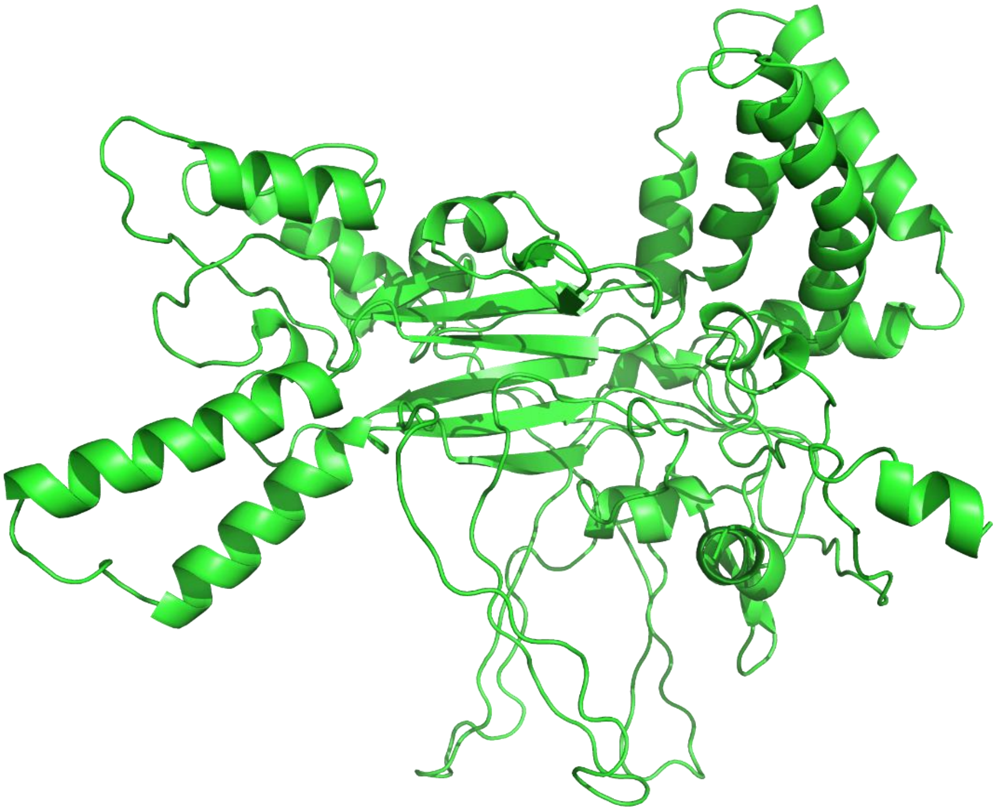

Figure 7: **Homology model of GPI inositol-deacylase (Q3UUQ7)** which is not well folded due to its low ratio of aligned and total residues

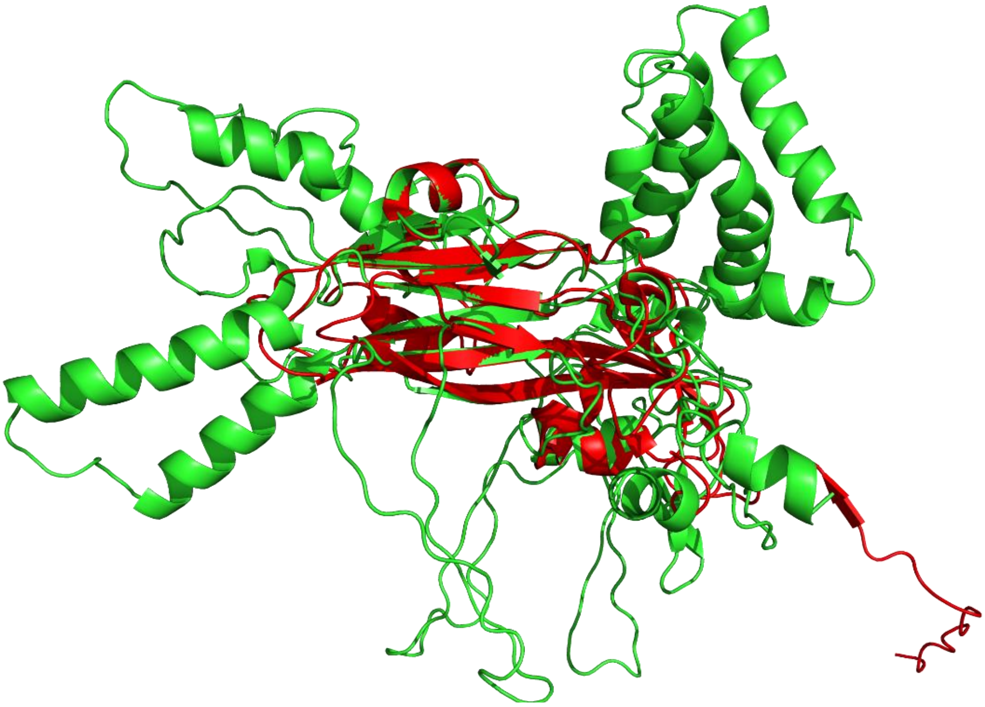

Figure 6: **Alignment of the homology model of GPI inositol-deacylase** in green (with Pymol super) with its template rice yellow mottle virus (1F2N-C) in red, which features a sequence similarity of 52.3%

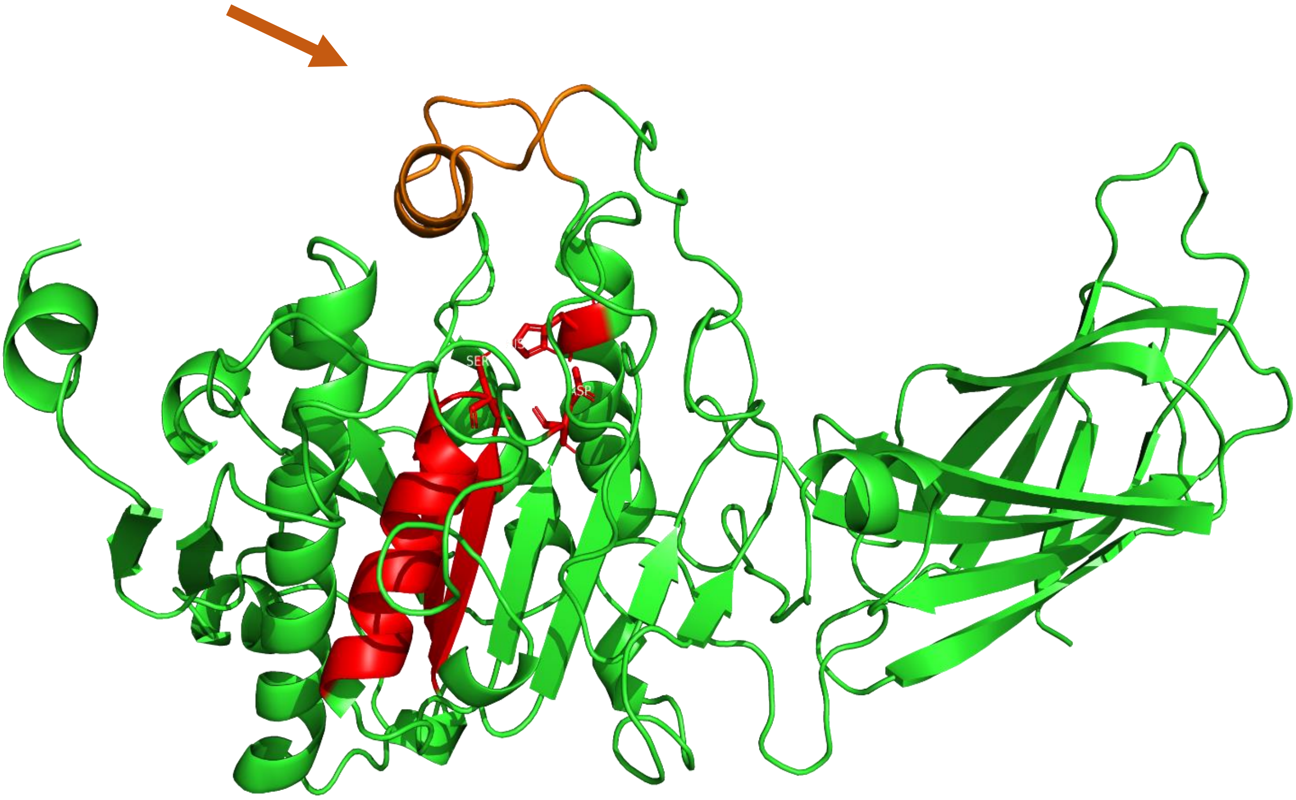

Figure 8: **Homology model of Pancreatic lipase-related protein 2 (P17892)**, featuring an α-β hydrolase fold, a serine on the nucleophilic elbow, and SER-HIS-ASP triad shown as red sticks and a lid coloured in orange (arrow). The area under the lid appears big and free but it is densely packed with side chain residues, which prevented cavity procreation.

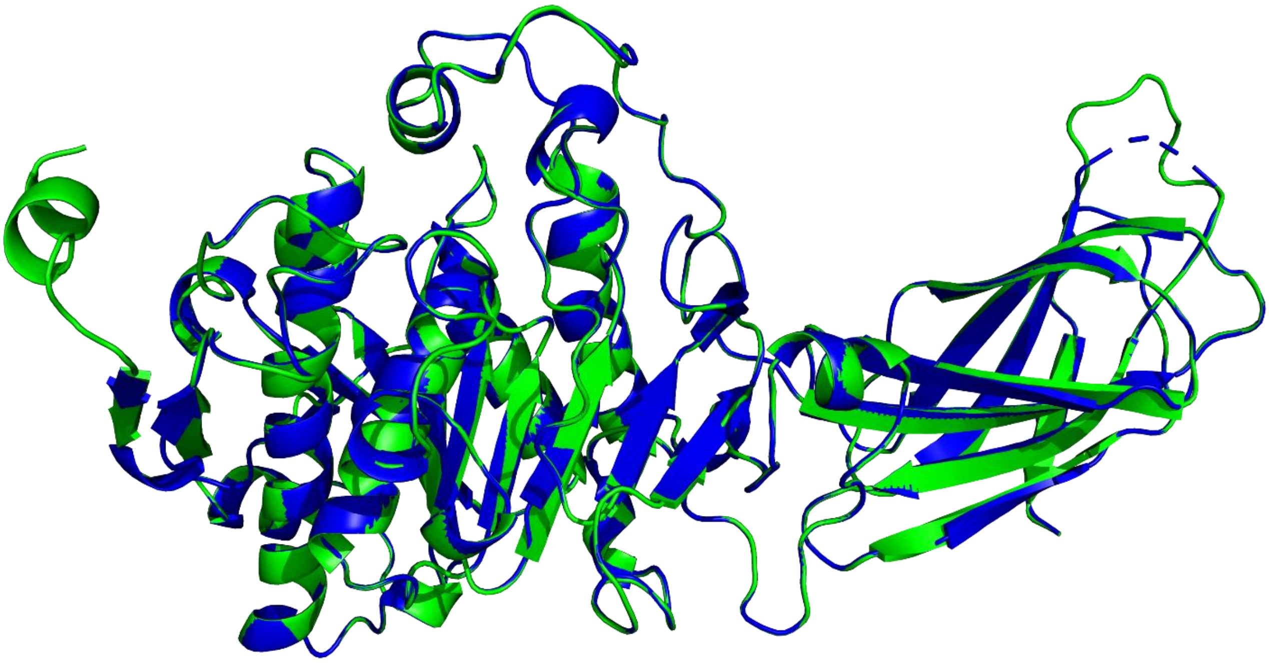

Figure 9: **Alignment of the homology model of Pancreatic lipase-related protein 2** in green with its template Rat pancreatic lipase related protein 2 (1BU8-A) in blue, which features a sequence similarity of 97.1%

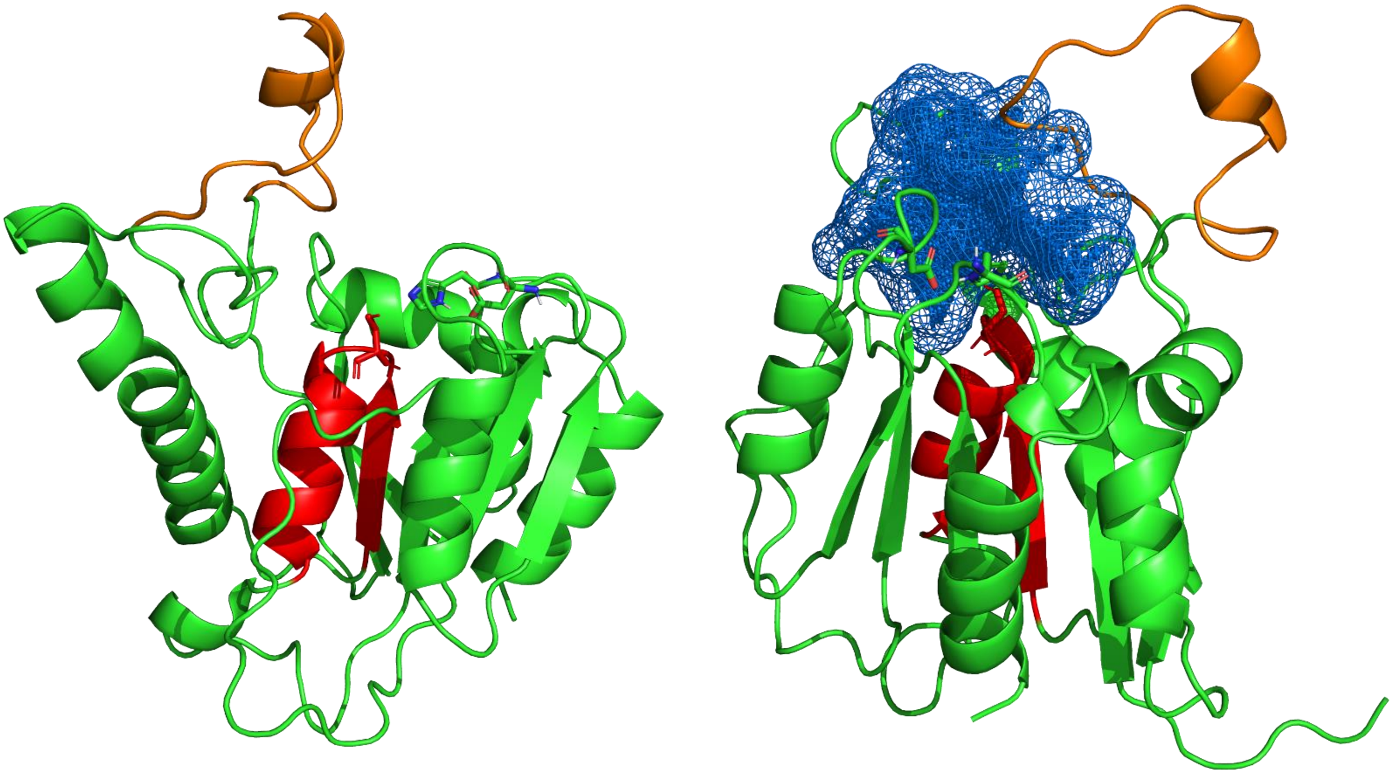

Figure 11: **Homology model of Esterase OVCA2 (Q9D7E3)** with an α- β hydrolase fold in red, a SER-HIS-ASP catalytic triad, with its serine placed on the loop of the nucleophilic elbow and an open lid in orange above the catalytic triad. On the right, the procreated cavity is illustrated in blue for better visibility under the lid and near the catalytic triad.

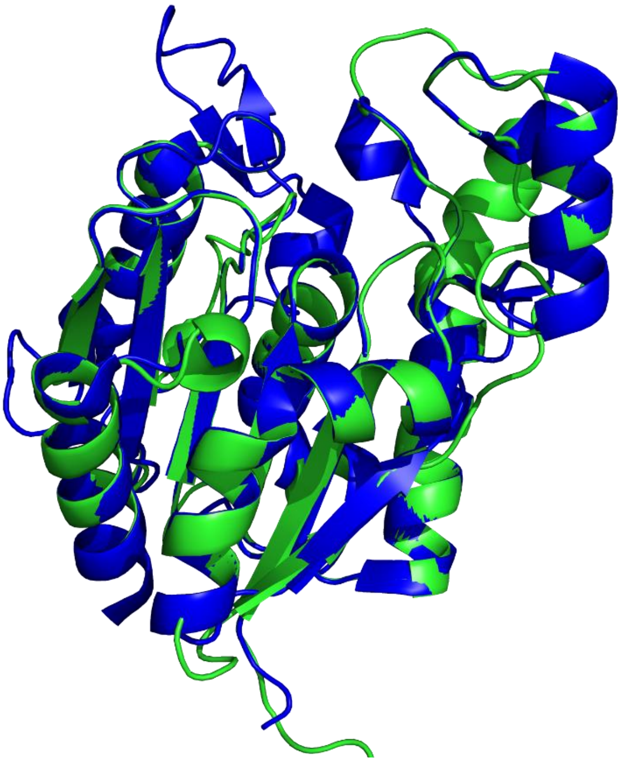

Figure 10: **Alignment of the homology model of Esterase OVCA2 (Q9D7E3)** in green with its template yeast FSH1/YHR049W, a member of the serine hydrolase family (1YCD-A), which features a sequence similarity of 52.9%

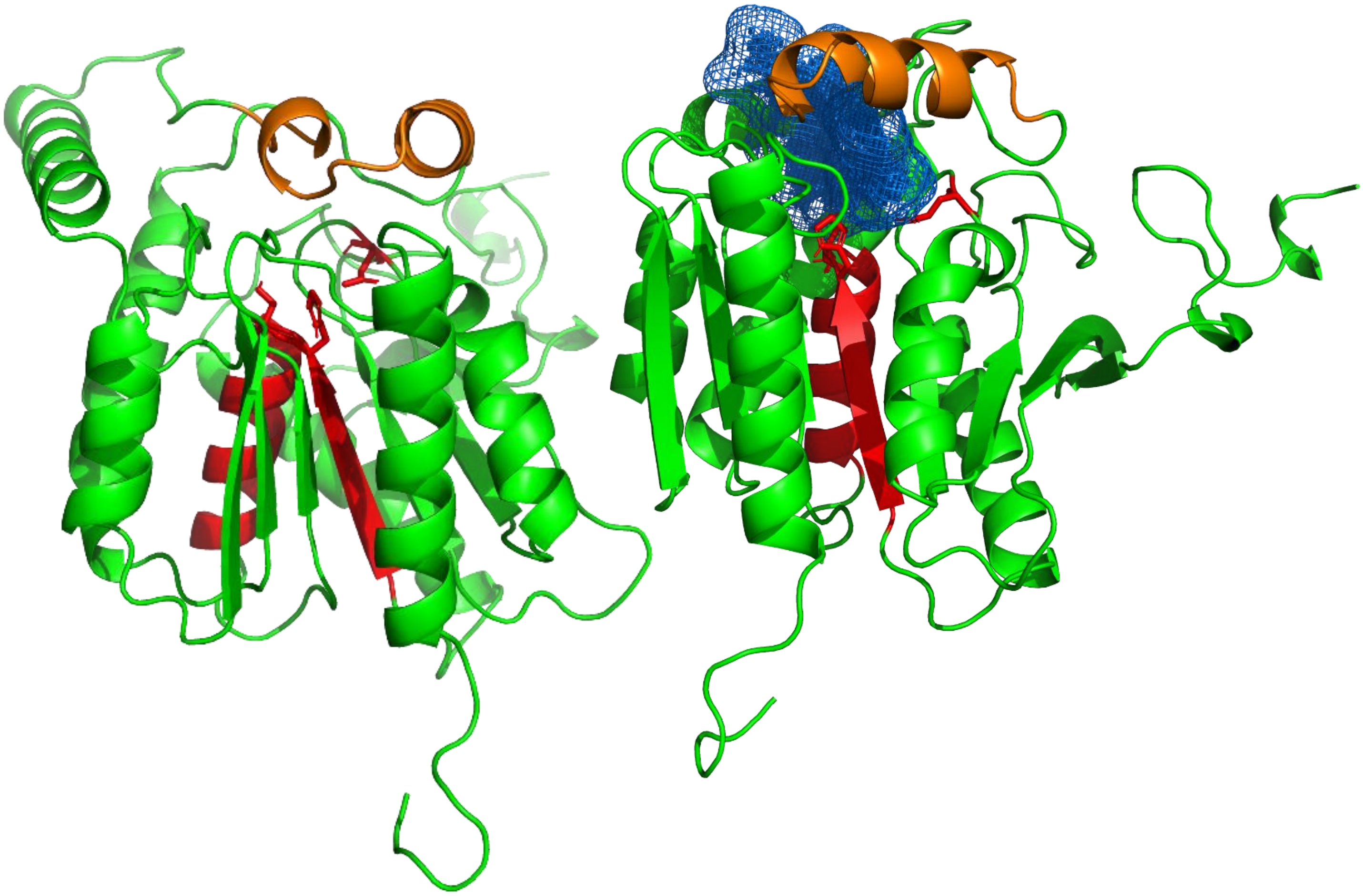

Figure 13: **Homology model of Monoglyceride lipase (O35678)** with an α-β hydrolase fold and a SER-HIS-GLU catalytic triad in red as well as a lid in orange. On the right side, the procreated cavity under the lid is coloured in blue for better visibility.

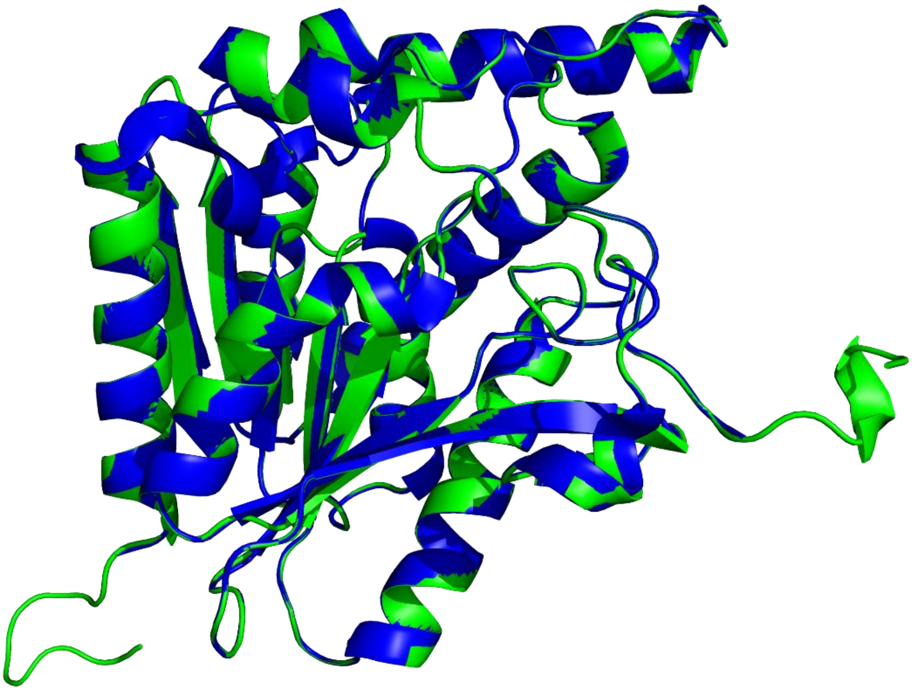

Figure 12: **Alignment of the homology model of Monoglyceride lipase** in green and its template a soluble form of human MGLL (3PE6-A), which features a sequence similarity of 94.1%

The homology model of GPI inositol-deacylase (Figure 7) features a sequence similarity of 52.3% with its template Rice yellow mottle virus (1F2N-C) and 151 aligned residues compared to 922 total residues(Figure 6). It sets an example for two possible obstacles. On the one hand, the low sequence similarity to its template and on the other hand the big discrepancy in aligned to total residues. These lead to the formation of a model that can not be used for further analysis.

The homology model of Pancreatic lipase-related protein 2 (P17892) (Figure 8) can be seen as an example for a model with high sequence similarity and therefore resulting in reliable folding throughout the modelling process. However, it also shows that the template choice plays an important factor. This model features an α-β hydrolase fold with a serine on the nucleophilic elbow and a catalytic triad where the fold is located (Figure 8). Nevertheless, it was not possible to procreate a cavity in this area because of the closed lid and the dense area in the active site due to the side chain rotation. Therefore, a highly representative model is not suitable for further analysis.

Although the homology model of Esterase OVCA2 (Q9D7E3) has a low sequence similarity of 52.9% to its template yeast FSH1/YHR049W, a member of the serine hydrolase family (1YCD-A) (Figure 10), the model turned out well folded. Compared to GPI inositol-deacylase, which features an analogue sequence similarity of 52.3% (Figure 6), the outcome of the modelling process is better. The reason therefore is the difference in the ratio of aligned to total residues, which is in this case 172 aligned to 225 total residues. This model can also be viewed as an example on how the template has a high impact on the modelling outcome. In the template, the lid (Figure 11) is wider open and thus provides the space needed for cavity procreation which in turn makes the model suitable for further analysis and cavity comparison with other models. Having a lid that is positioned such as in this case makes the procreated cavity and its properties more reliable than a version where the lid would be completely open. If this is the case, the cavity would lack its influence on its properties originating from the lid residues.

Looking at the homology model of Monoglyceride lipase (O35678) (Figure 13), one can see that the high sequence similarity to its template soluble form of human MGLL (3PE6-A) (Figure 12) leads to a reliable model, which is very likely to represent the protein in the real world. This model too, features the characteristic α- β hydrolase fold and a SER-HIS-GLU catalytic triad, where the serine is located at the nucleophilic elbow.

Cavity procreation

For all 185 homology models, cavities could be procreated which resulted in an overall number of 1328 cavities. The highest number of cavities found in a protein were 32 and the lowest number of cavities in one protein is one.

For the proteins, Carboxylic ester hydrolase (E9PV38), Pyrethroid hydrolase Ces2a (Q8QZR3), Carboxylesterase 3B (Q8VCU1), Carboxylic ester hydrolase (Q6PDB7), Epoxide hydrolase 3 (Q3V1F8), Protein ABHD8 (Q8R0P8), Retinoid-inducible serine carboxypeptidase (Q920A5), Testis-expressed protein 30 (Q3TUU5), Arylacetamide deacetylase-like 2 (B2RWD2), Protein ABHD13 (Q80UX8), Phospholipase A1 member A (Q8VI78) , Palmitoyl-protein thioesterase 1 (O88531), Arylacetamide deacetylase-like 3 (A2A7Z8), Diacylglycerol lipase-β (Q91WC9), Neuroligin-1 (Q99K10), Epoxide hydrolase 4 (Q6IE26), 1-acylglycerol-3-phosphate O-acyltransferase ABHD5 (Q9DBL9), which are also later used in the cavity matching, the procreated cavities enclose more space than only the active site and its substrate entries.

Motive Search

By applying the motif search algorithms brute force enhanced as well as robust motif search with an allowed deviation of +/- 2 A of the ideal distance of a motif, a high number of motifs could be found in the set of homology models (Figure 14). Before the interpretation, it must be kept in mind that motifs which feature less residues are more likely to be present in a protein, since fewer distance requirements need to be fulfilled (Table 12). Furthermore, the allowed distances play a role in the likelihood of finding a distinct motif too as well as the natural frequency of amino acids to occur in a protein. The higher the maximum allowed distances are, the more likely it is for a motif to be present in a protein. In general, the position in the protein like on the N-terminus for the N-terminal Threonine hydrolases is not considered in this survey. The most abundant motif is the “N-terminal threonine hydrolases” with 174 proteins featuring this motif, directly followed by “Glycosidases with retaining mechanism” which makes an appearance in 173 proteins. Both of the most abundant motifs feature two residues. A highly abundant motif is the three residues containing “Mannosidase”, which allows for a big distance between residues. Despite the “Mannosidase” motif, the motif with the highest number of appearances that consists out of more than two residues is the “triad based on homology models” which is a SER-HIS-ASP motif.

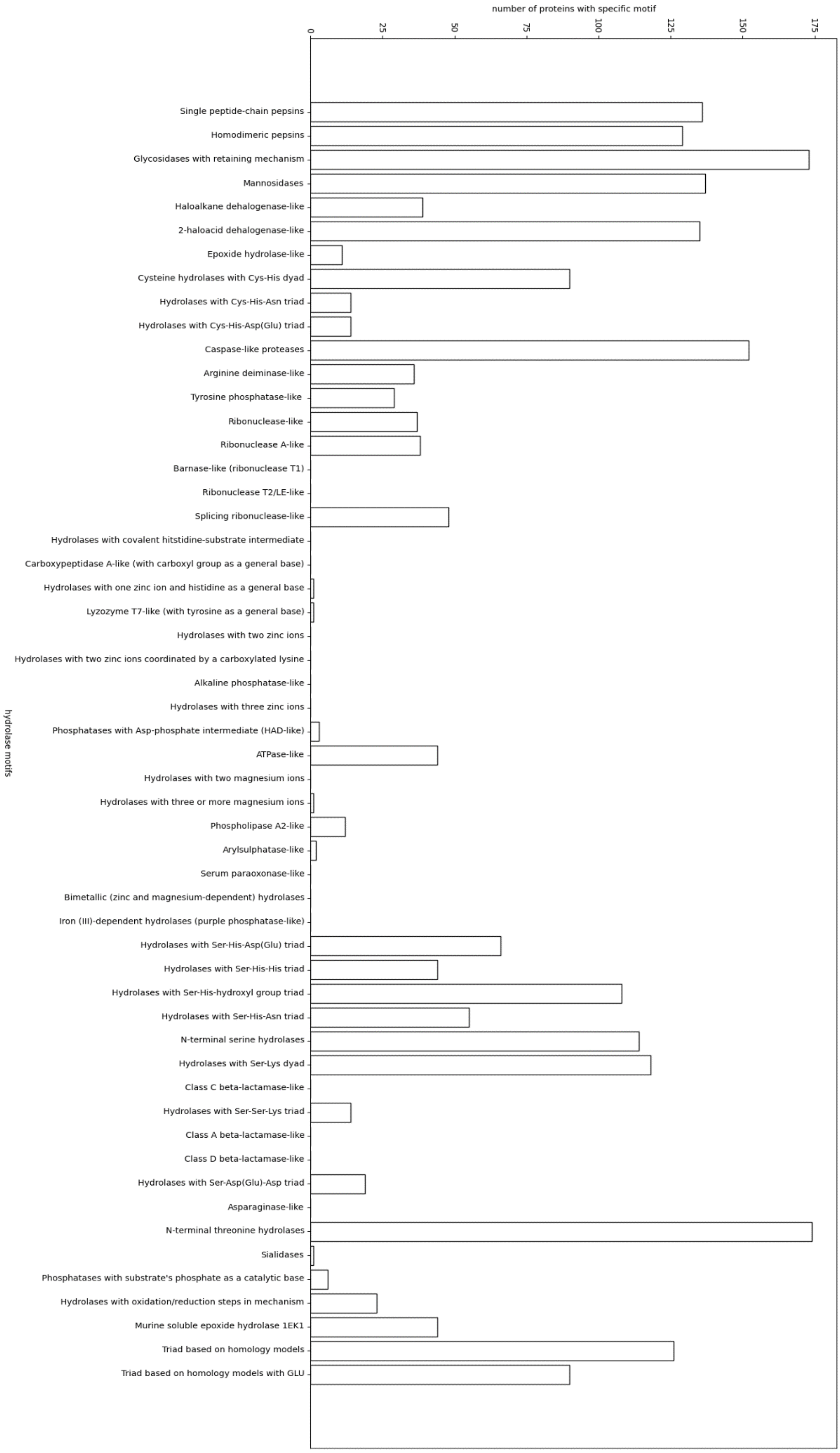

Figure 14: **Occurrence of a motif in the dataset.** One protein can feature more than one motif and therefore the number of motifs found exceeds the number of proteins. In general, bigger motifs are less likely to be found.

The number of proteins that feature a valid cavity is 119. That cavity is defined as a point cloud near an α-β hydrolase fold or for Phospholipase A a combination of annotations as Phospholipase A in the Uniprot database and the presence of the Phospholipase motif. For the rest 66 proteins, no cavity could be validated based on these requirements. Only motifs with more than two residues were used. The reason therefore is the frequent and multiple appearance of motifs with two residues in almost every protein which thus does not make them a good indicator for cavity validation. The most abundant motif is the "triad based on homology models", a SER-HIS-ASP motif, which is the most common one in combination with an α-β hydrolase fold. It is followed by the "hydrolase with Ser-His-hydroxyl group" which is a SER-HIS-SER motif and "triad based on homology model GLU" which is the same motif as the first motif, but the ASP substituted by a GLU (Figure 15) instead. The least abundant motif near a cavity is the "ATPase- like" motif followed by the "arginine deaminase-like" motif. Since one cavity can be validated by more than one motif, for instance if they share the cavity validating residue, the results show more motifs that validate cavities than there are proteins with valid cavities. 6 cavities were validated without a known motive but with a SER placed on the nucleophilic elbow of an α-β hydrolase fold which was indeed near a cavity.

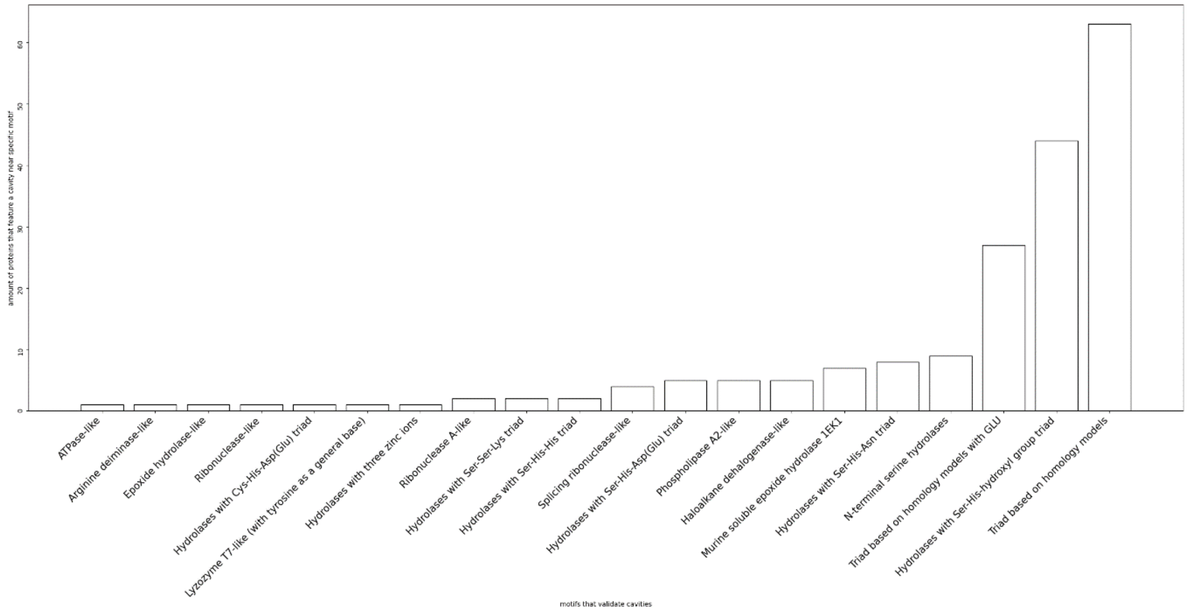

Figure 15: **Number of motifs found for validating a cavity**. The sum of motifs validating a cavity excel the number of proteins with a confirmed cavity. The reason therefore is that more than one motif can be in proximity to a cavity and therefore validate it. An additional constrain was that these cavities are located near an α-β hydrolase fold.

Table 1: Occurrences, the number of cavities motifs validate and the relative validation in percent of six motifs with the highest number of cavity validations

Motif	Occurrence	Cavity validation	% validation
Murine soluble epoxide hydrolase 1EK1	44	7	15.91
Hydrolase with SER-HIS-ASN	55	8	14.55
N-terminal serine hydrolases	114	9	7.89
Triad based on homology models GLU	90	27	30
Hydrolase with SER-HIS- hydroxyl group	108	44	40.74
Triad based on homology models	90	63	70

Comparing the occurrence of the 6 motifs with the highest cavity validation rate, “triad based on homology models” validates the most cavities relative to its appearance in all proteins, followed by “hydrolase with SER-HIS hydroxyl group”, whereas “N-terminal serine hydrolase” is more frequent, it is not as often found near a cavity that is near an α-β hydrolase fold (Table 1). In turn “Murine soluble epoxide hydrolase 1EK1”, which features the overall lowest occurrence in this selection of 6 motifs, has a higher relative validation than the motif with the highest occurrence.

Cavity matching

The process of cavity matching resulted in 16641 matches, of which 317 failed. Of these 9 results had to be excluded because matching the cavities against each other failed in both ways. The following proteins had to be excluded: Dipeptidyl peptidase 2 (Q9ET22), Carboxylic ester hydrolase (Q08ED5), Protein ABHD8 (Q8R0P8), Epoxide hydrolase 4 (Q6IE26), Diacylglycerol lipase-β (Q91WC9), Lysosomal Pro-X carboxypeptidase (Q7TMR0), Group 10 secretory phospholipase A2 (Q9QXX3), Acyl-coenzyme A thioesterase 1 (O55137). The cloud overlap represents the percentage of how much of a cavity overlaps with the cavity it got matched against. The mean overlap between all cavities is 34% with a standard deviation of 21% (Figure 16), the minimal overlap is 0.216% and the maximum overlap, apart from being 100% for identical cavities is

99.997%.The best overlap combination that is not 100% and 100% for identical cavities is 91.488% and 98.323%, whereas the worst, apart from being 0% and 0% when absolutely no overlap can be found, is 0.278% and 2.091%. Regarding Q8VEB4 and Q9QXE5 (Figure 17,Figure 18) the quality of cavity alignments is measured to be in the middle of the scale. Alignments where both cloud overlaps are very high are usually cavities of homology models that originate from the same template.

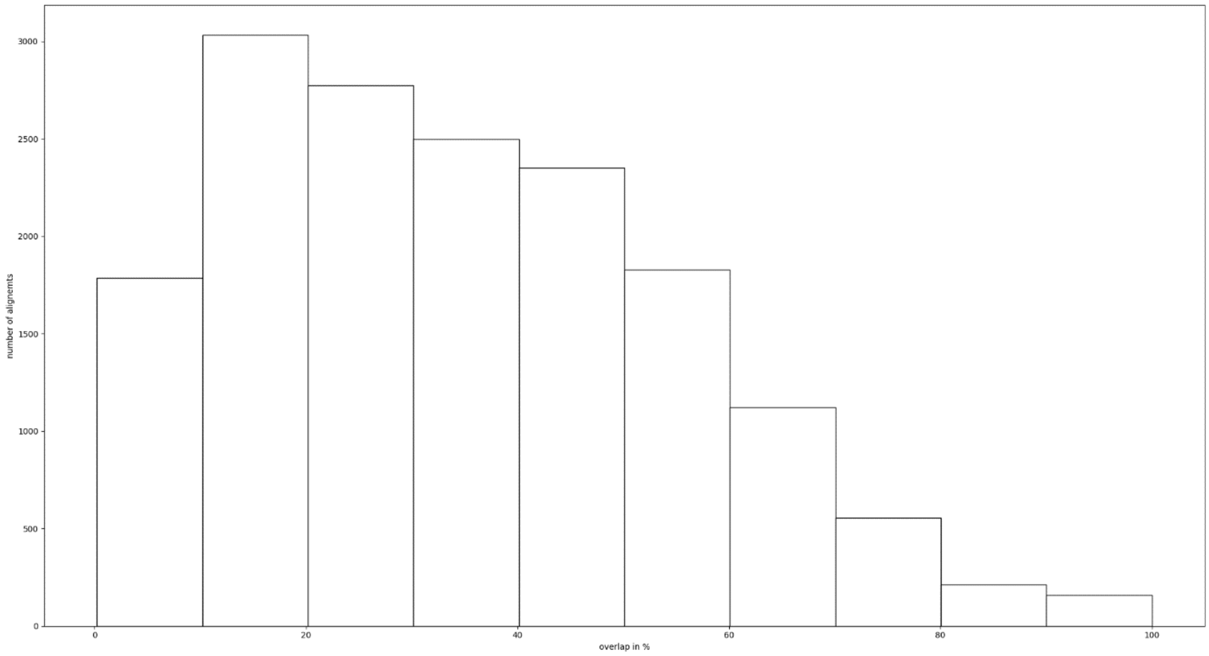

Figure 16: **Histogram of overlaps between cavities found in cavity matching**. The mean overlap during the process of matching is 34% with a standard deviation of 21%.

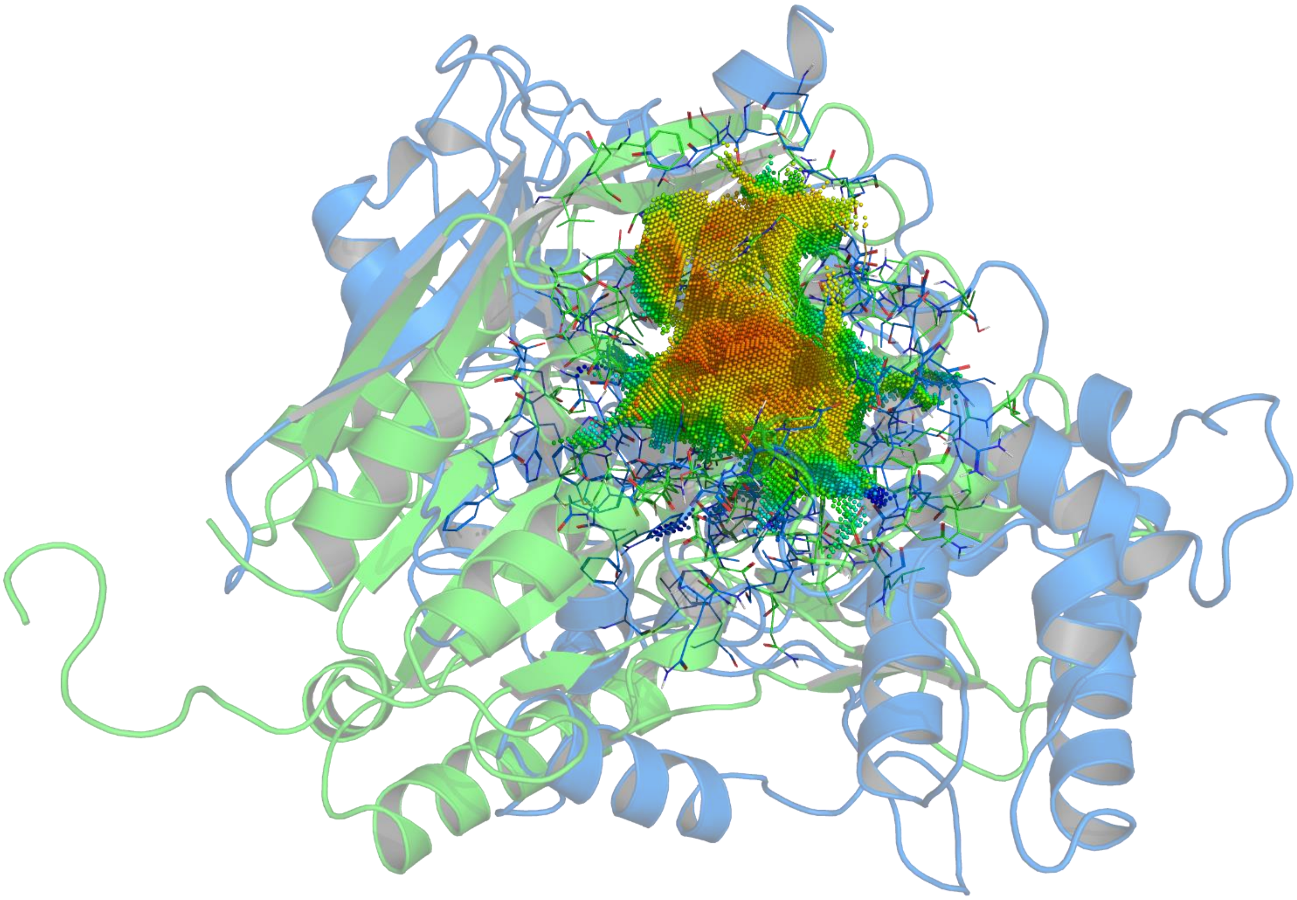

Figure 17: Alignment of Q8VEB4 in green and Q9QXE5 in blue based on their cavities.

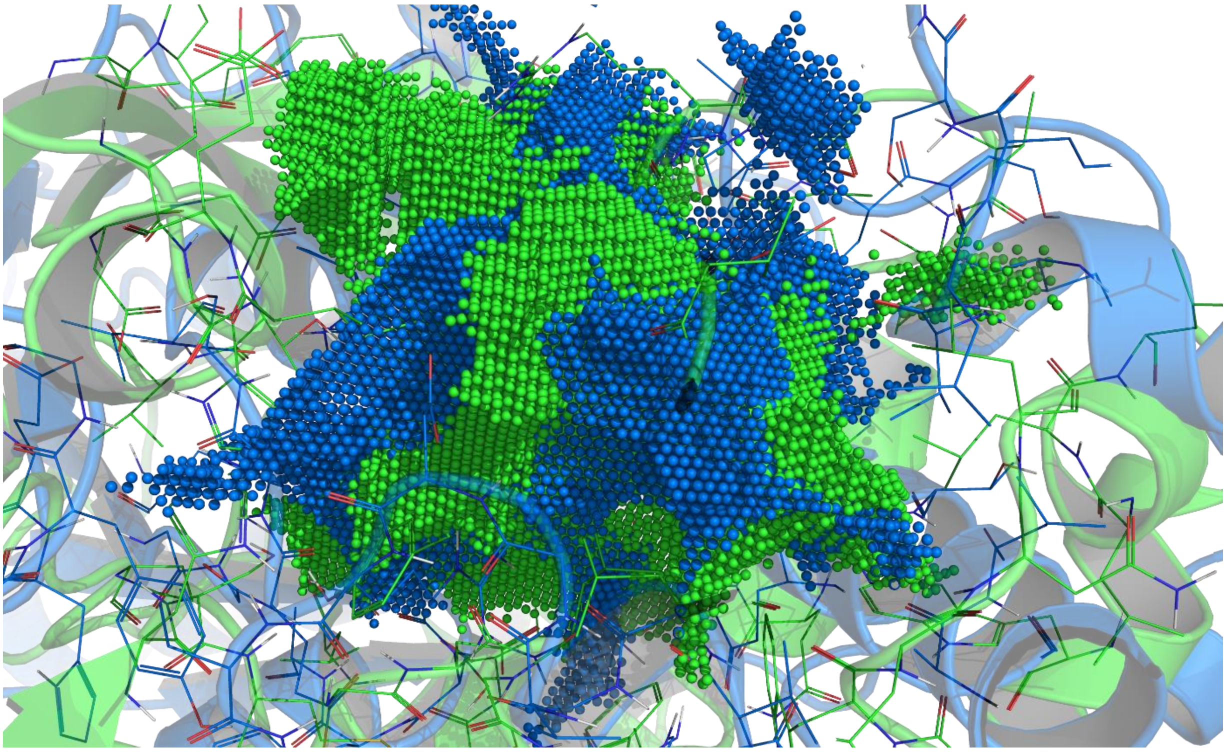

Figure 18: **Detailed view of the overlap of the cavities** of Q8VEB4 in green and Q9QXE5 in blue, which feature an overlap of 57% and respectively 61%

Result clustering

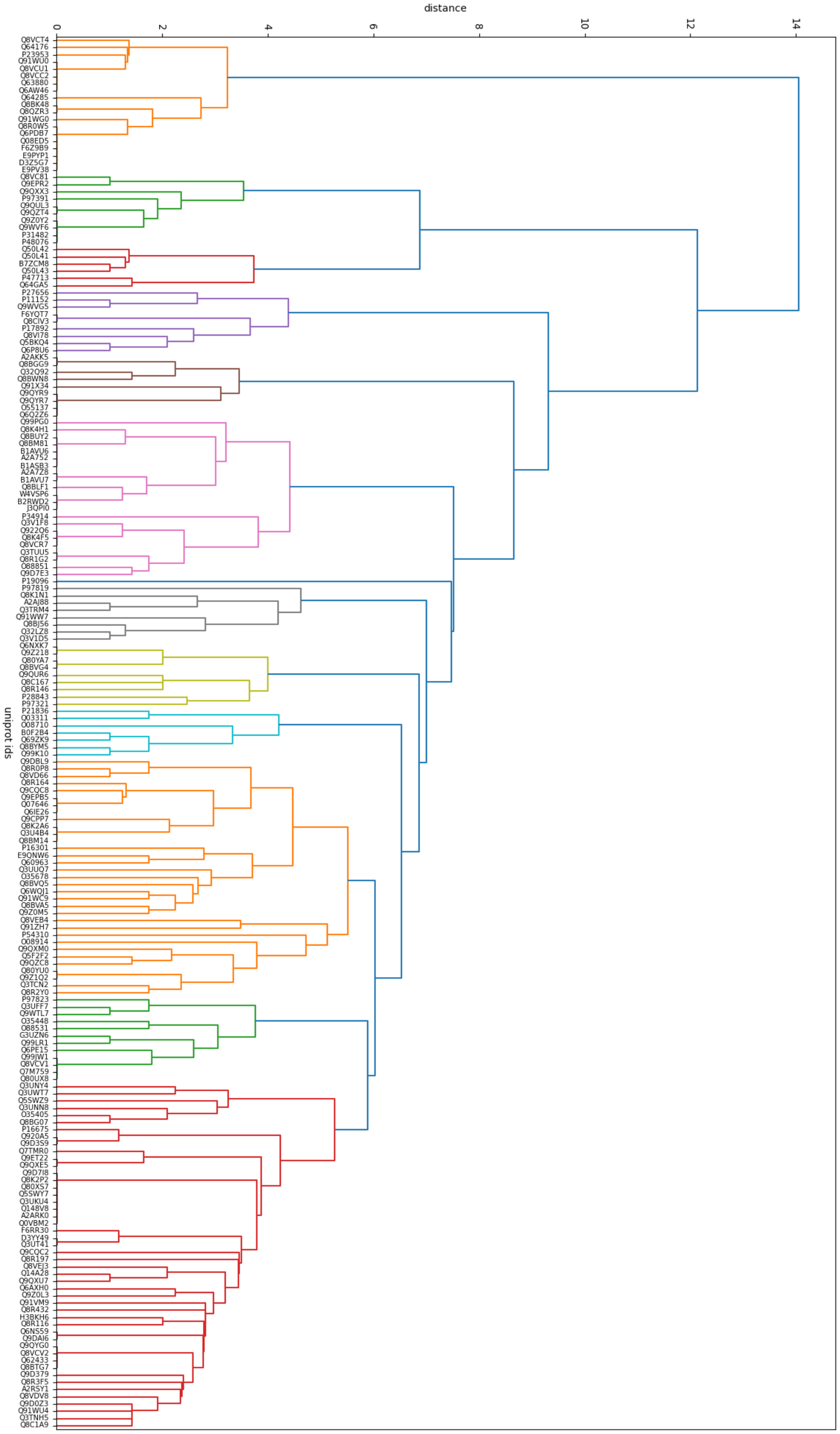

Figure 19: **Dendrogram based on Prosite, PFAM and GO entries of the data set**. The absence and presence of each pattern was used to cluster the proteins, which lead to 13 clusters, which are presented in different colours compared to their neighbouring cluster.

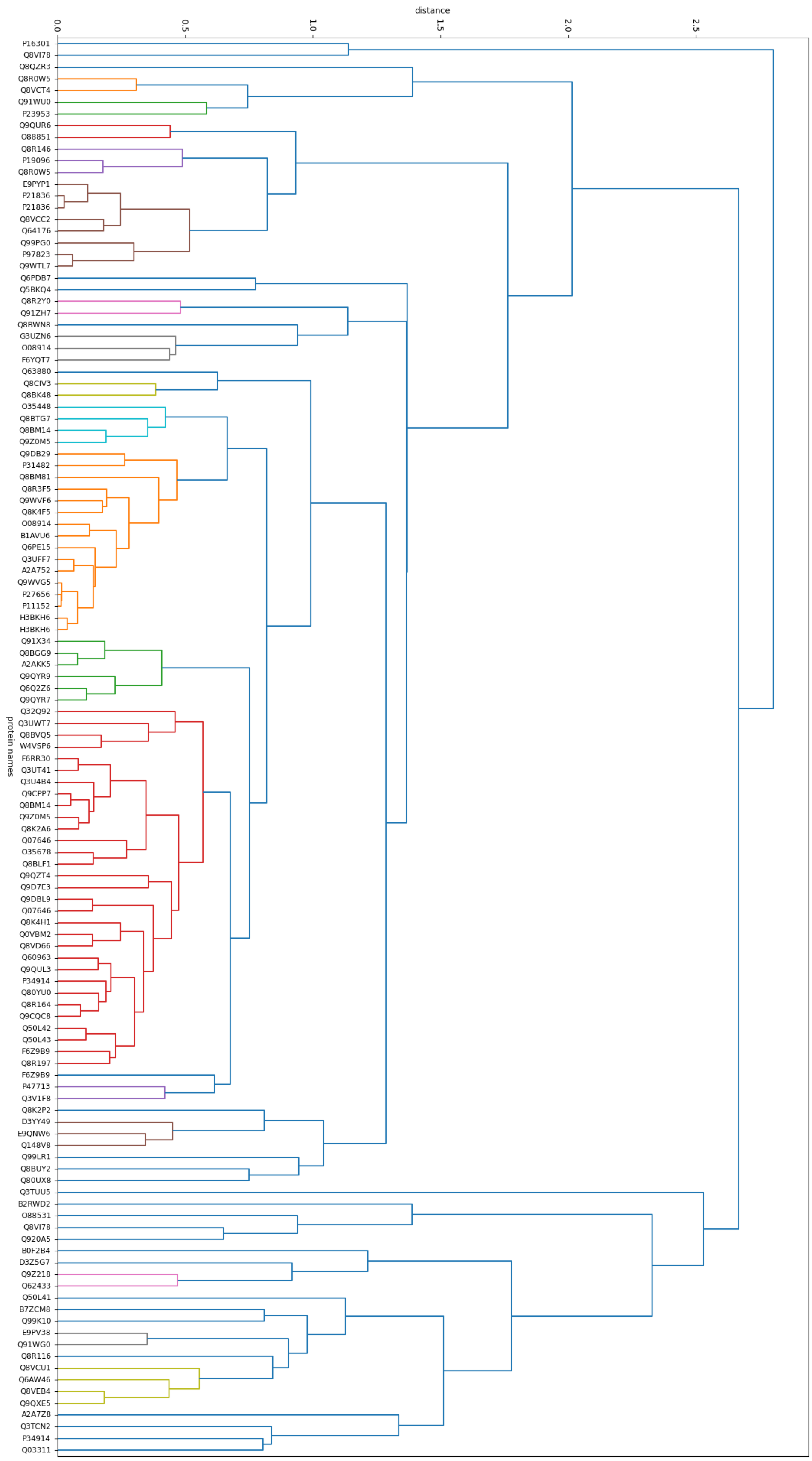

Figure 20: **Dendrogram based on similarities in the protein cavities** derived from the cavity matching

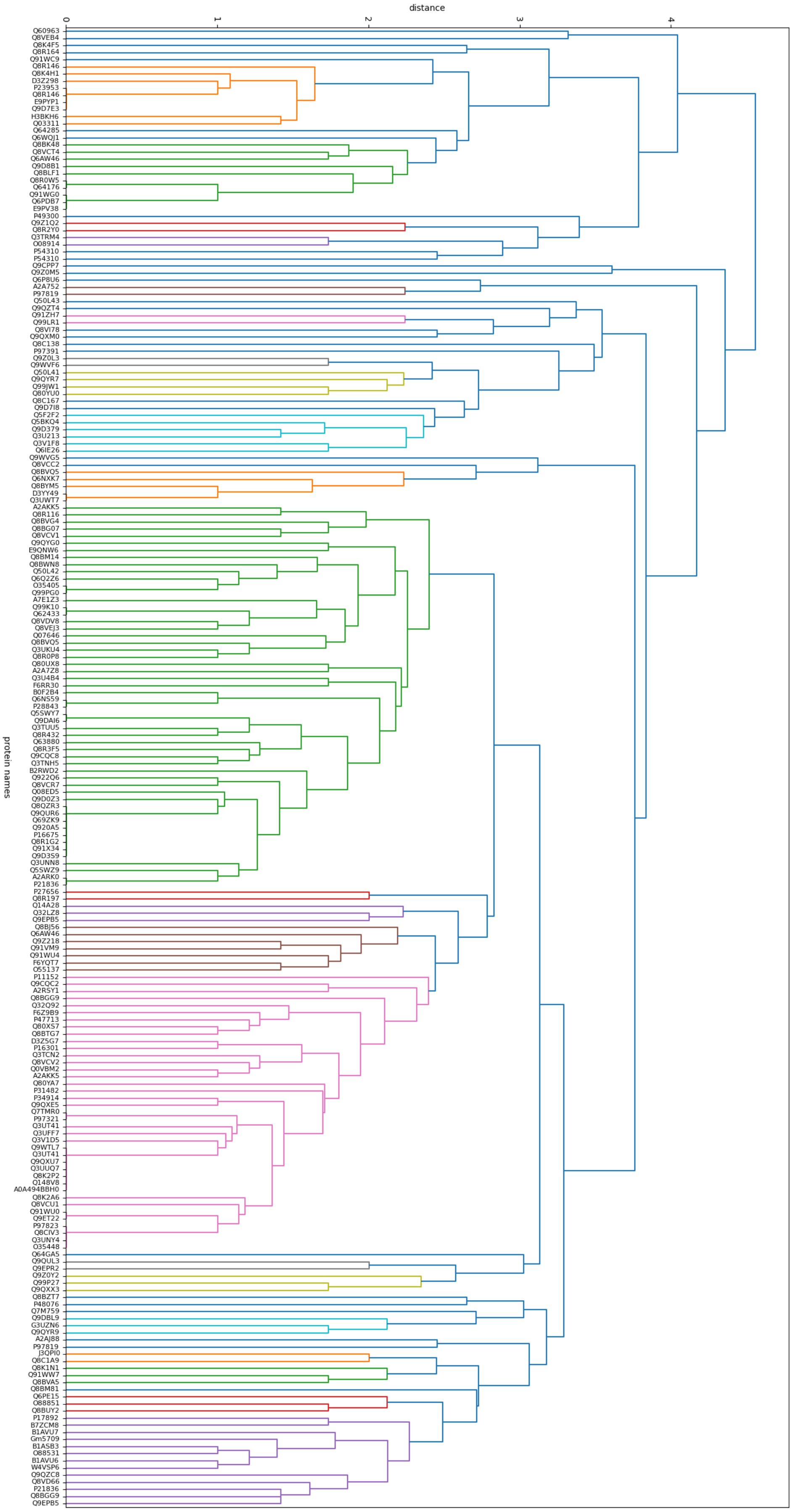

Figure 21: **Dendrogram based on the similarity in the proteins assay results**

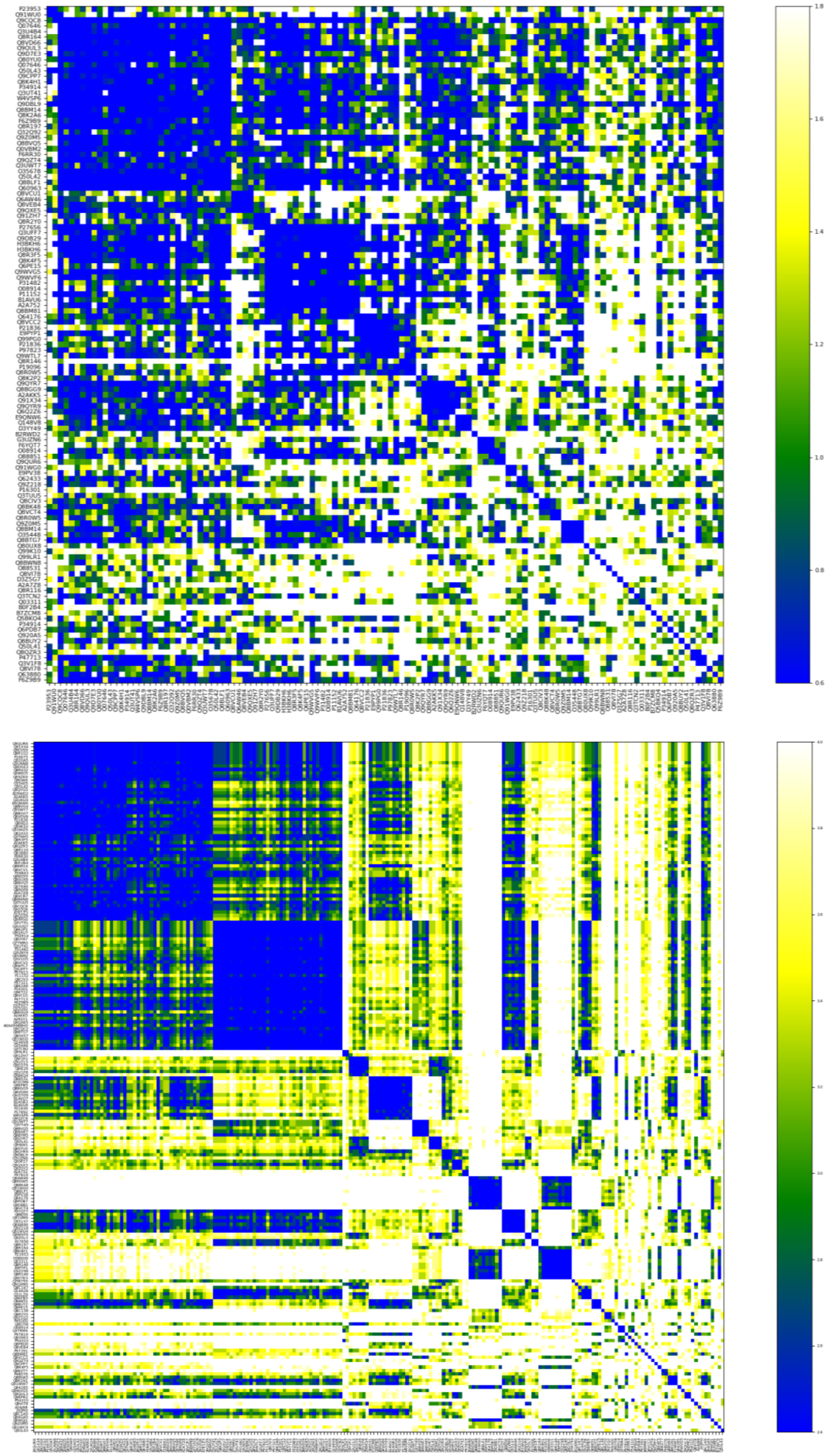

Figure 22: **Clustered heatmaps** with results from clustering based on cavity matching (top) and clustering based on assay results (bottom)

Table 2: Cluster derived from agglomerative clustering based on cavity matching results with the protein's Uniprot ID attached if they do not feature a unique name in this data set. The last row contains proteins can not be classified into one cluster.

Nr.	proteins in cluster	mean	max	min
0	Carboxylesterase 1F, Carboxylesterase 1C	0.584	0.585	582
1	Lipase (F6RR30), Group IIF secretory phospholipase A2, Cytosolic phospholipase A2 epsilon, Platelet-activating factor acetylhydrolase (Q60963), Protein FAM83B, Monoglyceride lipase, Lysosomal acid lipase/cholesteryl ester hydrolase, Mesoderm-specific transcript protein, Valacyclovir hydrolase, (Lyso)-N-acylphosphatidylethanolamine lipase, Group IIE secretory phospholipase A2, Esterase OVCA2, Gastric triacylglycerol lipase, Kynurenine formamidase, Bifunctional epoxide hydrolase 2, Lipase (Q3UT41), Lipase member M, Carboxylic ester hydrolase (F6Z9B9), S-acyl fatty acid synthase thioesterase, medium chain, Protein phosphatase methylesterase 1, Lipase member K, 1-acylglycerol-3-phosphate O-acyltransferase ABHD5, AADACL2 family member 2, Maspardin, Acyl-coenzyme A thioesterase 6, Phospholipase D, Lipase member N, Cytosolic phospholipase A2 delta, Mesoderm-specific transcript protein, Protein ABHD16B, Neutral cholesterol ester hydrolase 1	0.431	2.362	0.05
2	Carboxylesterase 5A, Phospholipase A2 group XV, Thymus-specific serine protease, Carboxylesterase 3B	0.451	0.607	0.181
3	Monoacylglycerol lipase ABHD6, Phospholipase ABHD3	0.48	0.48	0.479
4	Isoamyl acetate-hydrolyzing esterase 1 homolog, Lysophospholipase-like protein 1, Hepatic triacylglycerol lipase, Arylacetamide deacetylase-like 4, AADACL4 family member 2, Fatty-acid amide hydrolase 1, Palmitoyl-protein thioesterase ABHD10- mitochondrial, Group IID secretory phospholipase A2, AADACL4 family member 4, Phospholipase A2- membrane associated, Protein ABHD11, S-formylglutathione hydrolase, S-formylglutathione hydrolase, Malonyl-CoA-acyl carrier protein transacylase- mitochondrial, Lipoprotein lipase, Endothelial lipase	0.284	1.707	0.011
5	Acyl-protein thioesterase 1, Carboxylic ester hydrolase (E9PYP1), Liver carboxylesterase 1, Acyl-protein thioesterase 2, Acetylcholinesterase, Carboxylesterase 1E, Arylacetamide deacetylase, Acetylcholinesterase	0.368	0.909	0.024
6	Fatty acid synthase, Carboxylesterase 4A, Acylamino-acid-releasing enzyme	0.382	0.731	0.176
8	Acyl-coenzyme A thioesterase 5, Bile acid-CoA:amino acid N-acyltransferase, Acyl-coenzyme A thioesterase 2- mitochondrial, Acyl-coenzyme A thioesterase 3, Acyl-coenzyme A amino acid N-acyltransferase 2, Acyl-coenzyme A amino acid N-acyltransferase 1	0.31	0.689	0.077
9	Platelet-activating factor acetylhydrolase (E9QNW6), Protein FAM83H, Lipase (D3YY49)	0.414	0.563	0.339
11	Abhydrolase domain-containing 12B, Fatty-acid amide hydrolase 1, Lipase, member I	0.453	0.462	0.437
12	Prolyl endopeptidase, Putative hydrolase RBBP9	0.441	0.442	0.44

13	Carboxylic ester hydrolase, Acylcarnitine hydrolase	0.347	0.348	0.347
14	Dipeptidyl aminopeptidase-like protein 6, Protein NDRG1	0.469	0.469	0.469
17	Lipase member H, Pyrethroid hydrolase Ces2e	0.381	0.382	0.38
18	Carboxylesterase 4A, Carboxylesterase 1D	0.307	0.307	0.307
19	Lipase member K, Lysosomal thioesterase PPT2, Lysosomal acid lipase/cholesteryl ester hydrolase, Protein NDRG4	0.36	0.544	0.188
40	Cytosolic phospholipase A2, Epoxide hydrolase 3	0.419	0.42	0.419
nan	Protein FAM83A, Arylacetamide deacetylase-like 2, Phosphatidylcholine-sterol acyltransferase, Testis-expressed protein 30, Protein ABHD13, Neuroligin-1, Lysophosphatidylserine lipase ABHD12, Peroxisomal succinyl-coenzyme A thioesterase, Palmitoyl-protein thioesterase 1, Phospholipase A1 member A, Carboxylic ester hydrolase (D3Z5G7), Arylacetamide deacetylase-like 3, Palmitoleoyl-protein carboxylesterase NOTUM, Putative phospholipase B-like 2, Cholinesterase, Neuroligin 4-like, Phospholipase A2 (B7ZCM8), Inactive pancreatic lipase-related protein 1, Bifunctional epoxide hydrolase 2, Carboxylic ester hydrolase (Q6PDB7), Retinoid-inducible serine carboxypeptidase, AADACL2 family member 1, Cytosolic phospholipase A2 zeta, Pyrethroid hydrolase Ces2a, Phospholipase A1 member A, Carboxylesterase 3A, Carboxylic ester hydrolase (F6Z9B9)			

Table 3: Cluster based on agglomerative clustering of the assay results with the protein's Uniprot ID attached if they do not feature a unique name in the mouse. The last row contains proteins that can not be classified into one cluster

Nr.	proteins in cluster	act	inact	mean act	mean dist
0	Protein phosphatase methylesterase 1, 9430007A20Rik protein, Maspardin, Testis-expressed protein 30, Protein ABHD14A, Protein ABHD14B, Protein ABHD13, Acyl-coenzyme A amino acid N-acyltransferase 1, Bile acid-CoA:amino acid N-acyltransferase, Probable serine carboxypeptidase CPVL, Carboxymethylenebutenolidase homolog, Lysosomal protective protein, Retinoid-inducible serine carboxypeptidase, Transmembrane protein 53, Carboxylic ester hydrolase (Q08ED5), Neuroligin-2, Prolyl endopeptidase, 5'-3' exonuclease PLD3, Cytosolic phospholipase A2 epsilon, Protein NDRG2, Arylacetamide deacetylase-like 2, Protein FAM83C, Protein FAM83F, Platelet-activating factor acetylhydrolase (E9QNW6), Dipeptidyl peptidase 9, 5'-3' exonuclease PLD4, MIT domain-containing protein 1, Dickkopf-related protein 4, Neuroligin-1, Protein NDRG1, Mesoderm-specific transcript protein, Mitochondrial cardiolipin hydrolase, Acetylcholinesterase, Protein FAM83G, Protein FAM135B, Glycosyl-phosphatidylinositol-specific phospholipase D, Cotranscriptional regulator FAM172A, Malonyl-CoA-acyl carrier protein transacylase, mitochondrial, Protein FAM135A, Dipeptidyl peptidase 4, A/β hydrolase domain-containing protein 17C, Neuroligin 4-like, Carboxylesterase 3A, Palmitoleoyl-protein carboxylesterase NOTUM, Inactive phospholipase D5, Arylacetamide deacetylase-like 3, Arylacetamide deacetylase, Acyl-coenzyme A thioesterase 5, Protein ABHD8, Peroxisomal succinyl-coenzyme A thioesterase,	537	1743	0.236	2.043

	Pyrethroid hydrolase Ces2a, Lipase member K, Lipase member N, Lipase (F6RR30)				
1	Protein FAM83H, Phospholipase (Q3UNY4), Phospholipase A2-membrane associated, Lipase (Q3UT41), Protein FAM83A, GPI inositol-deacylase, Prokineticin-2, Bifunctional epoxide hydrolase 2, Lysosomal thioesterase PPT2, Protein FAM83B, Lipase (Q3UT41), Dipeptidyl peptidase 8, Lysosomal Pro-X carboxypeptidase, Cytosolic phospholipase A2, N-myc downstream-regulated gene 3 protein, Protein FAM83E, Protein NDRG4, Colipase, Thymus-specific serine protease, Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A amino acid N-acyltransferase 2, Acyl-coenzyme A thioesterase 6, Carboxylic ester hydrolase (D3Z5G7), Carboxylesterase 1F, Carboxylic ester hydrolase (F6Z9B9), Carboxylesterase 3B, Dipeptidyl peptidase 2, Phosphatidylcholine-sterol acyltransferase, Lipase member M, Prolyl endopeptidase FAP, Lipase member H, Lipoprotein lipase, Acyl-protein thioesterase 1, Lysophospholipase-like protein 1, Acyl-protein thioesterase 2, Putative phospholipase B-like 2, Omega-hydroxyceramide transacylase, A0A494BBH0, KAT8 regulatory NSL complex subunit 3	137	1503	0.084	1.692
2	Phospholipase ABHD3, Lysophosphatidylserine lipase ABHD12	49	31	0.612	2.236
3	Inactive pancreatic lipase-related protein 1, Epoxide hydrolase 3, Epoxide hydrolase 4, Epoxide hydrolase 1, Protein SERAC1, Protein ABHD15	72	168	0.3	2.123
4	Pancreatic lipase-related protein 2, AADACL4 family member 4, AADACL4 family member 5, AADACL2 family member 2, (Lyso)-N-acylphosphatidylethanolamine lipase, Acyl-coenzyme A amino acid N-acyltransferase 2, Serine hydrolase-like protein, Acetylcholinesterase, Protein ABHD1, Palmitoyl-protein thioesterase 1, Phospholipase A2 (B7ZCM8), Gm5709, AADACL4 family member 3	209	431	0.327	1.961
5	Protein phosphatase methylesterase 1, Neuroligin-3, Lipase (D3YY49), Phospholipase D, Inactive dipeptidyl peptidase 10	53	187	0.221	1.58
6	A/β hydrolase domain-containing protein 17A, Protein ABHD16B, Acyl-coenzyme A thioesterase 3, Cytosolic phospholipase A2 zeta	51	109	0.319	2.109
7	1-acylglycerol-3-phosphate O-acyltransferase ABHD5, Abhydrolase domain-containing 12B, Acyl-coenzyme A thioesterase 2- mitochondrial	37	83	0.308	1.989
8	Group 10 secretory phospholipase A2, Phospholipase A2, Group XIIB secretory phospholipase A2-like protein	24	96	0.2	2.139
9	AADACL4 family member 2, 85/88 kDa calcium-independent phospholipase A2	48	72	0.4	2.236
10	Carboxylesterase 4A, Carboxylic ester hydrolase (Q6PDB7), Carboxylesterase 1D, Carboxylic ester hydrolase (E9PV38), Acylcarnitine hydrolase, Carboxylesterase 1E, Neutral cholesterol ester hydrolase 1, Carboxylesterase 5A, Pyrethroid hydrolase Ces2e, Androgen-induced gene 1 protein	258	182	0.586	1.804

11	Patatin-like phospholipase domain-containing protein 2, Lipase, member I, Acyl-coenzyme A thioesterase 1, Dipeptidyl aminopeptidase-like protein 6, Inorganic pyrophosphatase 2, mitochondrial, Carboxylesterase 5A, Transmembrane and coiled-coil domain-containing protein 4	60	260	0.188	1.907
12	Otoconin-90, Group IID secretory phospholipase A2	29	51	0.362	1.732
13	Hepatic triacylglycerol lipase, S-acyl fatty acid synthase thioesterase- medium chain	14	66	0.175	2.0
15	Carboxylic ester hydrolase (E9PYP1), S-formylglutathione hydrolase, Esterase OVCA2, Acylamino-acid-releasing enzyme, Kynurenine formamidase, Cholinesterase, Carboxylesterase 1C, Acylamino-acid-releasing enzyme, Carboxylic ester hydrolase(D3Z298)	226	214	0.514	1.172
19	Serine hydrolase-like protein, Prokineticin-1, Patatin-like phospholipase domain-containing protein 5	32	128	0.2	2.15
20	Putative hydrolase RBBP9, AADACL2 family member 1, Palmitoyl-protein thioesterase ABHD10- mitochondrial	43	77	0.358	1.989
22	Phosphatidylserine lipase ABHD16A, Monoacylglycerol lipase ABHD6	43	37	0.538	2.236
25	Patatin-like phospholipase domain-containing protein 6, Fatty-acid amide hydrolase 1	63	17	0.788	1.732
39	1-acylglycerol-3-phosphate O-acyltransferase Pnpla3, Calcium-independent phospholipase A2-gamma, Lipid droplet-associated hydrolase	25	95	0.208	1.989
42	Group IIE secretory phospholipase A2, Group XIIA secretory phospholipase A2	12	68	0.15	2.0
46	AADACL2 family member 3, Protein ABHD18	18	62	0.225	2.0
nan	Valacyclovir hydrolase, A/β hydrolase domain-containing protein 17B, Monoacylglycerol lipase ABHD2, Prolyl endopeptidase-like, Androgen-dependent TFPI-regulating protein, C-type lectin domain family 10 member A, Protein FAM83D, 85/88 kDa calcium-independent phospholipase A2, Platelet-activating factor acetylhydrolase, Hormone-sensitive lipase, Pancreatic triacylglycerol lipase, Phospholipase A2 group XV, Phospholipase A2 group V, Arylacetamide deacetylase-like 4, Liver carboxylesterase 1, Group IIF secretory phospholipase A2, Gastric triacylglycerol lipase, Protein ABHD11, Group 3 secretory phospholipase A2, Group IIC secretory phospholipase A2, Bile salt-activated lipase, Endothelial lipase, Hormone-sensitive lipase, Phospholipase A1 member A, Patatin-like phospholipase domain-containing protein 7, Cytosolic phospholipase A2 gamma, Diacylglycerol lipase-α, Lysosomal acid lipase/cholesteryl ester hydrolase, Diacylglycerol lipase-β, Cytosolic phospholipase A2 delta				

Table 4: Cluster based on Uniprot cross references with the protein's Uniprot ID attached if they do not feature a unique name in the mouse

Nr.	Proteins in cluster
0	Carboxylic ester hydrolase (D3Z5G7), Carboxylic ester hydrolase (E9PV38), Carboxylic ester hydrolase (E9PYP1), Carboxylic ester hydrolase (F6Z9B9), Carboxylesterase 1C, Carboxylic ester hydrolase (Q08ED5), Carboxylesterase 3A, Carboxylesterase 1E, Bile salt-activated lipase, Carboxylesterase 5A, Carboxylic ester hydrolase (Q6PDB7), Pyrethroid hydrolase Ces2e, Pyrethroid hydrolase Ces2a, Carboxylesterase 4A, Liver carboxylesterase 1, Carboxylesterase 1D, Carboxylesterase 3B, Acylcarnitine hydrolase, Carboxylesterase 1F
1	Phospholipase A2 (Q9Z0Y2), membrane associated, Group IIC secretory phospholipase A2, Phospholipase A2 group V, Group XIIB phospholipase A2-like protein, Group XIIA secretory phospholipase A2, Group IIE secretory phospholipase A2, Group 10 secretory phospholipase A2, Group IIF secretory phospholipase A2, Group IID secretory phospholipase A2, Phospholipase A2
2	Phospholipase A2 (B7ZCM8), Cytosolic phospholipase A2, Cytosolic phospholipase A2 zeta, Cytosolic phospholipase A2 epsilon, Cytosolic phospholipase A2 delta, Cytosolic phospholipase A2 gamma
3	Lipase, member I, Lipoprotein lipase, Pancreatic lipase-related protein 2, Hepatic triacylglycerol lipase, Inactive pancreatic lipase-related protein 1, Pancreatic triacylglycerol lipase, Lipase member H, Phospholipase A1 member A, Endothelial lipase
4	Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A thioesterase 1, Acyl-coenzyme A thioesterase 6, Acyl-coenzyme A thioesterase 5, Acyl-coenzyme A amino acid N-acyltransferase 2, Peroxisomal succinyl-coenzyme A thioesterase, Bile acid-CoA:amino acid N-acyltransferase, Acyl-coenzyme A thioesterase 3, Acyl-coenzyme A thioesterase 2, mitochondrial
5	AADACL4 family member 2, Arylacetamide deacetylase-like 3, AADACL4 family member 5, AADACL4 family member 4, AADACL4 family member 3, Arylacetamide deacetylase-like 2, AADACL2 family member 3, Putative hydrolase RBBP9, Bifunctional epoxide hydrolase 2, Testis-expressed protein 30, Epoxide hydrolase 3, Neutral cholesterol ester hydrolase 1, Arylacetamide deacetylase-like 4, AADACL2 family member 1, Protein ABHD11, Kynurenine formamidase, Carboxymethylenebutenolidase homolog, Protein ABHD14B, Protein ABHD14A, Arylacetamide deacetylase, Esterase OVCA2, AADACL2 family member 2
6	Patatin-like phospholipase domain-containing protein 7, 85/88 kDa calcium-independent phospholipase A2, Patatin-like phospholipase domain-containing protein 5, Patatin-like phospholipase domain-containing protein 6, Omega-hydroxyceramide transacylase, Patatin-like phospholipase domain-containing protein 2, Calcium-independent phospholipase A2-gamma, 1-acylglycerol-3-phosphate O-acyltransferase Pnpla3
7	Dipeptidyl peptidase 4, Prolyl endopeptidase FAP, Inactive dipeptidyl peptidase 10, Dipeptidyl peptidase 8, Dipeptidyl peptidase 9, Prolyl endopeptidase-like, Acylamino-acid-releasing enzyme, Prolyl endopeptidase, Dipeptidyl aminopeptidase-like protein 6
8	Neuroligin 4-like, Thyroglobulin, Acetylcholinesterase, Cholinesterase, Neuroligin-2, Neuroligin-3, Neuroligin-1
9	Platelet-activating factor acetylhydrolase (E9QNW6), Fatty-acid amide hydrolase 1, Monoglyceride lipase, Phosphatidylcholine-sterol acyltransferase, Hormone-sensitive lipase, Mesoderm-specific transcript protein, Putative phospholipase B-like 2, Lipase member N, GPI inositol-deacylase, Protein ABHD15, Platelet-activating factor acetylhydrolase (Q60963), Epoxide hydrolase 4, Diacylglycerol lipase-α, Protein ABHD16B, Lipase member K, Lipid droplet-associated hydrolase, Protein phosphatase methylesterase 1, Lipase member M, Protein ABHD8, Valacyclovir hydrolase, Monoacylglycerol lipase ABHD6, (Lyso)-N-acylphosphatidylethanolamine lipase, Phospholipase A2 group XV, Diacylglycerol lipase-β, Phospholipase ABHD3, Gastric triacylglycerol lipase, Maspardin, 1-acylglycerol-3-phosphate O-

	acyltransferase ABHD5, Serine hydrolase-like protein, Monoacylglycerol lipase ABHD2, Protein ABHD1, Lysosomal acid lipase/cholesteryl ester hydrolase, Phosphatidylserine lipase ABHD16A
10	Abhydrolase domain-containing 12B, Lysosomal thioesterase PPT2, Palmitoyl-protein thioesterase 1, Acyl-protein thioesterase 1, Lysophospholipase-like protein 1, Palmitoyl-protein thioesterase ABHD10- mitochondrial, A/β hydrolase domain-containing protein 17B, Protein ABHD13, A/β hydrolase domain-containing protein 17C, A/β hydrolase domain-containing protein 17A, Lysophosphatidylserine lipase ABHD12, Acyl-protein thioesterase 2
11	Protein FAM83C, KAT8 regulatory NSL complex subunit 3, Lipase (D3YY49), Lipase (F6RR30), S-formylglutathione hydrolase, 5'-3' exonuclease PLD3, Lysosomal protective protein, Protein FAM83B, Protein FAM83H, Prokineticin-1, Cotranscriptional regulator FAM172A, Protein FAM83F, Inactive phospholipase D5, Phospholipase, Lipase (Q3UT41), Phospholipase D, Protein FAM83G, Mitochondrial cardiolipin hydrolase, Protein NDRG1, Pla2g3 protein, Protein FAM135A, Lysosomal Pro-X carboxypeptidase, Protein FAM83E, 5'-3' exonuclease PLD4, Protein NDRG4, Protein ABHD18, Protein FAM83A, Palmitoleoyl-protein carboxylesterase NOTUM, S-acyl fatty acid synthase thioesterase, medium chain, Malonyl-CoA-acyl carrier protein transacylase, mitochondrial, Glycosyl-phosphatidylinositol-specific phospholipase D, N-myc downstream-regulated gene 3 protein, MIT domain-containing protein 1, Dickkopf-related protein 4, Inorganic pyrophosphatase 2- mitochondrial, Transmembrane and coiled-coil domain-containing protein 4, Retinoid-inducible serine carboxypeptidase, Colipase, Transmembrane protein 53, Epoxide hydrolase 1, Probable serine carboxypeptidase CPVL, Protein FAM83D, Protein FAM135B, Dipeptidyl peptidase 2, Thymus-specific serine protease, Prokineticin-2, Protein NDRG2, Otoconin-90
12	Fatty acid synthase

Table 5: Consensus between cavity clustering and assay clustering

Cavity	Assay	Proteins that cluster in both methods in the same cluster
1	0	Lipase (F6RR30), Mesoderm-specific transcript protein, Lipase member N, Cytosolic phospholipase A2 epsilon, Lipase member K, Protein phosphatase methylesterase 1, Maspardin
1	1	Carboxylic ester hydrolase (F6Z9B9), Bifunctional epoxide hydrolase 2, Protein FAM83B, Acyl-coenzyme A thioesterase 6, Lipase (Q3UT41), Lipase member M
1	4	(Lyso)-N-acylphosphatidylethanolamine lipase, AADACL2 family member 2
1	5	Phospholipase D, Protein phosphatase methylesterase 1
1	15	Kynurenine formamidase, Esterase OVCA2
2	1	Carboxylesterase 3B, Thymus-specific serine protease
4	1	Lipoprotein lipase, Phospholipase A2- membrane associated, Lysophospholipase-like protein 1
5	0	Acetylcholinesterase, Arylacetamide deacetylase
5	1	Acyl-protein thioesterase 1, Acyl-protein thioesterase 2
8	0	Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A thioesterase 5, Bile acid-CoA:amino acid N-acyltransferase
8	1	Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A amino acid N-acyltransferase 2
13	10	Carboxylic ester hydrolase (E9PV38), Acylcarnitine hydrolase
18	10	Carboxylesterase 4A, Carboxylesterase 1D
19	1	Lysosomal thioesterase PPT2, Protein NDRG4

Table 6: Cluster consensus over all three clustering approaches

Cavity clustering	Assay clustering	Uniprot clustering	overlaps
1	0	9	Mesoderm-specific transcript protein, Lipase member N, Lipase member K, Protein phosphatase methylesterase 1, Maspardin
1	1	11	Protein FAM83B, Lipase (Q3UT41)
1	15	5	Kynurenine formamidase, Esterase OVCA2
5	1	10	Acyl-protein thioesterase 1, Acyl-protein thioesterase 2
8	0	4	Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A thioesterase 5, Bile acid-CoA amino acid N-acyltransferase
8	1	4	Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A amino acid N-acyltransferase 2
13	10	0	Carboxylic ester hydrolase (E9PV38), Acylcarnitine hydrolase
18	10	0	Carboxylesterase 4A, Carboxylesterase 1D

The dendrograms based on different characterization approaches (Figure 19, Figure 20, Figure 21) cannot be compared in terms of the absolute distances between entries and clusters because all three feature different metrics to obtain the distances between the proteins. Although what becomes clear is, that the dendrogram based on the comparison of the active site cavities (Figure 20) features cluster, that have inside a cluster shorter distances and therefore more similar entries than the other two approaches (Figure 19, Figure 21).

The comparison of the two heatmaps based on cavities and based on assay results respectively (Figure 22), show a higher resolution in the approach through cavity matching. The reason being that the assay results get reduced to 0 and 1 for their distance calculation, whereas cavity matching uses not reduced data to calculate the total score which is used in the distance matrix.

To evaluate the created cluster, the annotated catalytic activity of proteins in each cluster can be used. Here the annotation for "catalytic activity" in the Uniprot to date (27.05.2021) was taken to operate with. Taking this approach, the cavity comparison of clusters based on the active site (Table 2) show diverse results and is referred to as cavity clustering in the following sentences. The proteins in cluster 0 match in their annotated catalytic activity as carboxylesterases. For Carboxylesterase 1F its catalytic activity is proven by experimental data and for Carboxylesterase 1C it is annotated using the PROSITE-ProRule annotation (10). One of the biggest clusters in cavity clustering, the cluster 1, features a diverse group of proteins whose catalytic activity and substrate specificity is not homogenous. In cluster 2, the Carboxylesterase 5A and 3B whose catalytic activity annotation is based on PROSITE-ProRule pattern and the Phospholipase A2 group XV which features experimentally proven catalytic activity cleave a fatty acid whereas the Thymus-specific serine protease has no annotated reaction mechanism in contrast. Cluster 4 consists just as cluster 1 out of a group of proteins that have diverse catalytic functions. Cluster 5 features the same principal reaction mechanism, but the proteins differ in their substrate. For Arylacetamide deacetylase (Q99PG0), Acyl-protein thioesterase 1, Acyl-protein thioesterase 2 and Acetylcholinesterase (P21836) the annotated reaction mechanism is derived from experimental data. The proteins Carboxylesterase 1E and Liver carboxylesterase 1 feature the annotation of a catalytic activity from the Uniprot "by similarity" functionality. This uses the code ECO:0000250 for manually curated information which have been propagated from related experimentally characterized proteins (11). Furthermore, this cluster features a protein, Carboxylic ester hydrolase (E9PYP1), whose reaction mechanism has not been annotated in the Uniprot yet. In cluster 8, proteins that have a matching catalytic activity can be found. With one, the Acyl-coenzyme A amino acid N-acyltransferase 2 whose reaction mechanism is not annotated but has a sequence identity of 93.27% (12) to Acyl-coenzyme A amino acid N-acyltransferase 1, which is also present in this cluster. In cluster 13, two proteins are present: the Acylcarnitine hydrolase (Q91WG0), which features an experimentally proven catalytic activity, and the Carboxylic ester hydrolase (E9PV38), which in contrast lacks in this kind of evidence. Cluster 18 is like cluster 0 in the constellation of proteins. It too features one protein which has a experimentally proven catalytic activity (Carboxylesterase 1D) and another protein (Carboxylesterase 4A) that's only annotation is "probable carboxylesterase". Cluster 19 features not enough evidence regarding the catalytic activity for a conclusion.

Looking at the clustering based on the assay results (Table 3), which in the following is called assay clustering, two big clusters (cluster 0 and 1) can be seen which feature an inhomogeneous set of proteins based on their catalytic activity. A lower distance threshold which is defined by the maximum distance that proteins can have to each other in order to be counted in the same cluster would help to divide these big cluster in better matching clusters based on their catalytic activity. In the contrary, this would result in more proteins that do not fit into a cluster and therefore end up as an own cluster. Well matching cluster based on their catalytic function are clusters 2, 8 and 42. Cluster 10 in assay clustering features proteins with the same reaction mechanism but with different substrate sizes. Cluster 7 features two proteins that are involved in reactions containing acyl – CoA (1-acylglycerol-3-phosphate O-acyltransferase ABHD5, Acyl-coenzyme A thioesterase 2 - mitochondrial) and one protein (Abhydrolase domain-containing 12B) where the catalytic activity is not annotated yet. The rest show heterogenous collections of proteins

Clustering based on the Prosite, PFAM and GO entries, in the following called Uniprot clustering resulted in 13 clusters. Using this approach results in some well-defined clusters. The mentioned cluster numbers refer to clusters shown in Table 4. Clusters 1, 2, 3, 4 and 7 from the Uniprot clustering result in homogenous cluster, which match in their catalytic activity. The clusters 0, 5, 8 and 10 match for most of their entries catalytic activity, but feature some entries that do not indicate the same reaction mechanism. Whereas clusters 6, 9 and 11 build diverse groups of proteins.

Looking at the consensus between the cavity and the assay clustering (Table 5), one of the biggest and diverse cluster (cluster 1) from cavity clustering spreads over many clusters in the assay clustering approach. The same goes for cluster 1 from assay clustering. Overlaps that feature proteins that match based on their catalytic activity are between 5 and 0, 5 and 1, 8 and 0, 8 and 1 as well as 18 and 10, where the first cluster is always the one derived from cavity clustering and the second from assay clustering. The other cluster do not match in their catalytic activity. This classification is based on the annotated reaction mechanism for these proteins, which does not have to be complete since proteins have the capability to catalyse different reactions. Further, by looking for proteins that cluster in all three approaches in the same cluster (Table 6), overlaps that match in their catalytic activity can be found. These are Acyl-protein thioesterase 1 and Acyl-protein thioesterase, Acyl-coenzyme A amino acid N-acyltransferase 1 and Acyl-coenzyme A amino acid N-acyltransferase 2, Carboxylesterase 4A (annotated as "probable Carboxylesterase") and Carboxylesterase 1D and the group containing Acyl-coenzyme A amino acid N-acyltransferase 1, Acyl-coenzyme A thioesterase 5, Bile acid-CoA amino acid N-acyltransferase. Furthermore, the groups Kynurenine formamidase and Esterase OVCA2 as well as Carboxylic ester hydrolase (E9PV38) and Acylcarnitine hydrolase both feature one protein with an annotated reaction mechanism (Kynurenine formamidase and Acylcarnitine hydrolase) and one without. The group containing Protein FAM83B and Lipase (Q3UT41) as well as the group with the proteins Mesoderm-specific transcript protein, Lipase member N, Lipase member K, Protein phosphatase methylesterase 1 and Maspardin do not show a uniform reaction mechanism.

Discussing conclusion

The sequence identity of these proteins is not suitable to predict catalytic similarity because the pairwise sequence identities are too small (Figure 2) and the clusters do not involve enough proteins. The fact that the fold is more conserved than the sequence is supported by the results presented in the set of proteins used in this thesis.

The high diversity of ABHD containing proteins is highlighted by the wide variety of PFAM, Prosite, Gene Ontology and ExPASy enzyme entries for this set of proteins. When using these annotations, originating from mostly sequence based information, categorical clustering of proteins worked well and resulted in clusters that were expected (Figure 19). It is important to point out that out of all these annotations, GO terms are the most abundant and therefore have the most influence on the outcome of the clustering process, since some proteins lack PFAM or Prosite annotations.

Considering that different homology modelling programs result in different outcomes of models due to the different selection of templates or the choice of force fields, the use of another program could result in different homology models and therefore change the outcome. The template selection in this set of proteins plays a significant role due to the flexible lid structure which is often present in these enzymes. The ideal selection for a template for each protein present would be with a lid that is not totally closed. Thus, making it possible to procreate a cavity and in addition gaining influence for the lid residues on the cavity properties (Figure 11). A completely closed lid is not optimal because it often leads to failure in cavity procreation. This is due to a too small room caused by residues that rotate in the area where the cavity is located during the energy minimization (Figure 8). Other factors that play a major role in the outcome of the modelling process are sequence similarity and the ratio of aligned to total residues. During the procedure of comparing models with similar sequence similarity to the respective template but a different ratio of aligned residues to total residues, the importance of this parameter becomes clear. An example would be the model of GPI inositol-deacylase (Q3UUQ7) (Figure 6), which features a sequence similarity of 52.3% and a 151:922 ratio of aligned to total residues, however results in a model that cannot be utilized. On the contrary, the model of Esterase OVCA2 (Q9D7E3) (Figure 11) features the almost same sequence similarity of 52.9% but a better ratio of aligned to total residues of 172:225 and therefore leads to a model that can be used despite the small sequence similarity.

Procreating cavities for a large number of proteins with the same settings can be a challenging task, owing to the difference in size of the proteins and the difference in cavity size. To receive the highest number of possible cavities, a high maximum size of the cavities of 2500 Å^3 was applied (Table 10). This worked out for most proteins but caused some oversized cavities that take up more space than only the true catalytic site and its substrate entrance area.

The ability to evaluate cavities based on the presence of catalytic motifs and therefore make finding the cavity of interest easier, was the main idea behind robust motif search enhances and brute force enhanced. However, smaller motifs occur more often in proteins (Figure 14). Hence the smaller the motif, the less specific is their appearance. The reason therefore is the need to meet less distance constrains, which in turn raises the probability of random appearances. This applies especially to motifs that feature only two residues. Meaning that this method alone cannot be operated in order to annotate a catalytic activity to a specific protein. In conjunction with the procreated cavities, it gives a clue where the cavity of interest could be located in. Its necessity is represented in the fact that most of the time proteins have more than one cavity procreated and the ability to determine the cavity of interest is a useful tool. This statement can be verified throughout the group of 210 proteins. Nevertheless, beside the presence of a motif was of interest, the location of cavities relative to the α-β hydrolase fold could be used as a criterion for the choice of the active site cavity as well. With the use of 2 Å maximum deviation from the original motif, many motifs could be found, sometimes more than one of the same motif-type near different cavities. This high possible deviation, however, was chosen since this work is based on homology models, where the energy minimization in the building process of the models can lead to bigger distances than in crystal structures. To get a higher discriminatory power during the motif search approach and less false positive hits, a smaller maximum deviation should be taken into consideration when dealing with crystal structures.

The procedure of cavity matching and the results after clustering show mixed results. In the different clustering results (Table 2, Table 3, Table 4), the most cluster matching clusters based on their catalytic activity annotation is derived from the Uniprot clustering. It shows, analogous to the other two approaches, clusters that fit well and clusters that are diverse in their catalytic activity. An example for an expected cluster would be the five proteins that are annotated as Phospholipase A2 and feature the Phospholipase-A like motif in their structure. These are Phospholipase A2- membrane associated, Group IIE secretory phospholipase A2, Group 10 secretory phospholipase A2 (failed two times in cavity matching and therefore had to be excluded), Group IIF secretory phospholipase A2, Group IID secretory phospholipase A2. They can only be found in the Uniprot clustering in one cluster.

The largely diverse cluster in cavity clustering could be a result of the size of the cavities. A large cut- off for the cavity size (2500 Å) was required to obtain cavities for all proteins, which caused overlaps in cavities in areas that are not directly involved in the catalytic action of the protein and therefore do not represent catalytic similarity in proteins. A method to prevent this error from occurring, would be to use shaped cavities. Here a ligand is set into the active site of the protein and the procreated cavity would be cut around that specific ligand. Consequently, only the segment which is important for the catalytic function would be operated on for matching. Another reason for heterogenous clusters is cavity overlap during the matching process (Figure 16). An optimal overlap would be 75% and above for both cavities. This does not appear often in this set of cavities. Furthermore, the templates have an impact on cavity properties and therefore on the matching results. Since 113 templates were used for the 210 proteins, proteins that do not catalyse the same reaction can have a similar cavity due to the same template and therefore create an unwanted optimized match. To obtain better defined clusters with less heterogenous catalytic activities, squaring the matching- scores results was tested. In theory, this would impact proteins in closer proximity less than proteins with bigger distances to each other. Thus, keeping related proteins closer together and spread proteins with less relation further from each other. However, this did not work as expected.

The results of the assay clustering follow the same scheme as cavity clustering. In order to obtain better matching clusters, more assays that are tailored and more specific to different subgroups of enzymes in this set of 210 proteins would be an advantage. In addition, the activity should be a quantitative measure. Since the activity results used in this analysis were converted to 0 (not active) and 1 (active), the resolution of the results was lowered which can lead to proteins having the same distance and therefore interfere with the clustering process. Moreover, different clustering methods like K-means, spectral clustering, DBSCAN, optics and affinity propagation, as implemented in scikit-learn (13), were tested. Here the distance matrix was not supplied but rather the "coordinates" of the proteins in "activity space". The outcome of the clustering was similar which underlines again the importance of data with high discriminatory power.

Function prediction

The results of the assay clustering as well as the cavity clustering feature some clusters that share a common catalytic activity for all entries, except for one protein which is not yet fully annotated. This can be used to get a hint to the catalytic activity when an annotation for a protein is missing. This information can therefore be used to indicate the catalytic activity of the non-annotated protein.

Cavity cluster 13 features the following two proteins, Carboxylic ester hydrolase (E9PV38) and Acylcarnitine hydrolase. These proteins are present in one cluster in assay clustering and in Uniprot clustering as well (Table 6). Since the Carboxylic ester hydrolase (E9PV38) has no annotated catalytic activity yet, one could assume that it can also catalyse the same reaction (Figure 23).

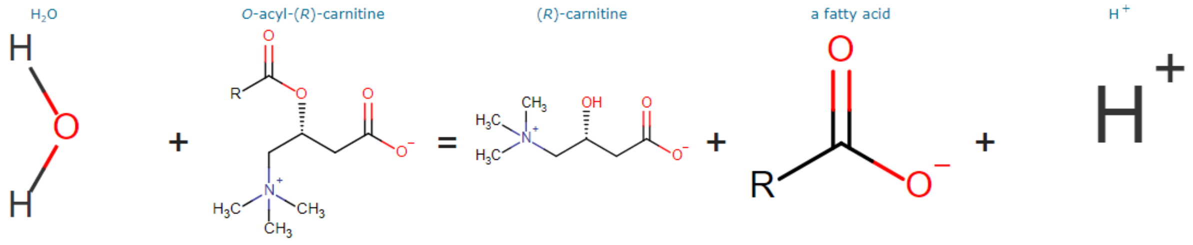

Figure 23: Reaction prediction for Carboxylic ester hydrolase (E9PV38) - reaction scheme from (24)

The same situation is present in part of cavity cluster 1 which agrees with assay cluster 15 and Uniport cluster 5 with the proteins Kynurenine formamidase and Esterase OVCA2. Here the Esterase OVCA2 has no annotated catalytic activity whereas the Kynurenine formamidase catalyses the following reaction (Figure 25).

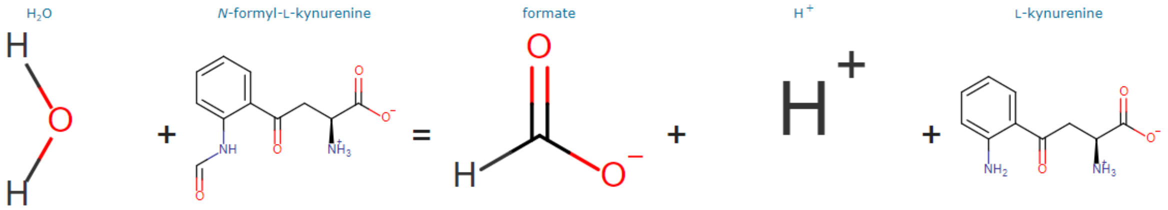

Figure 25: Reaction prediction for Esterase OVCA2 - reaction scheme from (25)

This leads to the assumption that the Esterase OVCA2 could catalyse the same reaction The same goes for cavity cluster 18 which features the proteins Carboxylesterase 4A and Carboxylesterase 1D. Here the Carboxylesterase 1D has an annotated catalytic activity (Figure 24) and the Carboxylesterase 4A is annotated as “probable Carboxylesterase”.

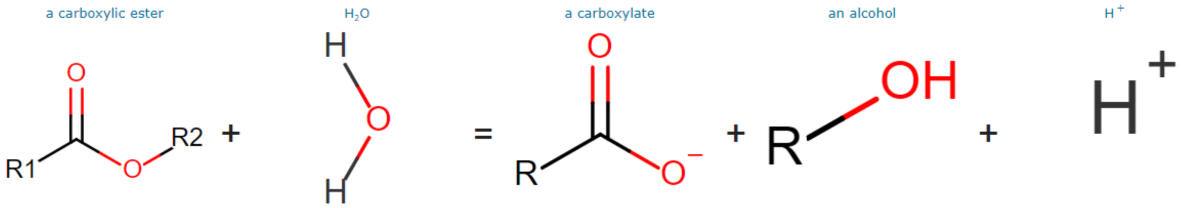

Figure 24: Reaction prediction for Carboxylesterase 4A - reaction scheme from (26)

This can lead to the assumption that the Carboxylesterase 4A catalyses the same reaction. For assay cluster 7, which also features a similar condition, a prediction is rather unclear since this cluster does not appear in the cavity clustering.

Taking the same approach as described above for only the cavity clustering this would lead to the following predictions: In cluster 8 Acyl-coenzyme A amino acid N-acyltransferase 2 could catalyse the same reaction as the other proteins in the cluster.

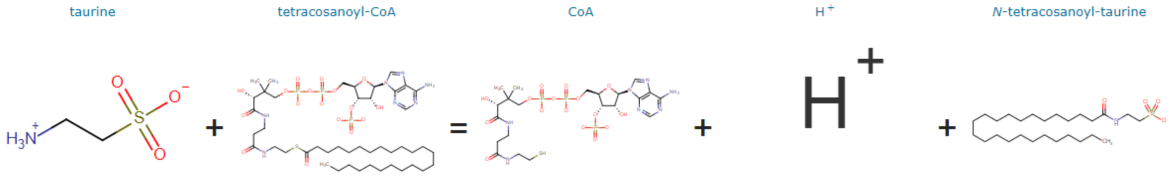

Figure 26: Reaction prediction for Acyl-coenzyme A amino acid N-acyltransferase 2 - reaction scheme from (27)

Especially since it features a high sequence identity to Acyl-coenzyme A amino acid N-acyltransferase 1, the reaction it catalyses can be the same (Figure 26).

For cluster 0 and 18, this is also indicated by the consensus of all three approaches. For assay cluster 7, the proteins 1-acylglycerol-3-phosphate O-acyltransferase ABHD5 and Acyl-coenzyme A thioesterase 2- mitochondrial both hydrolyse an acyl-CoA and transfer the acyl group to a donor (14) (15). Therefore, the conclusion that AB hydrolase domain-containing 12B (G3UZN6) could catalyse a similar reaction is acceptable.

Methods

The code used for working with the data outside of the Innophore platform can be found online https://git.innophore.com/gregorw/a_b_hydrolase_analysis. This includes the code used in the motif search algorithms.

Sequence identity

To compare all given sequences the multiple sequence alignment tool Clustal w (https://www.ebi.ac.uk/Tools/msa/clustalo/) (12) which is provided as a web application by the European Bioinformatics Institute (EMBL-EBI) with its default settings (Table 8) was utilized. The obtained information, specifically the percent identity matrix was used to obtain all pairs of proteins that have a higher sequence similarity than 80%, a heatmap of the percent identity matrix as well as the mean, minimum and maximum sequence identity.

Cross References

Since most proteins in the list are annotated in the Uniprot database, the search was targeted at cross references to other protein family and domain databases. The selected databases of interest were the following:

Pfam

This database features a large collection of protein families, which are each represented by multiple sequence alignments and hidden Markov models. (16) Its sequence data is based on the UniProtKB reference proteomes. PFAM entries can feature only unique sequences to avoid the appearance of a region of a sequence in more than one PFAM family. A small overlap between families is allowed, because resolving these would be a time-consuming task. All PFAM entries are manually annotated including functional information from available literature if possible (17).

ExPASy Enzyme

ExPASy Enzyme is a nomenclature database meaning information relative to the nomenclature of enzymes is stored. This mechanism is primarily based on recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (18).

Gene ontology – Molecular function

The Gene ontology project provides a computational representation of the current scientific knowledge about the function of proteins and non-coding RNA molecules built by genes. One of its aims is to grasp a better understanding on how individual genes contribute to the biology of an organism at the molecular, cellular and organism level (19). Gene ontology features three categories, which are the biological process, the molecular function and the cellular component. To put data into use that refers to the molecule itself, the "molecular function" data is used. It is defined by the biochemical activity of a gene product, including specific binding to ligands or structures. Factors such as where or when this happens are not taken into consideration, only its function play a role. (20) To receive for the project relevant annotations that are directed towards enzyme activity, GO entries which contain "activity" in their name were exclusively used.

Prosite

Prosite entries describe protein domains, families and functional sites as well as associated patterns and profiles to identify them. These are based on the observation that proteins can be grouped into families based on similarity in their sequences. The regions that are conserved for the function or for the three-dimensional structure of a protein can be analysed for its constant and variable properties in groups of similar sequences respectively. From that, a unique signature for a protein family or a domain can be extracted (21).

The annotations of PFAM, Gene ontology and Prosite for all proteins were then used to cluster them. Here hierarchical, categorical clustering was applied, where the relationship of one entry to another is not taken into consideration. Instead, all protein entries are converted into a binary list with the length of all possible cross references found in the given proteins, where 0 stands for the absence and 1 for the presence of a motif. For clustering the proteins, the ward method was used as a linkage method.

Additionally, the number of occurrences of all database entries in the stated sequences was checked.

Homology models

Since for only 13 (Q9DBL9, Q6PE15, Q3TCN2, P21836, P19096, P34914, Q9QYG0, Q8VDV8, Q99K10, Q5SWZ9, Q69ZK9, P63005, O35298) of the 210 sequences, crystal structures are available, homology models were built for all sequences. The homology model building functionality of the Catalophore platform was chosen to build the models (https://uni-graz.catalophore.com/CATALObase2/) with its default settings (Table 9).

Cavity procreation

The comparison of the active site of all proteins was made possible by the cavity procreation of the Catalophore platform. Through the help of a modified LIGSITE algorithm the CavFind algorithm creates point clouds in areas of the protein of interest where an active site might be. These cavities represent the physicochemical properties of the space they occupy. Since cavities in some of the proteins are quite large, a maximum volume of 2500 Å^3 was used (Table 10). This can result in multiple procreated cavities where not every cavity is automatically the cavity of interest. Therefore, it is necessary to examine the proteins manually and select the appropriate cavity. The correct cavity in this context is defined as the point cloud that represents the substrate binding cavity in a protein. Furthermore, the substrate binding cavity was estimated to be the cavity that is near the α- β hydrolase fold in a protein.

Motif search

As mentioned above, multiple cavities need to be inspected manually. Therefore, it is important to be aware where a procreated point cloud represents the cavity of interest and where it does not fulfil the requirements. One the one hand, the α-β hydrolase fold can operate as an indication for this purpose. On the other hand, different motifs in proteins can be attributed to a protein's functionality. These motifs can feature starting from two up to nine residues that are in a distinct distance to one and another. The geometric constellation of these motifs is consequential for the interaction of amino acids with each other. The data of the hydrolase atlas (22) provided a wide variety of hydrolase motifs. The catalytic residues of several motifs are stored as PDB files. This data was employed to calculate the pseudo atom positions of these residues and their distances to one other. The presence of a motif near a cavity verifies that this point cloud embodies the cavity that represents the active site of the protein. Brute force enhanced as well as robust motif search enhanced were run to search these motifs. The maximum deviation of a motif in a protein based on the template motive (Table 11), which still classifies as a valid motive was 2 Å. This high cushion was chosen since the data purely consists of homology models: The energy minimisation during the building process of these homology models can lead to the rotation of residues that causes bigger distances between the residues.

Cavity validation

The final validation of cavities was done by searching for point clouds near the α-β hydrolase fold and comparing these with the results of the algorithm. The requirement for the recognition of a point cloud as the correct cavity, at least one residue of the detected motif needs to be as close as 3 Å or closer to at least one point in the point cloud and near the α- β hydrolase fold. The only exception applies for proteins, where the point cloud was near a Phospholipase A like motif and where the protein is already annotated as Phospholipase A in the Uniprot. In this case, the cavity near that motif was used even when there was no α-β hydrolase fold present. This was done since this motif consists out of five residues, this makes a random presence unlikely. The combination of the presence of a motif containing multiple residues in combination with the same Uniprot annotation enable a strong indication for the correct cavity to be found during the search despite the absence of an α- β hydrolase fold.

Algorithms

Both algorithms brute force enhanced and robust motive search feature the same mechanism to deal with amino acid positions. To reduce the position of each amino acid to one point, every type of amino acid is assigned to a designated pseudo atom position. The mean of all catalytically important atoms in an amino acid is defined in Table 12. Compared to the use of one specific atom in each amino acid, the utilization of the pseudo atom position grants the algorithm more independency in terms rotation of the amino acid side chain. Regarding the position to where the side chain of the amino acid is pointed to, this specific approach is more precise than the position of the Cα atom.

Beside the pdb-file of the protein, both algorithms depend on the motif of interest and the magnitude of the variation regarding the distances between the residues in the motif. Therefore, two parameters need to be given to define the motif. Firstly, amino acids that build the motive such as SER-ASN-ARG-SER are needed. Secondly, the distances between all residues must be stated, for example 7.3 Å – 7.8 Å – 7.7 Å - 9.0 Å - 9.0 Å - 9.0 Å. For motifs, the number of distances d_n to fully describe a motive is defined in the equation $d_n = \frac{n*(n-1)}{2}$,where n is the number of residues in a motif. The distances are the upper triangle of the distance matrix.

Brute force enhanced

This algorithm was created by Univ.-Prof. Dr. Karl Gruber.

It starts with calculating all distances between all amino acids and creates a distance matrix. Then, the indices off residues of the type of amino acids that are present in the motif get extracted. The next step is to reduce the possible combinations of that later need to be checked. This is done by searching for all pairs of the right amino acid type that obey the distance criterion and iteratively updating the list of possible residues that can be used to form pairs.

After that, the lists of possible residues get joined together to form all possible combinations. These get checked to only feature unique residues and only the number of residues present in the motif. For every combination, the distances between all residues get extracted from the distance matrix. From this newly generated list of distances between these residues, the obeying of all distance criteria get checked. If all distances are within the optimal distances +/- the selected cushion, the found motif gets tagged as found motif and later returned.

Robust motive search enhanced

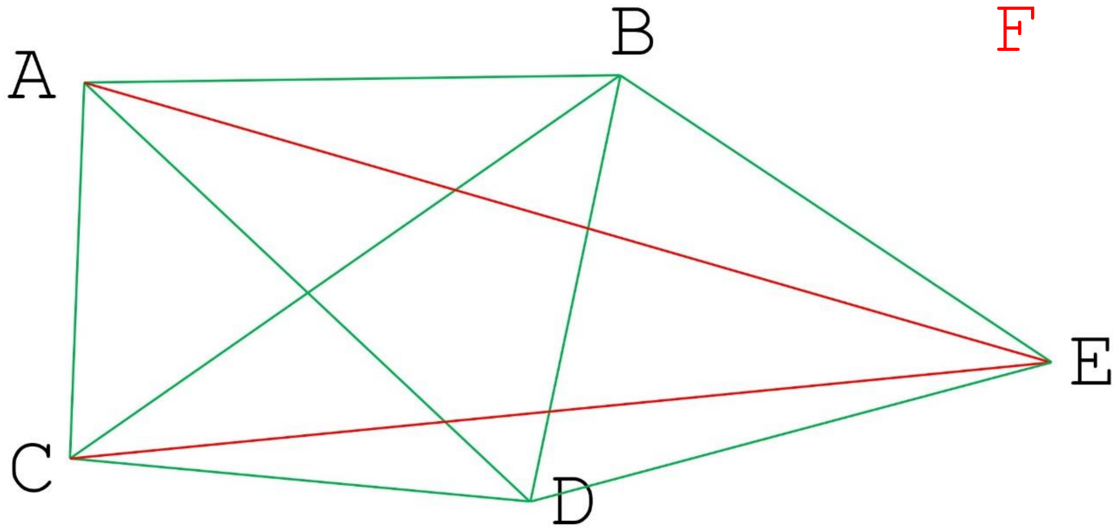

Figure 27: **Motif constellation of a hypothetical arrangement of residues**, where the length of the vectors AB, AC, AD, BC, BD, BE, CD, DE in green obey the given distance criteria and the points A, B, C, D and E fulfilling the requirement of the correct amino acid needed for the given motif.

The stated distances, pairs and triangles in the algorithm description refer to Figure 27. In the following example, the motif of interest consists of 4 residues.

In the beginning, the settings are checked. Then, the distances between all amino acids are calculated in one step. To continue, extracting amino acid pairs featured in the motif and saving them as pairs if they obey the distance criterion is necessary. The following pairs, which further are referred to as primary pairs are:

AB, AC, AD, BC, BD, BE, CD, DE

Out of these mentioned all possible pairs get built, which in turn go by the term possible triangles:

ABAC, ABAD, ABBC, ABBD, ABBE, ABCD, ABDE, ACAD, ACBC, ACBD, ACBE, ACCD, ACDE, ADBC, ADBD, ADBE, ADCD, ADDE, BCBD, BCBE, BCCD, BCDE, BDBE, BDCD, BDDE, BECD, BEDE, CDDE

To verify whether these are triangles are truly present and not artefacts of the random combination, all possible triangles are examined as follows:

In the first step, a possible triangle is inspected whether it features a common point. For ABAC this would be A. In contrast, ABCD would fail because it does not fulfil this requirement and therefore can not form a triangle. Staying with the example of ABAC, the unique entries ABC are split into all possible pairs:

AB, AC, BC

These are re-examined whether they are present in primary pairs. Since this is the case, the triangle ABAC is stored as a valid triangle ABC. It is worth mentioning that the list of these triangles is called "valid triangles". On the contrary, the possible triangle ABBE would pass the first condition of featuring a common point, but would fail the second one because this would result in:

AB, AE, BE

Here AB and BE are present in primary pairs, however AE is absent, which is why ABBE is not considered as a valid triangle.

All valid triangles, with the character of being sorted and unique would be the following, which furthermore are classified as valid triangles:

ABC, ABD, ACD, BCD, BDE

In the next step, all valid triangles are grouped into motifs that possesses as many triangles as needed to feature enough points to build the motif. For the sample motif, combinations of two triangles are required to be built because the motif consists of four residues and the minimal number of triangles to feature at least four residues is two triangles.

Every possible combination of triangles would be as follows, which are defined as possible motifs:

ABCABD, ABCACD, ABCBCD, ABCBDE, ABDACD, ABDBCD, ABDBDE, ACDBCD, ACDBDE, BCDBDE

Each combination of possible motifs is examined in the same manner as the possible triangles above. The amount of unique points and the number of residues in the motif get reviewed. After fulfilling this criterion, all possible triangles from the possible motif are created. For a possible motif like ABCABD which features four unique residues ABCD, this would result in the following triangles:

ABC, ABD, ACD, BCD

Due to the fact that these triangles are present in valid triangles, this motif is confirmed and get stored as a valid motif.

The possible motif ABDBDE would meet the conditions of the unique entries test with ABDE and result in the following triangles:

ABD, ABE, ADE, BDE

ABD, BDE are present in valid triangles, but ABE and ADE are absent. This displays another example for a combination that cannot be categorized as a valid motif.

In the end all unique valid motifs get split into all possible pair combinations once again and checked whether they originate from different primary pairs. If so, the outcome is the return of the valid motif.

When testing brute force enhanced and robust motive search against each other with 10000 different motif- protein combinations, which are combinations of motifs and the proteins in the data set, and 5000 combinations of randomly created motifs with the proteins of the data set, they differ only once in a result. In this case, it was impossible to determine which algorithm is correct. The motif consisted of ASN-ASP-ASP-ASP-ASP and the distances could not be mapped to the ASP, therefore impeding the determination of a false positive result or classifying the outcome as correct.

After finding valid motifs with one of the algorithms, the distance of every residue's pseudo atom position to each point in all cavities is calculated using the cavity distance calculator. When a point in a cavity is as close or closer than 3 A to a residue in the motif, the cavity gets tagged as validated by the specific motif. This information can further be used as an assistance to find the true cavity that represents the active site of the proteins or to give a clue on cavities that might fail to be noticed.

Cavity matching
This method compares the physicochemical properties of point clouds against each other, which in turn allows multidimensional comparison. When the lowest root mean square deviation (RMSD) regarding the multidimensional comparison of the physicochemical properties is calculated, the point clouds get matched based on their special properties to determine the biggest overlap of the cavities in total. This was done using the Catalophore platform and its default settings (Table 13). Matching is implemented for all cavities against each other, which means that every cavity matching is run twice. As an example, cavity protein A against cavity protein B will be matched at first and in the second time cavity protein B against cavity protein A. Cavity protein A against cavity protein A will be matched as well, which show a score of 0 because they are identical.

Table 7: Proteins used for cavity procreation with proteins where two cavities were used (marked as **bold**)

A2A752, A2A7Z8, A2AKK5, B0F2B4, B1AVU6, B2RWD2, B7ZCM8, D3YY49,
D3Z5G7, E9PV38, E9PYP1, E9QNW6, F6RR30, F6YQT7, **F6Z9B9**, G3UZN6,
H3BKH6, **O08914**, O35448, O35678, O55137, O88531, O88851, P11152,
P16301, P19096, **P21836**, P23953, P27656, P31482, **P34914**, P47713,
P97823, Q03311, **Q07646**, Q08ED5, Q0VBM2, Q148V8, Q32Q92, Q3TCN2,
Q3TUU5, Q3U4B4, Q3UFF7, Q3UT41, Q3UWT7, Q3V1F8, Q50L41, Q50L42,
Q50L43, Q5BKQ4, Q60963, Q62433, Q63880, Q64176, Q6AW46, Q6IE26,
Q6PDB7, Q6PE15, Q6Q2Z6, Q7TMR0, Q80UX8, Q80YU0, Q8BGG9, Q8BK48,
Q8BLF1, **Q8BM14**, Q8BM81, Q8BTG7, Q8BUY2, Q8BVQ5, Q8BWN8, Q8CIV3,
Q8K2A6, Q8K2P2, Q8K4F5, Q8K4H1, Q8QZR3, Q8R0P8, **Q8R0W5**, Q8R116,
Q8R146, Q8R164, Q8R197, Q8R2Y0, Q8R3F5, Q8VCC2, Q8VCT4, Q8VCU1,
Q8VD66, Q8VEB4, **Q8VI78**, Q91WC9, Q91WG0, Q91WU0, Q91X34, Q91ZH7,
Q920A5, Q99K10, Q99LR1, Q99PG0, Q9CPP7, Q9CQC8, Q9D7E3, Q9DB29,
Q9DBL9, Q9ET22, Q9QUL3, Q9QUR6, Q9QXE5, Q9QXX3, Q9QYR7, Q9QYR9,
Q9QZT4, Q9WTL7, Q9WVF6, Q9WVG5, **Q9Z0M5**, Q9Z218, W4VSP6

Furthermore, 119 different proteins (Table 7) were selected and from them 129 different cavities were matched against each other. The therefor selected cavities are near an α-β hydrolase fold. This does not apply for proteins that feature a Phospholipase A motif and is annotated as a Phospholipase A in the Uniprot. Thus the cavity near the motif was used instead even if the α- β hydrolase fold was absent.

Assay based distance calculation

All 210 proteins were screened in the laboratory of Assoz. Prof. Mag. Dr.rer.nat. Zimmermann at the Institute of Molecular Biosciences of the University of Graz by Johannes Breithofer, BSc with the following screens:

- **pNPV** Assay pH7.4 - Lysate
- **pNPV** Assay pH7.4 - Supernatant
- 1 mM **Triolein** (2% BSA; 0.3% NP40) in 25 mM NaCitrate buffer pH 5 – Lysate - FFA release
- 1 mM **Triolein** (2% BSA; 0.3% NP40) in PBS pH 7.4 - Lysate - FFA release
- Human **Lipoprotein** Screen (0.8 mg/ml protein 2% BSA) - Supernatant - FFA release
- Human **Lipoprotein** Screen (0.58 mg/ml protein 2% BSA) - Lysate- FFA release
- **Lipiddroplet** assay (2 mM TG, pH 7.4) -Lysate - FFA release
- **Lipiddroplet** assay (1.35 mM TG, pH 5) - Lysate - FFA release
- 0.5 mM **BMP (S,S)** Assay pH 7.4 (2% BSA 2.5 mM CHAPS) - Lysate – FFA release
- 0.25 mM **Hemi-BMP (S,R)** Assay pH 7.4 (2% BSA 2.5 mM CHAPS) - Lysate - FFA release
- **18:1 Monoglycerid** Assay pH 7.4 (2%BSA 1.69 mM CHAPS) - Lysate - Glycerol release/FFA release
- **18:1 Diglycerid** assay (+PC:PI) pH 7.4 + 2% BSA - Lysate - Glycerol release (modified Glycerol reagent!)

The description mentioned below regarding the execution of assays was provided by Johannes Breithofer, BSc.

Triolein assay:

The Expi lysates were screened with 1 mM Triolein (TO) in Phosphate-Buffered Saline(PBS) (pH 7.4) and 0.25mM NaCitrate buffer +137mM NaCl + 2.7 mM KCl (pH 5). 10 µl (=20 µg) of Expi lysates (2 mg/ml) were incubated for two hours at 37 °C with 30 µl of substrate. Afterwards, the substrate was prepared with 2 % BSA and 0.3% NP40. The "Wako NEFA kit" (100 µl Reagent 1 + 50 µl reagent 2) was utilized to measure the release of free fatty acid.

Lipoprotein assay:

The supernatant of the Expi transfections were incubated with human lipoproteins. In the following sentences, the assay conditions are portrayed: 30 µl of SN + 10 µl of lipoprotein substrate; incubation for four hours at 37 °C; measurement of the release of free fatty acid (FFA) with the "WAKO NEFA Kit" (100 µl reagent1 + 50 µl reagent2). Thereafter, desalination of the lipoproteins in a size-exclusion column (elution in 4 ml PBS pH 7.4) performed. The protein concentration of the lipoprotein solution was determined and diluted to 2.4 mg/ml. Subsequently, 6% of BSA were added to this solution. The assays were performed once with and without 3mM CaCl2/ 3mM MgCl2.

Lipid droplet assay:

Lipid droplets isolated from a fasted mouse liver were incubated with the Expi cell lysates. The lipid droplets were diluted to a triglyceride concentration of 4 mM with PBS pH 7,4 ("lipid droplet stock concentration" in PBS pH 7,4 is 10 mM). This LD solution was prepared once with 4 % BSA (essential fatty acid free) and once with 2mM CaCl2/MgCl2. 20 µl of Expi lysate (protein conc. = 2 mg/ml) were mixed for the assay with 20 µl of the LD substrate, againonce with CaCl2/MgCl2 and once without, and incubated at 37°C for one hour. During this incubation period the plates were carefully mixed twice by using the Vortex to ensure even distribution of the lipid droplets in the wells. FFA release was measured with the "WAKO NEFA kit" (100 µl reagent 1 + 50 µl reagent 2). Since the LD solution in wells is cloudy, photometric measurement of the plates needs to be enabled. To achieve this goal, the LDs areseperated by centrifuging the plates at 3000 rpm for 10 min (after the NEFA reaction). The positive control for the assay is shown as follows: 1. 20 µl ATGL expi lysate + 0,5 µl recombinant CGI58 (from Gernot); 2. 10 µl ATGL expi lysate + 10 µl HSL expi lysate 3. 10 µl ATGL expi lysate + 10 µl CGI58 expi lysate.

BMP assay:

20 µl of Expi lysates (protein conc. 2 mg/ml) were incubated with 20 µl of 18:1 BMP (S,S) (sn-(3-oleoyl-2-hydroxy)-glycerol-1-phospho-sn-1'-(3'-oleoyl-2'-hydroxy)-glycerol) for one hour at 37 °C. The FFA release was measured with the "Wako Nefa Kit" (100 µl R1; 50 µl R2). To prepare the substrate, 18:1 BMP (S,S) powder (10 mg) was dissolved in 1 ml Chloroform. After Chloroform was fully evaporated, the lipid was subsequently sonicated in PBS pH 7.4 + 4% BSA (essential free fatty acid free) + 5 mM CHAPS so that the 18:1

BMP (S,S) concentration reaches 1.14 mM in the solution. The assay conditions were 0,57 mM 18:1 BMP (S,S), 2% BSA and 2.5 mM CHAPS once with and once without 1 mM CaCl2 1 mM MgCl2.

Hemi BMP assay:

20 µl of Expi Lysates (protein conc. 2 mg/ml) were incubated with 20 µl of 18:1 Hemi BMP (S,R) (sn-(3-oleoyl-2-hydroxy)-glycerol-1-phospho-sn-3'-(1',2'-dioleoyl)-glycerol for one hour at 37 °C. The FFA release was measured with the "Wako Nefa Kit" (100 µl R1; 50 µl R2). To prepare the substrate, the 18:1 Hemi BMP (S,R)(5 mg) was dissolved in 1 ml Chloroform. After Chloroform was fully evaporated and the lipid was subsequently sonicated in PBS pH 7.4 + 4% BSA (essential free fatty acid free) + 5 mM CHAPS so that the 18:1 Hemi-BMP (S,R) concentration reaches 0.5 mM in the solution. The assay conditions were 0,.25 mM 18:1 Hemi-BMP (S,R), 2% BSA and 2.5 mM CHAPS once with and once without 1 mM CaCl2 1 mM MgCl2.

Monoglyceride assay:

10 µl of Expi Lysates (protein conc. 2 mg/ml) were incubated with 20 µl of 1 mM 18:1 MG (-oleoyl-ra c-glycerol). Substrate preparation was carried out after Chloroform (solvent of the lipid) was fully evaporated and the lipid was subsequently sonicated in PBS pH 7.4 + 3 % BS (essential fatty acid free) + 2.5 mM CHAPS with and without 1 mM CaCl2/ MgCl2. The assay was incubated for 30 min at 37 °C. All assays were prepared in duplicates to enable the measurement of Glycerol release and FFA release with the respective detection kit.

Diglyceride assay:

20 µl of Expi lysates (protein concentration 2 mg/ml) were incubated with 20 µl of 1,67 mM rac-Dioleoylglycerol . The substrate was prepared with PC:PI 3:1 (7.5 µl of 20 mg/ml stock for every ml of substrate solution) in PBS pH 7.4 containing 2% BSA (FFA free). After incubating the assay for 30 min at 37 °C, Glycerol release was measured with the Glycerol reagent (100 µl/well) containing 1 µl of purified bacterial Monoglyceride Lipase per 100 µl glycerol reagent. With this modified glycerol reagent, the assay can detect reactions where monoglycerides are the reaction product, for exampleDAG --> H20 + FFA + MAG; Bacterial monoglyceride LipaseMAG --> H20 + FFA + Glycerol

These assays are based on the free fatty acid release after hydrolysis through the enzyme. This causes a change in colour, which in turn can be measured through the change in absorbance. In each screen the resulting absorbance was quantified and compared to the absorbance when only an empty vector HisMaxC was expressed. If the absorbance of a specific sample was 1.5 times higher than the absorbance (abs.) of the negative control (neg. control) HisMaxC, the protein was tagged as active in the assay.

Furthermore, this data was used to cluster the proteins based on their assay results. The results of the assays were converted to 0 for not active (abs. sample < abs. neg. control*1.5) or 1 for active (abs. sample > abs. neg. control*1.5). Since some neg. controls had a negative value, the mean of all neg. controls was used for comparison. Consequently, each proteins' data consists of a list assigning 0 and 1 for each assay and condition. In total, every protein's entry is built on 40 data points. To calculate the similarity among the proteins, the Euclidean distance between each protein was evaluated in order to obtain a distance matrix. This was then used to cluster the proteins, which is further described in the chapter "result clustering" (page 54).

Result clustering

Clustering of the results was operated through agglomerative clustering with a precomputed distance matrix (23). Hierarchical clustering is a type of clustering algorithm that build nested clusters by merging or splitting them successively. A dendrogram helps to visualize the hierarchy of the clusters. The root of the dendrogram is a cluster containing all samples, whereas the leaves display the smallest clusters with only one sample. This was subsequently used to set the right distance cut off for the clustering algorithm to obtain clusters with the desired size. The agglomerative clustering algorithm applies the bottom-up approach. Here each sample starts as its own cluster, which are merged consecutively (13). The precomputed matrices are either obtained from the file "cavity matching result" or by calculating the Euclidean distance between the proteins based on their assay results. In the case of cavity matching, the distance used is the total score divided by the smaller overlap of the matched cavities.

The clusters obtained from the dendrogram, which is based on Uniprot cross references, are specified as clusters with the requirement of a distance range smaller than 5.7 (Figure 19).

Literature

1. Ollis D., Cheah E., Cygler M., Dijkstra B., Frolow F., Franken S., Harel M., Remington J., Silman I., Schrag J., Sussman J., Verschueren K. ,Goldman A. *The α/β hydrolase fold.* [Protein Engineering, Design and Selection] April 1992. Vol. 5.

2. Bachovchin D, Ji T., Li W., Simon G., Blankman J., Adibekian A., Hoover H., Niessen S., Cravatt B. *Superfamily-wide portrait of serine hydrolase inhibition achieved by library-versus-library screening.* [PNAS] 7 12 2010. Vol. 107.

3. Shahiduzzaman M., Coombs K. *Activity based protein profiling to detect serine hydrolase alterations in virus infected cells.* [frontiers in Microbiology] 2012. Vol. 3.

4. Callaway, E. *'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures.* [Nature] 2020. Vol. 588.

5. uniprot. [Online] https://www.uniprot.org/uniprot/.

6. Proteine Data Bank. [Online] https://www.rcsb.org/search.

7. yasara.org. [Online] [Cited: 03 05 2021.] http://yasara.org/homologymodeling.htm.

8. Huang B., Schroeder M. *M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.* [BMC Struct Biol] 2006. Vol. 6.

9. python.org. [Online] [Cited: 03 05 2021.] https://www.python.org/about/.

10. prosite expasy prorule. [Online] [Cited: 27 05 2021.] https://prosite.expasy.org/prorule.html.

11. Consortium, UniProt. uniprot evidence. [Online] 06 04 2021. [Cited: 27 05 2021.] https://www.uniprot.org/help/evidences#ECO:0000250.

12. Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D. and Higgins D.G. *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* [Mol. Syst. Biol] 2011. Vol. 7.

13. scikit-learn. [Online] [Cited: 06 05 2021.] https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering.

14. uniprot Q9DBL9. [Online] https://www.uniprot.org/uniprot/Q9DBL9.

15. uniprot Q9QYR9. [Online] https://www.uniprot.org/uniprot/Q9QYR9.

16. PFAM. [Online] https://pfam.xfam.org/.

17. Mistry J., Chuguransky S., Williams L., Qureshi M., Salazar G., Sonnhammer E., Tosatto S., Paladin L., Raj S., Richardson L, Finn R., Bateman A. *Pfam: The protein families database in 2021.* [Nucleic Acids Res.] 2021. Vol. 49.

18. Expasy enzyme. [Online] https://enzyme.expasy.org/EC/.

19. gene ontology. [Online] http://geneontology.org/.

20. Ashburner M., Ball C., Blake J., Botstein D., Butler H., Cherry M., Davis A., Dolinski K., Dwight S., Eppig J., Harris M., Hill D., Issel-Tarver L., Kasarskis A., Lewi S. *Gene Ontology: tool for the unification of biology.* [nature genetics] 2000. Vol. 25.

21. Expasy Prosite. [Online] [Cited: 20 05 2021.] https://prosite.expasy.org/prosite_details.html.

22. hydrolase atlas. [Online] http://www.enzyme.chem.msu.ru/hcs/classes.html.

23. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. *Scikit-learn: Machine Learning in Python.* [JMLR] 2011. Vol. 12.

24. uniprot Q91WG0. [Online] https://www.uniprot.org/uniprot/Q91WG0.

25. uniprot Q8K4H1. [Online] https://www.uniprot.org/uniprot/Q8K4H1.

26. uniprot Q8VCT4. [Online] https://www.uniprot.org/uniprot/Q8VCT4.

27. uniprot A2AKK5. [Online] https://www.uniprot.org/uniprot/A2AKK5.

Figures

Additional Data

Table 8: setting used for the sequence alignment with Clustal w

parameter	value
Dealign input sequences	no
MBED-Like clustering guide-tree	Yes
MBED-Like clustering iteration	Yes
Number of combined iterations	Default(0)
Max guide tree iterations	Default
Max hmm iterations	Default
Order	aligned

Table 9: Settings for homology modelling on the Catalophore platform

parameter	value
Priority	Medium
Number of PSI-BLAST iterations	6
The maximum PSI-BLAST Evalue	0.5
Maximum number of templates to consider	5
Maximum number of ambiguous alignments to consider	5
Maximum oligomerization state	4
Maximum number of terminal loop residues	10
Use re-refined templates from PDB-Redo additionally	No
Homology model accuracy	Fast
Number of samples to try per loop	50
Use structure-based template profiles from RCSB	No

Table 10: Cavity procreation settings on the Catalophore platform

parameter	value
Run CavFind	True
Min. cavity volume [A^3]	50
Max. cavity volume [A^3]	2500
Ligsite cut-off	5
Probe radius [A]	1.4
Grid spacing	0.375
Softshell [Grid Units]	0.5
Remove hydrogens	False
Remove hetgroups	False
Remove waters	False
Keep all alternates	False

Table 11: Distances between residues to for a valid motif. The whole list can be found in the code under new_triad_dict or in the hydrolase atlas (22) where the motifs are stored except the three mentioned here. Distances are listed as in the upper triangle of a distance matrix.

motif	residues	distances in A
Murine soluble epoxide hydrolase 1EK1	ASP, HIS, ASP	3.9, 7.8, 4.6
Triad based on homology models	SER, HIS, ASP	4.5, 8.0, 5.5
Triad based on homology models with GLU	SER, HIS, GLU	4.5, 8.0, 5.5

Table 12: Amino acids and their catalytically active atoms

amino acid	catalytically important atom(s)
ALA	CB
CYS	SG
ASP	OD2, OD1
GLU	OE1, OE2
PHE	CG ,CD1 ,CE1, CD2, CE2, CZ
GLY	CA
HIS	CG, ND1, CE1, NE2 ,CD2
ILE	CB, CG1, CD1, CG2
LYS	NZ
LEU	CB, CG, CD1, CD2
MET	SD, CE
ASN	OD1, ND2
PRO	N, CA, CB, CG, CD
GLN	OE1, NE2
ARG	NE, CZ, NH1, NH2
SER	OG
THR	OG1
VAL	CB, CG1, CG2
TRP	CE3, CZ3, CH2, CZ2, CE2, NE1, CD1, CG, CD2
TYR	OH

Table 13: Cavity matching parameters

Parameter	value
Maximum tries per entry	5
Query point cloud UUID	
ICP iterations	100
Timeout per match(s)	300
Priority	1
Run LSQMAN	
Cavity shape	0.01
Aromatic carbon point-cloud	1
Carbon point-cloud	1
Hydrogen-bond donor point-cloud	1
Non-hydrogen bonding nitrogens point-cloud	1
Nitrogen as H-bond acceptor point-cloud	1
Oxygen as H-bond acceptor point-cloud	1
Sulfur as H-bond acceptor point-cloud	1
Phosphor point-cloud	
Desolvation point-cloud	0.1
Accessibility point-cloud	
Electrostatics point-cloud	0.1
Hydrophobicity point-cloud	
Flexibility point-cloud	
Chains point-cloud	
Sulfur point-cloud	
Bromine point-cloud	
Chlorine point-cloud	
Fluorine point-cloud	
Iodine point-cloud	
To be found out	

Code

For the code to run, the following libraries are needed:

```
urllib.request (urllib3 version 1.26.2)

BeautifulSoup (beautifulsoup4 version 4.9.3)

numpy (numpy version 1.20.1)
```