Horizon 2020 Work programme

Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy

Call

H2020-FNR-2020: Food and Natural Resources

Topic name

FNR-16-2020: ENZYMES FOR MORE ENVIRONMENT-FRIENDLY CONSUMER PRODUCTS

FuturEnzyme:

Technologies of the Future for Low-Cost Enzymes for Environment-Friendly Products

30/05/2022

SET OF 1,000 ENZYMES SELECTED USING MOTIF SCREENS

D2.3

VÍCTOR GUALLAR
BSC
Jordi Girona 31, BARCELONA 08034,Spain

Document information sheet

Work package:	WP2, Machine learning enzyme bioprospecting integrated into an industrial context
Authors:	CSIC (Manuel Ferrer, Patricia Molina)
Document version:	1
Date:	30/05/2022
Starting date:	01/06/2021
Duration:	48 months
Lead beneficiary:	CSIC
Participant(s):	CSIC, BSC, Bangor, UHAM, UDUS
Dissemination Level:	Confidential, only for consortium's members (including the Commission Services)
Type	Other
Due date (months)	12
Contact details:	Manuel Ferrer, mferrer@icp.csic.es

Summary

SET OF 1,000 ENZYMES SELECTED USING MOTIF SCREENS

1. Scope of Deliverable

This deliverable consists in a fasta file containing 1,000 full-length candidate sequences encoding enzymes with high probability to fulfil manufacturers' specifications. These sequences are selected in Task 2.3 by machine learning techniques applied to the 250,000 full-length candidate sequences delivered in D2.2. The fasta file will be deposited in the FuturEnzyme internal repository with a report detailing the selection procedure, as well as the annotations and details of each sequence.

2. Introduction & Methodology

2.1. Source and profiling of enzymes

We have established and manually curated a database with 37,403 taxonomically diverse protein sequences (Annex, **Table 1**) featuring the key enzyme families, potentially targeting enzymes relevant to the detergent, textile and cosmetic sectors. The sequences are available in fasta files, one per each of the target enzymes. We downloaded and/or compiled about 670 million sequences from 12 public and internal metagenomes, and 48 genomes, 12 public and internal metagenomes (for details see Annex, **Table 2**). The sequences are available in fasta files, one per each of the sequence repositories. The sequences encoding enzymes relevant to FuturEnzyme were selected by the DIAMOND search tool, using the sequences in Annex, **Table 2**. For DIAMOND searches, default parameters were used (percent identity >60%; alignment length >70; e-value < $1e^{-5}$). A total of 3,153,537 sequences have been selected (Annex, **Table 1**), which are available in fasta files, one per each of the target enzymes.

2.2. Network analysis for selecting best candidates

For further selecting the priority enzymes, we have then taken all the selected sequences we performed blastp (default parameters) against, keeping only the alignments with a percentage of identity higher than 50%. With these results we built the identity percentage network. Then, we clustered the sequences using the MCL algorithm, implemented in the software of the same name (Markov Cluster Algorithm: Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584 (2002), using the parameter Inflation = 1.4). This method is widely used to obtain clusters in sequence networks. With the sequences of each cluster we performed a multiple alignment using ClustalW (default parameters), obtaining from it the consensus sequence and a list of reference sequences conforming each of the clusters. A total of 481 clusters, each containing enzymes that most likely do show similar properties, were identified (Annex, **Table 3** and **Figure 1**).

2.3. Constraint Network Analysis

For gaining insights into how the flexibility of enzymes is linked to their thermal stability (a property of interest for the FuturEnzyme project), we applied Constraint Network Analysis (CNA), a rigidity theory-based approach to analyse biomolecular statics. To improve the robustness and investigate the statistical uncertainty, for each of the enzyme input structures, we carried out CNA on ensembles of network topologies (ENT^MD) generated from molecular dynamics trajectories. We then predicted Tp, the phase transition temperature previously applied as a measure of structural stability of a protein, for each enzyme using a constraint dilution approach.

2.4 Computational pipeline for filtering the best candidates

We developed a pipeline to characterise different enzyme families, having their sequences as the only input to find which enzyme sequences could be potential candidates to fulfil manufacturers' specifications. First,

we checked whether the sequence contained the proper domain, the catalytic residues, whether it was patented, and its conservation (along with MSA) based on bioinformatic tools.

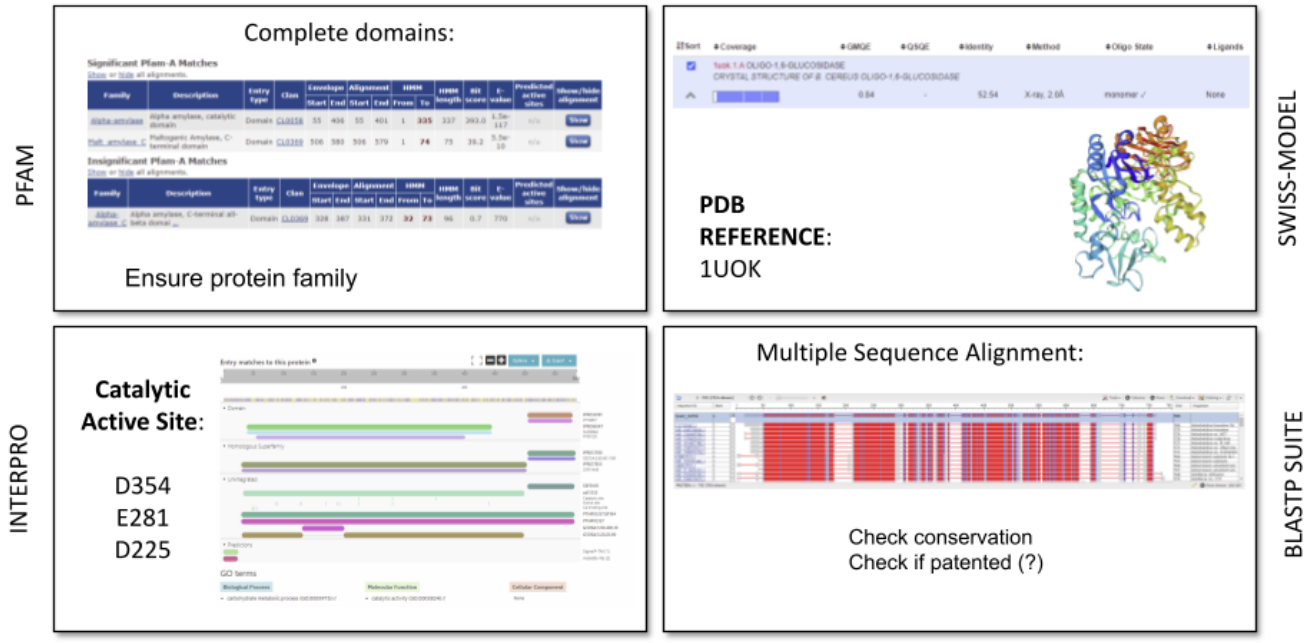


Figure 1. Illustration representing the softwares used to check the sequences with bioinformatic tools.

The sequences that passed this first filtering were modeled with AlphaFold 2.0 to obtain their 3D structure. Once the structure was obtained, substrates specified by the manufacturers' specifications were docked with the Glide software from the Schrödinger company in the active site of these enzymes. Subsequently, the substrate positioning around the active site was further explored with the software from BSC (Electronic and Atomic Protein Modelling group), Protein Energy Landscape Exploration (PELE). To account for the goodness of an enzyme-substrate interaction, we extract the measure of the catalytic events (those presenting catalytic-like distances) taking into account just the accepted Monte Carlo PELE steps "accepted catalytic events" or all (accepted and rejected) PELE steps "all catalytic events".

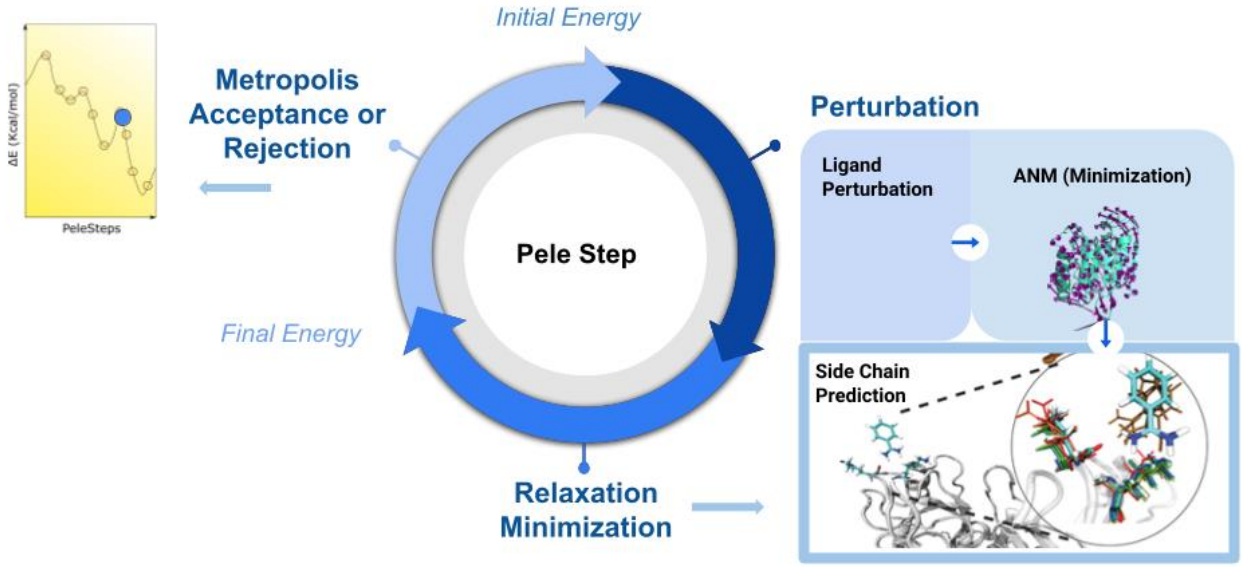


Figure 2. Scheme explaining the workflow of PELE's software.

The mentioned substrates were downloaded from the PubChem database, and their electrostatic point (ESP) charges were calculated from a quantum mechanics single point energy calculation with the Jaguar software from the Schrödinger company. These ESP charges were used in the mentioned induced-fit PELE simulations to have a higher precision in predicting the catalytic binding of the substrate in the active site of the enzyme.

Table 1. Summary workflow.

Protein Sequence Characterization	Structure Generation	Substrates Preparation and Docking	Comparative Catalytic Efficient
PFAM and INTERPRO databases	Structure Models with Alphafold 2.0 And Visual Inspection	Single Point Energy Calculation (Jaguar)	PELE Induced Fit Exhaustive
Complete sequence Protein Family Domains BLASTP SUITE Homology Sequences with desired properties	• Model 1.1.1 (Jumper et al. (2021)) • Predicted TM-score (pTM) Preset: full databases Catalytic Active Site Binding Residues Other Structure Features	• Basis set: CC-PVTZ • Theory: DFT (default) • Accuracy: Accurate • Maximum iterations: 100 • Properties: ESP charges • Solvation: solvent model PBF, Water Glide Docking (100 iterations, XP)	• CPUs 128 • 10 Iterations • 100 PeleSteps • SideChain Res. 10 • BoxRadius (25-35 A) • Peleffy: true Best Candidates Selection

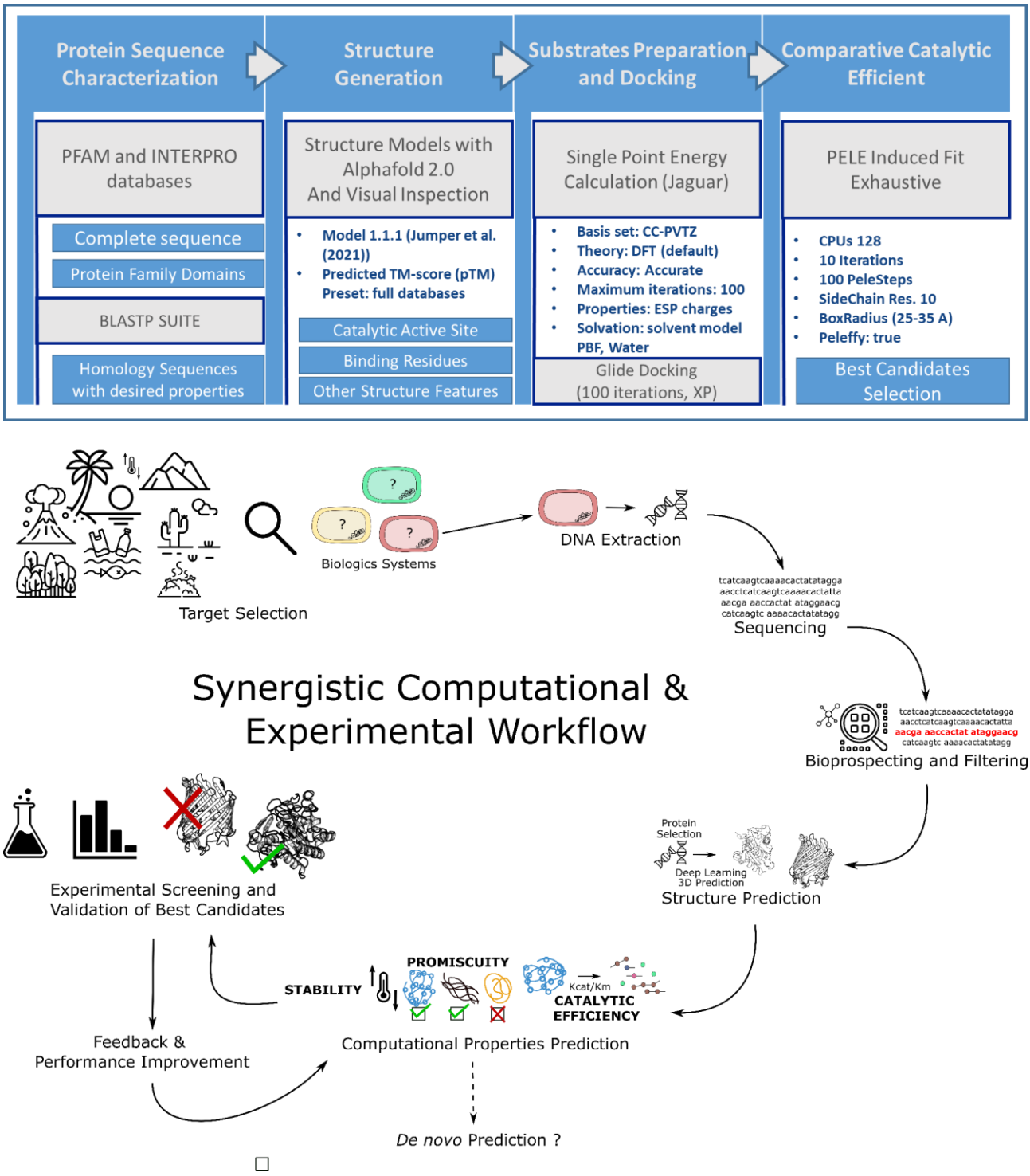


Figure 3. Experimental and computational workflow to search for new enzymes.

3. Results

3.1 In silico screen for enzymes of interest

We used a reference manually curated database with 37,403 diverse protein sequences (Annex, **Table 1**; Annex, **File 1**) featuring enzyme families relevant to the project to screen a total of 670 million sequences from 12 public and internal metagenomes, and 48 genomes, 12 public and internal metagenomes (Annex, **Table 2**; Annex, **File 2**). A total 3,153,537 sequences were selected (Annex, **Table 1**), which are available in fasta files, one per each of the target enzymes. Network analysis further revealed that they grouped into 481 clusters, each containing enzymes that most likely do show similar properties (Annex, **Table 1**; Annex, **Figure 1**; Annex, **File 3**).

After the filtering of sequences, a well-defined excel file with the information of 108 selected sequences was sent to BSC to perform the computational pipeline summarized in **Table 1** for the different needs in the project.

In the case of Evonik's needs, they want an enzyme that is able to efficiently generate hyaluronate of specific molecular weight (between 1-2 kDa). This enzyme should be either from the following EC numbers; 3.2.1.35, 3.2.1.36, 3.2.1.166, and 4.2.2.1. These include two types of big enzyme families: hydrolases (EC 3*) and lyases (EC 4*). Each type of hyaluronate degrading enzyme has its own catalytic residues and catalytic mechanism. Thus, we considered this notion when counting the number of catalytic poses in the PELE simulations.

In the case of both 3.2.1.36 and 3.2.1.166 enzyme sequences, the used substrate was a trimer of the hyaluronate molecule (focusing on the β-(1→3) glycosidic bond, which is the one that these enzymes break. One of the sequences stood out above the rest, which is the one that is closest to being a 3.2.1.36 classified enzyme. In contrast, the other sequences have closer homologs that belong to the 3.2.1.166 enzyme family. The problem is that this enzyme family defines the heparanases enzymes. Thus, they are specific towards heparan sulfate with a promiscuous (residual) activity towards hyaluronate due to the similarities in chemical motifs between both polymers (although heparan sulfate contains 2 to 3 more sulfate groups per disaccharide unit).

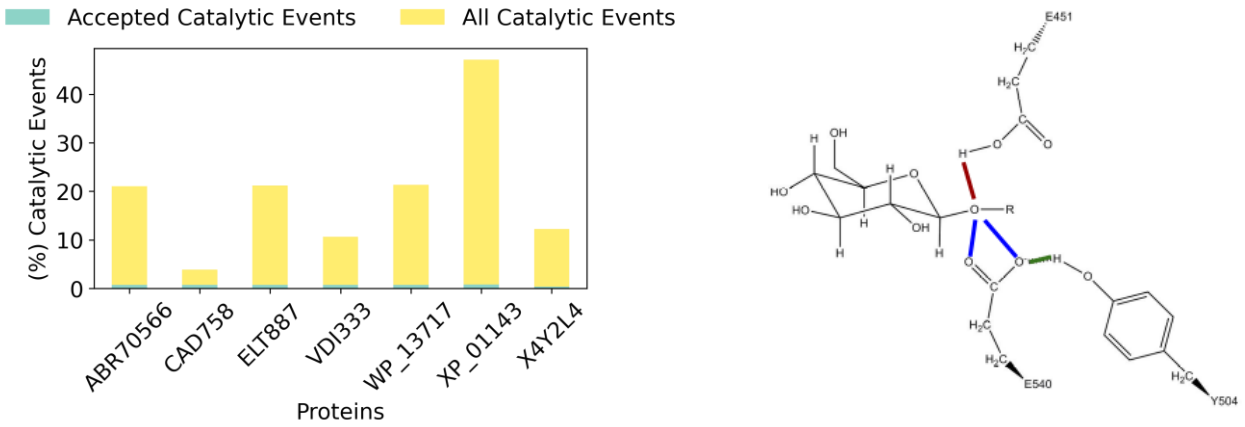


Figure 4. Plot showing the number of catalytic events in the 3.2.1.36/166 hyaluronidases compared to a control from UniProt entry; X4Y2L4 (left). Catalytic residues and the catalytic distances of 3.2.1.36/166 hyaluronidases highlighted (right).

Regarding 4.2.2.1 enzyme sequences, the used substrate was a hexamer of the hyaluronate molecule (since the active site's cavity is bigger compared to 3.2.1.36/166 enzymes). None of the sequences shined over the others. Only WP_070668766 showed promising results, but it was not a 4.2.2.1 enzyme nor a 3.2.1.36/166 one. This enzyme sequence belongs to the glycoside hydrolase family 16 and should be labelled as a 3.2.1.39 enzyme sequence. Thus, it is a hydrolase and has the typical catalytic dyad constituted by 2 Glu residues.

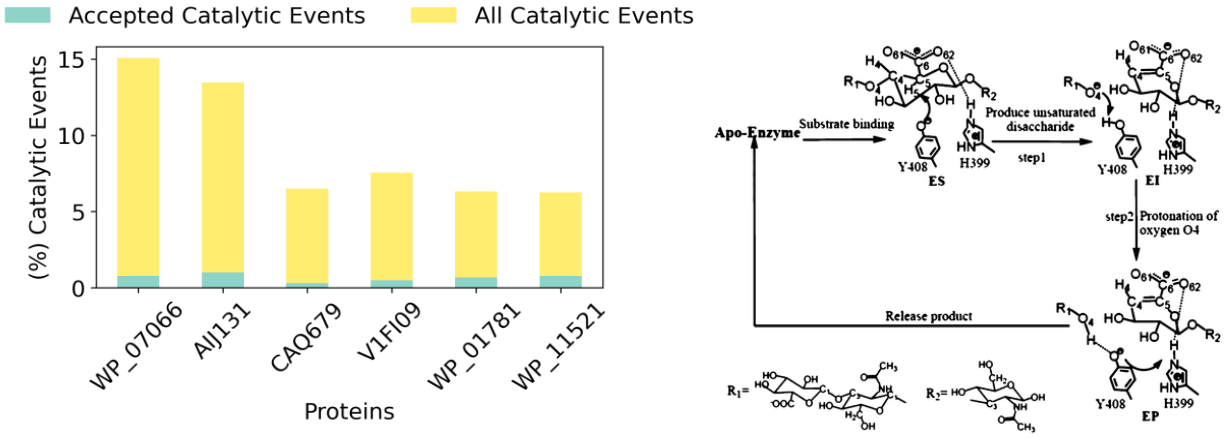


Figure 5. Plot showing the number of catalytic events in the 4.2.2.1 hyaluronate lyases. Catalytic residues and the catalytic mechanism of 4.2.2.1 hyaluronidases (right). Image taken from https://pubs.acs.org/doi/10.1021/jp406206s.

In the case of the needs in the textile industry, requested by Schoeller, they wanted several enzymes involved in different processes. The high priority demands were the cleaning/pretreatment of synthetic fibers process, which needs cutinases, polyurethanases and amidases; the problem of the chalk marks, which needs lipases, esterases, polyurethanases, amidases and cellulases; the solvent cleaning process, which needs lipases, cutinases, polyurethanases, amidases and proteases; the higher amounts of chemicals problem, which needs lipases, cutinases, polyurethanases, amidases, and proteases; and the fewer water consumption in the dyeing process, which needs lipases, cutinases and cellulases.

In summary, the requested enzymes are amidases (E.C.: 3.5.2.12), esterases (E.C.: 3.1), polyurethanases (E.C.: 3.1.1.3), cutinases (E.C.: 3.1.1.74), proteases (E.C.: 3.4) and cellulases (E.C.: 3.2.1.4). Esterases, polyurethanases, and cutinases require a catalytic triad formed by serine, histidine, and aspartic acid. Amidases work with a serine, serine, and lysine triad. Cellulases work with two acidic residues like aspartic acid or glutamic acid. Finally, there are different mechanisms for the proteases: serine, histidine, and aspartic acid triad; cysteine, histidine, and asparagine triad; and metalloproteases with a Zn^{2+} as the main catalytic element.

The substrates used for the computational simulations with PELE were: polyurethane dimer, MHET (**Figure 6),** several ester polymers like PLA, PCL, and aliphatic polyurethane, and two types of proteins: 6-units of nylon and 7-units of polyglycine. The proteases request is shared with the detergent needs.

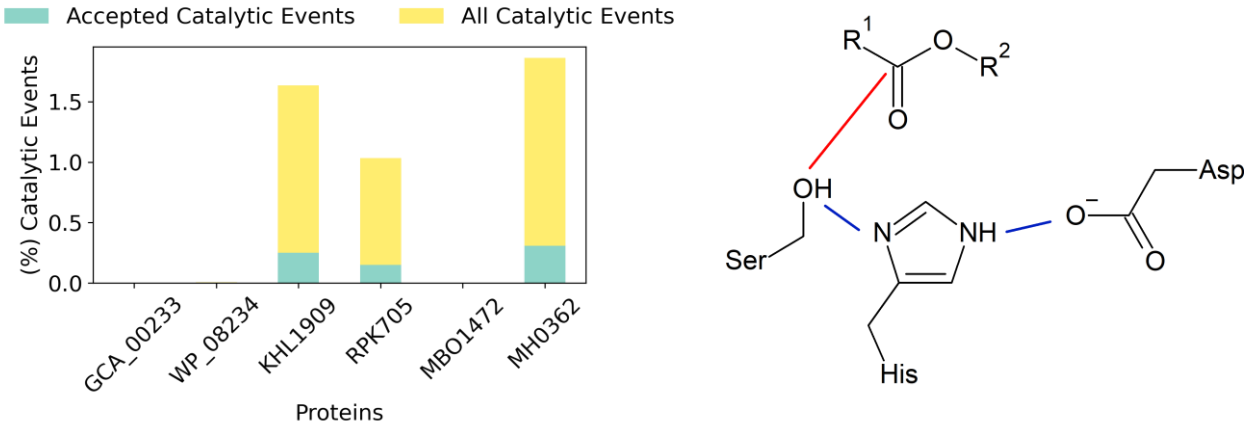


Figure 6. Accepted and total catalytic events for the 6 selected MHETases (left). In 3 of them the ligand never reaches catalytic positions. Catalytic mechanism for esterases (right).

In the detergent industry, the priority targets are enzymes for removing specific fatty oil stains, which are mainly true lipases (E.C: 3.1.1.3). Other relevant enzymes which have been considered are proteases/peptidases (E.C: 3.4) and amylases (E.C: 3.2.1.1). The catalytic mechanism of lipases involves a catalytic triad formed by serine, histidine and aspartic/glutamic acid residue. Histidine activates serine through general base catalysis to deprotonate serine, which transforms it into a nucleophile with the ability to attack the ester bond of triacylglycerides. Histidine donates a proton to the leaving group and then activates a water molecule to allow the hydrolysis of the intermediate. The acid residue, which can be an aspartic acid or glutamic acid residue, activates the histidine residue. Alpha-amylase catalyses the hydrolysis of internal alpha-glycosidic linkages in starch. The chemical reaction involves two aspartic acid residues and a glutamic acid. A nucleophilic aspartic acid side chain attacks the sugar anomeric center assisted by acid catalysis of glutamic acid and aspartic acid. Finally, proteases are shared with the textile industry.

Simulation conditions are 30°C (range 20-40°C) and pH 7.75 (range 7.0-8.5) to accomplish the liquid detergent formulation conditions that Henkel specified. The substrates employed have been the triglyceride triolein (glycerol + three unsaturated oleic acid units) for lipases, a dimer and a tetramer of starch for alpha-amylases and two types of peptide substrates for proteases, 6-units of nylon and 7-units of polyglycine.

There are two types of lipases: with and without lid domain. Study of lid domain (**Figure 7**) movement using molecular motion algorithms software (MoMa loop sampling), that allows exhaustively sample protein loop conformations, has allowed the opening of the lid domain in most lipases which had the active site inaccessible for the substrate. These will be further analysed (**Figures 8-13**).

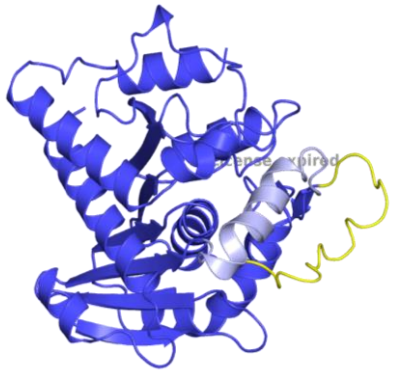


Figure 7. Lipase structure. The lid domain enclosing the catalytic active site is shown in grey, and the same lid domain in an open conformation is shown in yellow.

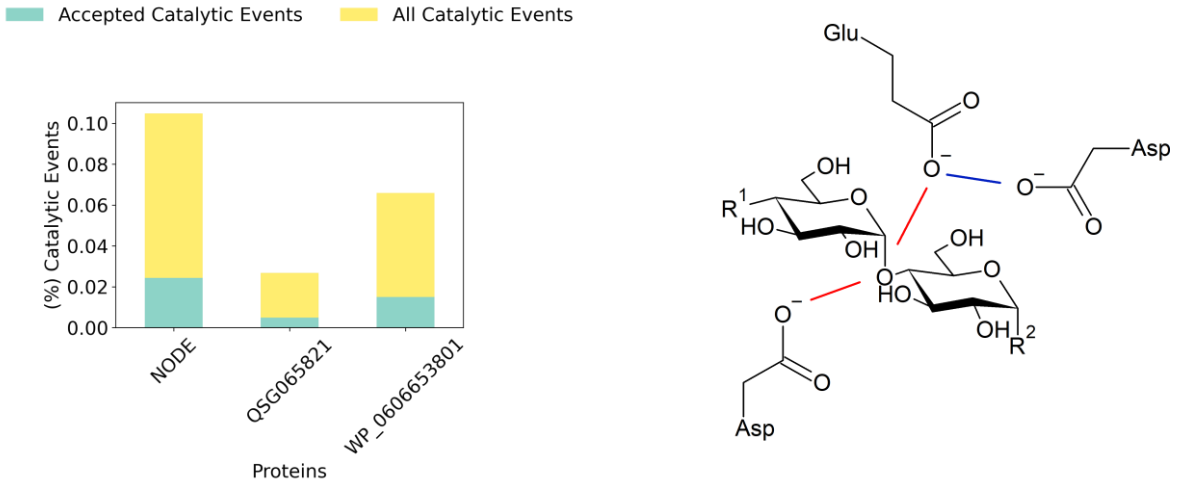
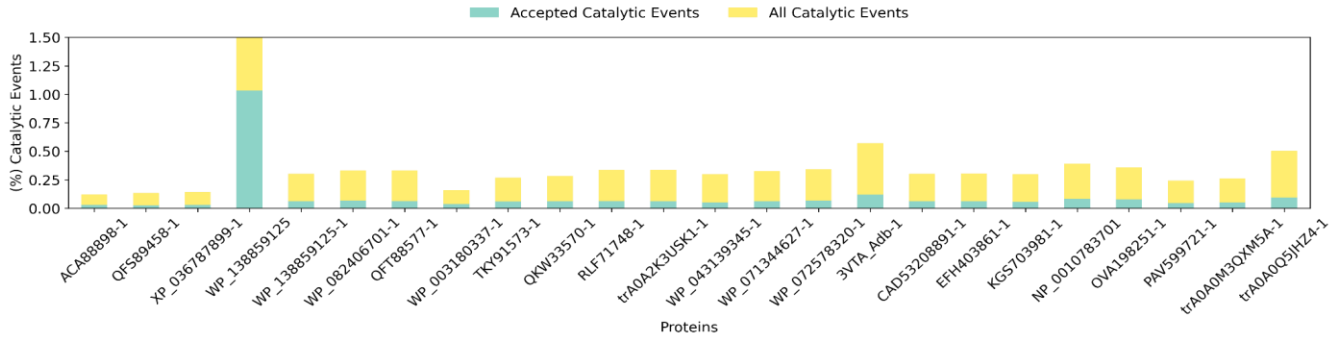
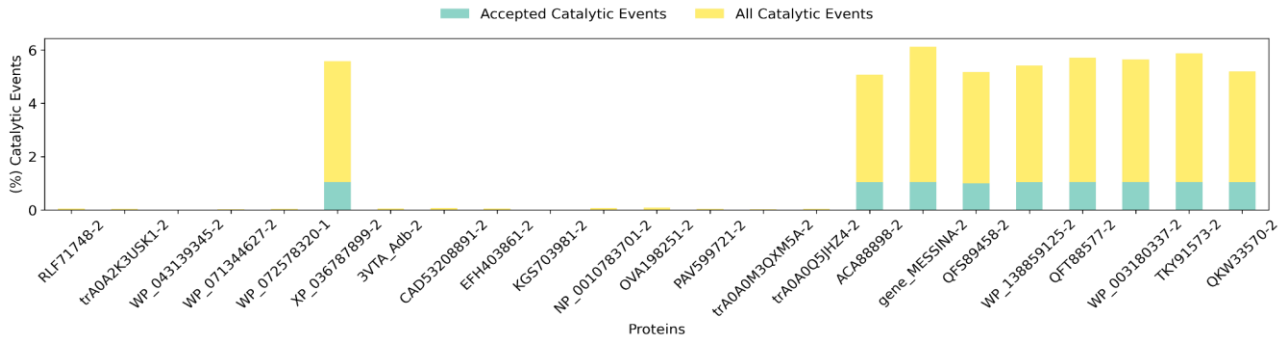


Figure 8. Accepted and all catalytic events for the selected amylases (left). Catalytic mechanism for these enzymes (right).



Nylon -6



PRG

Figure 9. Accepted and all catalytic events for the selected serine proteases against a 6-units nylon ligand (top) and a 7-units polyglycine (bottom).

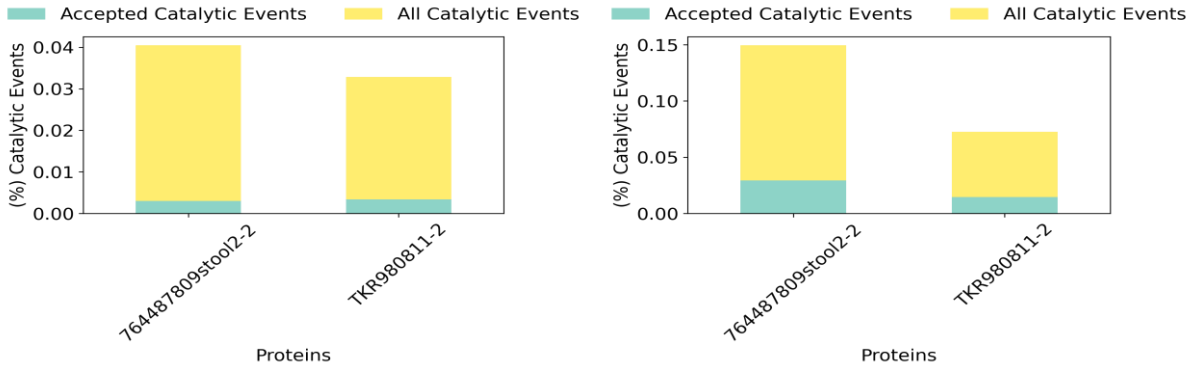


Figure 10. Accepted and all catalytic events for the selected cysteine proteases. From left to right, PELE simulations using 6-units nylon and 7-units polyglycine.

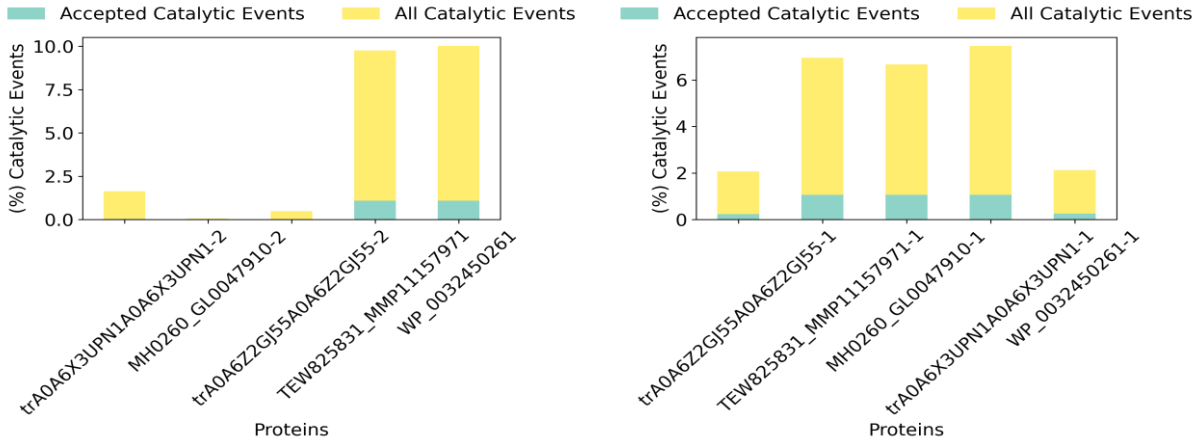


Figure 11. Accepted and all catalytic events for the zinc proteases. From left to right, PELE simulations using 6-units nylon and 7-units polyglycine.

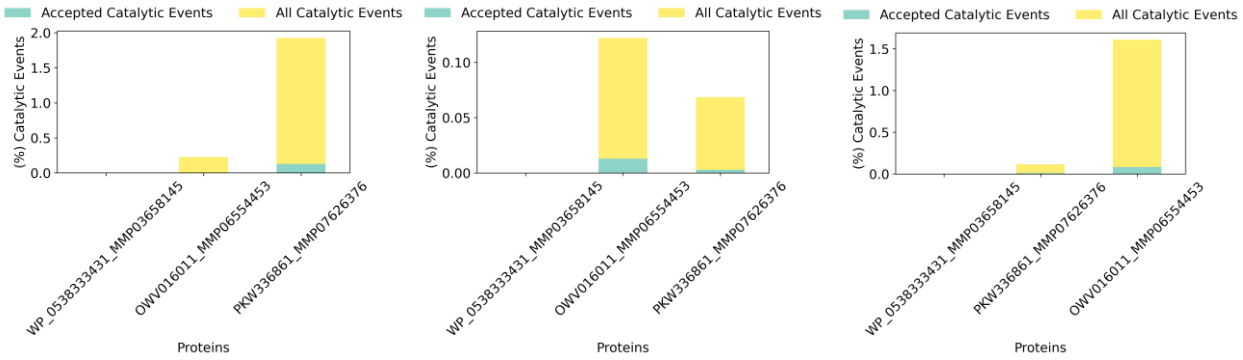


Figure 12. Accepted and all catalytic events for the polymer degrading enzymes. From left to right, the same proteins against polycaprolactone, polylactic acid, and aliphatic polyurethane.

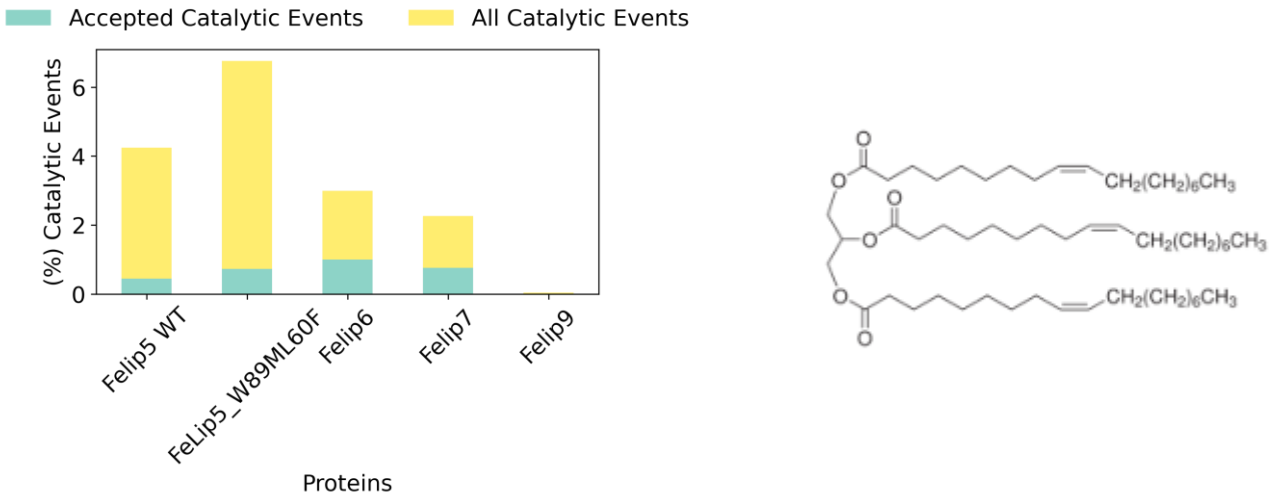


Figure 13. Accepted and all catalytic events for the selected lipases (left). Ligand used for the simulations: triolein (right).

4. Conclusions and outlook

For the simulations part, a lot of proteins have been tested with PELE, showing significant differences between them. A particular case, some lipases have not been simulated because its processing is a lot more difficult, since the lid domain is enclosing the catalytic site. Recently, a new server has been released, which helps opening the lid domains in proteins like in our case, and once all the remaining lipases have the lid opened, we will perform PELE simulations with them.

An experimental validation correlating these results would reinforce the methodology, and further simulations adding mutants and creating PluriZymes will be done (in WP5).

Once the first active enzymes from each class have been identified, their sequences will be handed over to UHAM to develop highly sensitive Hidden Markov Models (HMMs) with their AHATool pipeline, also developed within the frame of WP2. This tool automatises the processes of sequence alignments and HMM construction, *in silico* database screening and gathering of useful information for candidate selection, such as secretion signals or taxonomical origin of the hits. The constructed models will detect active enzymes with a higher success, since all the sequences used to build the models have been tested active previously. Thus, the process of expanding the diversity of active enzymes in the collection will be fast and efficient.

ANNEX

Annex File 1_FuturEnzyme Reference Sequences_to_do_BLAST

In-house database containing sequences encoding enzymes relevant to detergent, cosmetic and textile sectors. The sequences include those retrieved from bibliographic and patent search as well as one relevant sequence per taxonomic group. Because its extensive size, the file is available at the private area FuturEnzyme's web:

https://www.futurenzyme.eu/login/private-area/shared-data/

Annex File 2_Diamond_Results

Sequences encoding enzymes potentially relevant to detergent, cosmetic and textile sectors. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin. Because its extensive size, the file is available at the private area FuturEnzyme's web:

https://www.futurenzyme.eu/login/private-area/shared-data/

Annex File 3_Network Analysis Enzymes

Sequences encoding enzymes constituting each of the networks identified per enzyme family. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin. Because its extensive size the file is available at the private area FuturEnzyme's web:

https://www.futurenzyme.eu/login/private-area/shared-data/

Annex File 4_European_Project_Selected Enzymes_Table

Sequences encoding enzymes selected for computational analysis and gene synthesis Because its extensive size the file is available at the private area FuturEnzyme's web:

https://www.futurenzyme.eu/login/private-area/shared-data/

Annex **Table 1**. List of selected BLAST-hit candidates per each of the reference enzyme classes.

# Public databases	Sequences in the reference fasta	Sequences identified by BLAST
Amidase (FuturEnzyme - textile).fas	1	194
Amylase (COG0366 - FuturEnzyme - detergent).fas	21092	1048575
Amylase (EC3.2.1.1 - FuturEnzyme - detergent).fas	4	679
Cutinases (EC3.1.1.74 - FuturEnzyme - detergent & textile).fas	2572	255991
Cutinases (pfam01083 - FuturEnzyme - detergent & textile).fas	19	2175
Heparanase (EC 3.2.1.166 - FuturEnzyme - cosmetic).fas	4	386
Hyaluronate lyase (cd01083 - EC4.2.2.1 - FuturEnzyme - cosmetic).fas	355	41852
Hyaluronidase (EC3.2.1.36 - FuturEnzyme - cosmetic).fas	2	95
Hyaluronidase (EC4.2.2.1-FuturEnzyme - cosmetic).fas	292	36725
Hyaluronidase (pfam03662 - FuturEnzyme - cosmetic).fas	65	6701
Hyaluronidases (EC3.2.1.35 - FuturEnzyme - cosmetic).fas	4317	380042
Hyaluronidases (pfam01630 - FuturEnzyme - cosmetic).fas	5	2219
Lactonase (COG1735 - FuturEnzyme - detergent).fas	1069	119142
Lactonase (EC3.1.1.25 - FuturEnzyme - detergent).fas	24	2682
Lipase-Esterase (FuturEnzyme - detergent).fas	76	546
Mono(ethylene terephthalate) hydrolases (EC 3.1.1.102 - FuturEnzyme - detergent & textile).fas	70	824
Peptidase type Bromelain (EC3.4.22.32 - FuturEnzyme - textile).fas	2	179
Peptidase type family M04 (FuturEnzyme - detergent & textile).fas	225	32971
Peptidase type family S08 (alcalase - FuturEnzyme - detergent & textile).fas	1116	199971
Peptidase type Papain (EC3.4.22.2 - FuturEnzyme - detergent & textile).fas	41	5459
Peptidase type savinase-esperase (EC3.4.21.14 - FuturEnzyme - detergent & textile).fas	8	1515
Peptidase type subtilisin-alcalase (EC3.4.21.62 - FuturEnzyme - detergent & textile).fas	4703	804058
Peroxidases (FuturEnzyme - detergent).fas	159	16189
PLA, PCL, Impranil DNL hydrolases (FuturEnzyme - detergent & textile).fas	26	3022

Poly(ethylene terephthalate) hydrolases (FuturEnzyme - detergent & textile).fas	38	4615
Polyurethanase (1) (FuturEnzyme - detergent & textile).fas	50	4605
Polyurethanase (2) -lipase class 3 (FuturEnzyme - detergent & textile).fas	370	28415
Polyurethane degrading urease (EC3.5.1.5 - FuturEnzyme - textiles).fas	828	152894
Trypsin and protease inhibitor (FuturEnzyme - detergent).fas	3	136
TOTAL	37403	3152857

*For DIAMOND-BLASTP searches default parameters were used (percent identity ≥60%; alignment length ≥70; e-value ≤ $1e^{-5}$)

Annex **Table 2**. List of public and internal sequence repositories and genomes screened.

# Public databases*	Details	CDS
CAZyDB.07312020.dmnd`	http://bcb.unl.edu/dbCAN2/download/	1716043
mardb_proteins_V6.dmnd	https://public.sfb.uit.no/MarDB/; BLAST/proteins/mardb_proteins_V6.faa	46739080
marfunV3_proteins.dmnd	https://public.sfb.uit.no/MarFun/; BLAST/proteins/marfunV3_proteins.faa	71374
marref_proteins_V6.dmnd	https://public.sfb.uit.no/MarRef/; BLAST/proteins/marref_proteins_V6.faa	4726614
nr.dmnd	ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz	371327556
uniprot_sprot.dmnd	https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz	564638
uniprot_trembl.dmnd	https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.fasta.gz	214406399
IGC.dmnd	-	9878647
MAGProts.dmnd	-	208832
totalProtsMetaProBone.dmnd	-	10402509
Irish and Mediterranean.dmnd	https://bangoroffice365-my.sharepoint.com/personal/chsa18_bangor_ac_uk/	449245
Human microbiome	https://db.cngb.org/microbiome/genecatalog/genecatalog_human/)	10000000
# Additional genomes*	**Details**	**CDS**
HF571520-HF571521	Halorhabdus tiamatea SARL4B	3023
JFHS00000000.1	Psebau_v14	7839
LGTE00000000.1	ASM126341v1	3097
NC_015151.1	ASM19031v1	2320
NZ_AROI00000000.1	Pseudomonas pelagia CL-AP6	4112
NZ_NWMT00000000.1		

NZ_FOGN01000016	Pseudomonas bauzanensis	3241
NZ_LT629748.1	Pseudomonas litoralis	3717
NZ_NBYK00000000.1	Pseudomonas aestusnigri	3510
NZ_PPSK00000000.1	Pseudomonas oceani	3757
PRJEB12275	Cuniculiplasma divulgatum, C. divulgatum PM4	1816
PRJEB12276	Cuniculiplasma divulgatum (ASM90008351v1)	2750
	Thermosinus carboxydivorans Nor1, ASM16915v1 (AAWL00000000.1)	
ABXP00000000.2	Caldanaerobacter subterraneus subsp. pacificus DSM 12653 (ASM15627v2)	2511
ATYG00000000.1	Carboxydothermus ferrireducens DSM 11255, ASM42756v1	2492
BDJL00000000.1	Carboxydothermus islandicus, ASM195032v1	2480
CP000141.1	Carboxydothermus hydrogenoformans Z-2901, ASM1286v1	2620
CP001463.1	Thermococcus sibiricus MM 739, ASM2254v1	2036
CP002952.1	Thermococcus sp. AM4, ASM15120v2	2222
CP003321.1	Desulfurococcus amylolyticus DSM 16532, ASM23101v3	1421
CP003423.1	Fervidicoccus fontis Kam940, ASM25842v1	1385
CP003531.1	Thermogladius calderae 1633, ASM26449v1	1414
CP003557.1	Melioribacter roseus P3M-2, ASM27914v1	2840
CP006646.1	Thermofilum adornatum, ASM44601v1)	1896
CP007493.1	Thermofilum adornatus 1505, ASM81324v1	1924
CP009552.1	Geoglobus acetivorans, ASM78925v1	2218
CP009961.1	Thermofilum uzonense, ASM99380v1	1455
CP013050.1	Thermococcus barophilus, ASM143345v1	2634
CP018099	Caldithrix abyssi DSM 13497, ASM188681v1	4214
GCA_001306115.1	Ornatilinea apprima, ASM130611v1	3347
CP028858.1	Haloarculaceae archaeon HArcel1, ASM305836v1	2532
LJCQ00000000.1	Acidiplasma aeolicum, ASM139969v1	1722
LKBG00000000.1	Acidiplasma aeolicum, ASM140294v1	1696
NC_008260.1	Alcanivorax borkumensis SK2, ASM936v1	2755

CP005996.1, CP006601.1 (plasmid)	Cycloclasticus zancles 78-ME, ASM44259v1	2584
CP008874.1, CP008875.1 (plasmid)	Halanaeroarchaeum sulfurireducens, ASM101111v1	2228
CP011564.1, CP011565.1 (plasmid)	Halanaeroarchaeum sulfurireducens, ASM130565v1	2270
CP016804.1	Halodesulfurarchaeum formicicum, ASM188695v1	2100
CP016070.1	Halodesulfurarchaeum formicicum, ASM176731v1	2023
CP044129.1, CP044130.1 (plasmid)	Halomicrobium sp. LC1Hm, ASM961799v1	3447
CP025066.1	Halalkaliarchaeum desulfuricum, ASM295277v1	3232
CP064789.1, CP064790.1 (plasmid)	Haloarculaceae archaeon HSR-Bgl, ASM1709444v1	3117
CP064791.1, CP064792.1 (plasmid)	Haloarculaceae archaeon HSR-Est, ASM1709446v1	2859
CP064787.1	Haloarculaceae archaeon HSR12-1, ASM1709450v1	3055
CP064788.1	Haloarculaceae archaeon HSR12-2, ASM1709452v1	3024
CP040089.1	DPANN group archaeon LC1Nh, ASM961797v1	1162
CP064786.1	Halobacteriaceae archaeon AArc-S, ASM1709448v1	3120
CP024047.1, CP024045.1 (pla1); CP024046.1 (pla2)	Natrarchaeobaculum sulfurireducens, ASM343082v1	3708
CP027033.1, CP027032.1 (plasmid)	Natrarchaeobaculum sulfurireducens, ASM343080v1	3737
TOTAL		670619599

* List of databases for BLASTP searches.

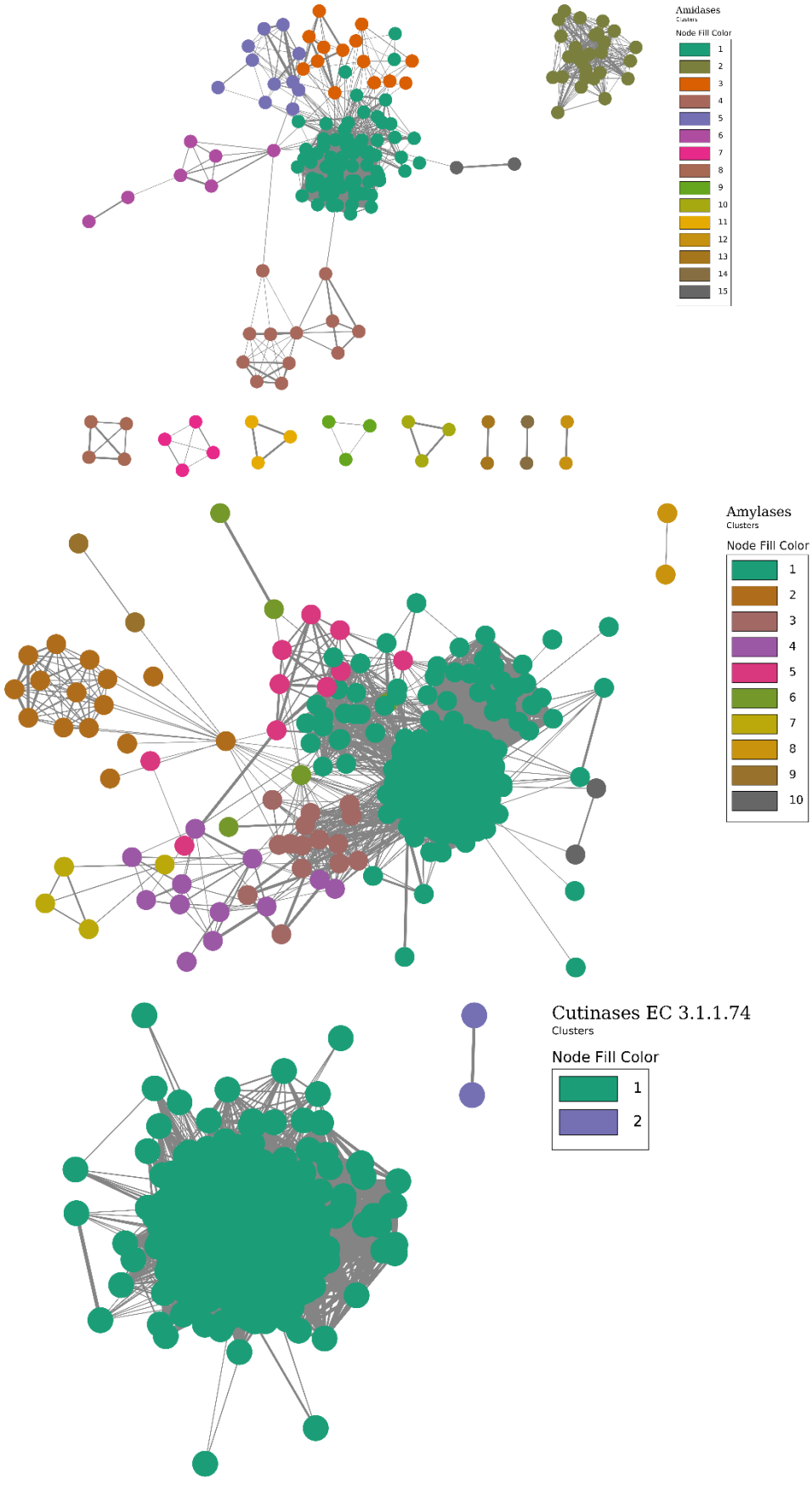
Annex **Table 3**. List of selected BLAST-hit candidates per each of the reference enzyme classes

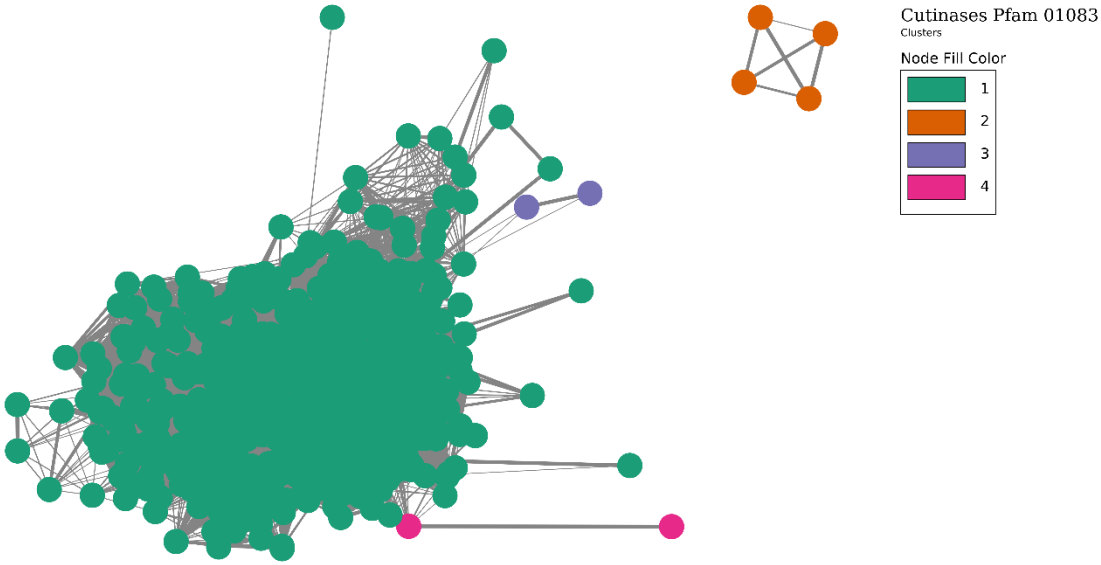
# Public databases	Sequences identified by BLAST	Nr of clusters
Amidase (FuturEnzyme - textile).fas	194	22
Amylase (COG0366 - FuturEnzyme - detergent).fas	1048575	21
Amylase (EC3.2.1.1 - FuturEnzyme - detergent).fas	679	
Cutinases (EC3.1.1.74 - FuturEnzyme - detergent & textile).fas	255991	3
Cutinases (pfam01083 - FuturEnzyme - detergent & textile).fas	2175	9
Heparanase (EC 3.2.1.166 - FuturEnzyme - cosmetic).fas	386	1
Hyaluronate lyase (cd01083 - EC4.2.2.1 - FuturEnzyme - cosmetic).fas	41852	87
Hyaluronidase (EC3.2.1.36 - FuturEnzyme - cosmetic).fas	95	22
Hyaluronidase (EC4.2.2.1-FuturEnzyme - cosmetic).fas	36725	38
Hyaluronidase (pfam03662 - FuturEnzyme - cosmetic).fas	6701	-
Hyaluronidases (EC3.2.1.35 - FuturEnzyme - cosmetic).fas	380042	14
Hyaluronidases (pfam01630 - FuturEnzyme - cosmetic).fas	2219	4
Lactonase (COG1735 - FuturEnzyme - detergent).fas	119142	-
Lactonase (EC3.1.1.25 - FuturEnzyme - detergent).fas	2682	-
Lipase-Esterase (FuturEnzyme - detergent).fas	680	112
Mono(ethylene terephthalate) hydrolases (EC 3.1.1.102 - FuturEnzyme - detergent & textile).fas	824	13
Peptidase type Bromelain (EC3.4.22.32 - FuturEnzyme - textile).fas	179	5
Peptidase type family M04 (FuturEnzyme - detergent & textile).fas	32971	8
Peptidase type family S08 (alcalase - FuturEnzyme - detergent & textile).fas	199971	55
Peptidase type Papain (EC3.4.22.2 - FuturEnzyme - detergent & textile).fas	5459	1
Peptidase type savinase-esperase (EC3.4.21.14 - FuturEnzyme - detergent & textile).fas	1515	34
Peptidase type subtilisin-alcalase (EC3.4.21.62 - FuturEnzyme - detergent & textile).fas	804058	7

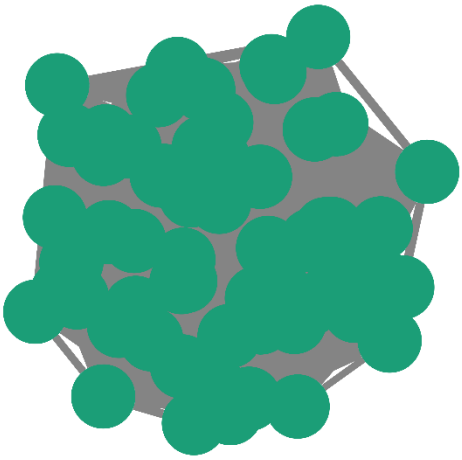
Peroxidases (FuturEnzyme - detergent).fas	16189	-
PLA, PCL, Impranil DNL hydrolases (FuturEnzyme - detergent & textile).fas	3022	19
Poly(ethylene terephthalate) hydrolases (FuturEnzyme - detergent & textile).fas	4615	4
Polyurethanase (1) (FuturEnzyme - detergent & textile).fas	4605	1
Polyurethanase (2) -lipase class 3 (FuturEnzyme - detergent & textile).fas	28415	1
Polyurethane degrading urease (EC3.5.1.5 - FuturEnzyme - textiles).fas	152894	-
Trypsin and protease inhibitor (FuturEnzyme - detergent).fas	136	-
TOTAL	3153537	

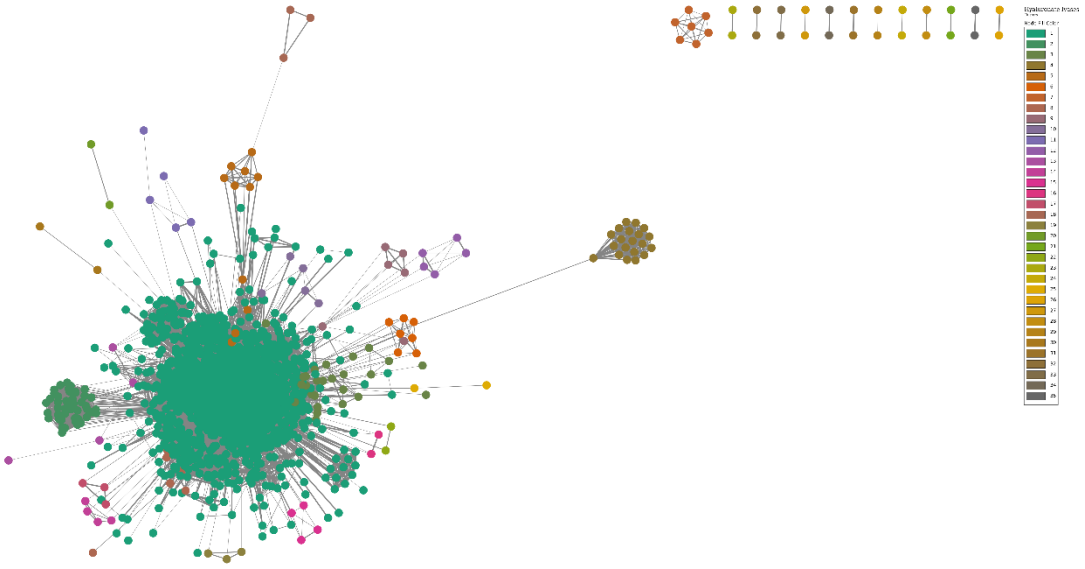
*For selection we have taken all the selected sequences we have blastp (default parameters) against all of them, keeping only the alignments with a percentage of identity higher than 50%. With these results we built the identity percentage network. Then we clustered the sequences using the MCL algorithm, implemented in the software of the same name (Markov Cluster Algorithm: Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584 (2002), using the parameter Inflation = 1.4). This method is widely used to obtain clusters in sequence networks. With the sequences of each cluster we performed a multiple alignment using ClustalW (default parameters), obtaining from it the consensus sequence and a list of reference sequences conforming each of the clusters.

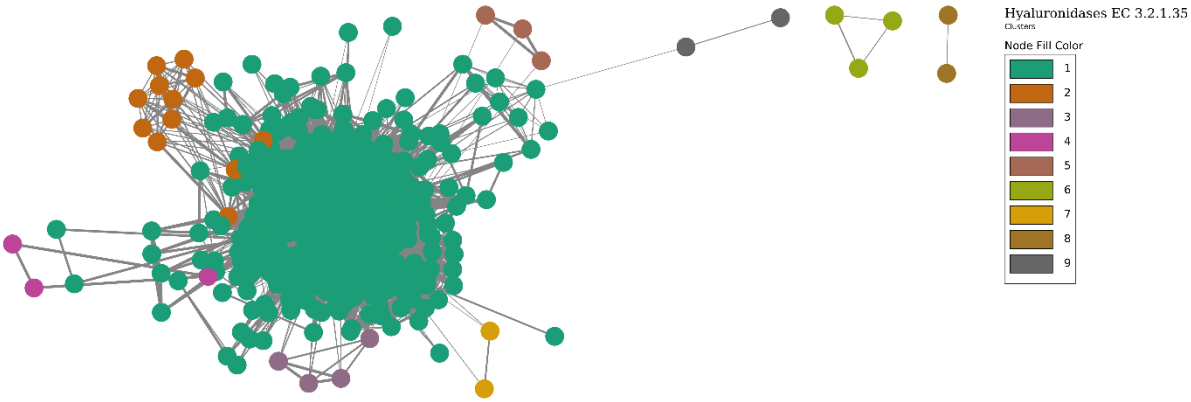
Annex **Figure 1**. Image representing the different clusters within enzyme classes

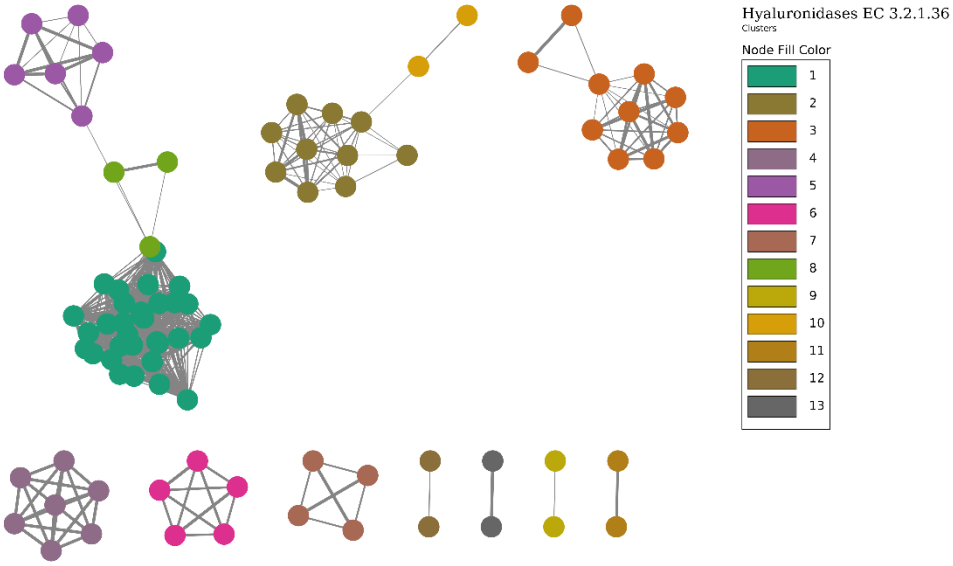


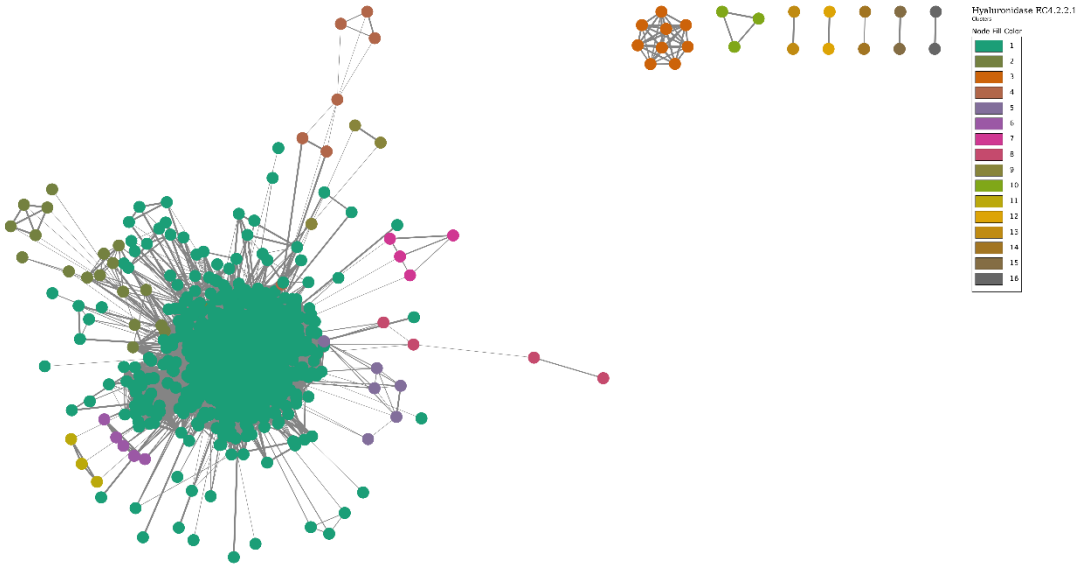

Cutinases Pfam 01083
Clusters
Node Fill Color
1
2
3
4

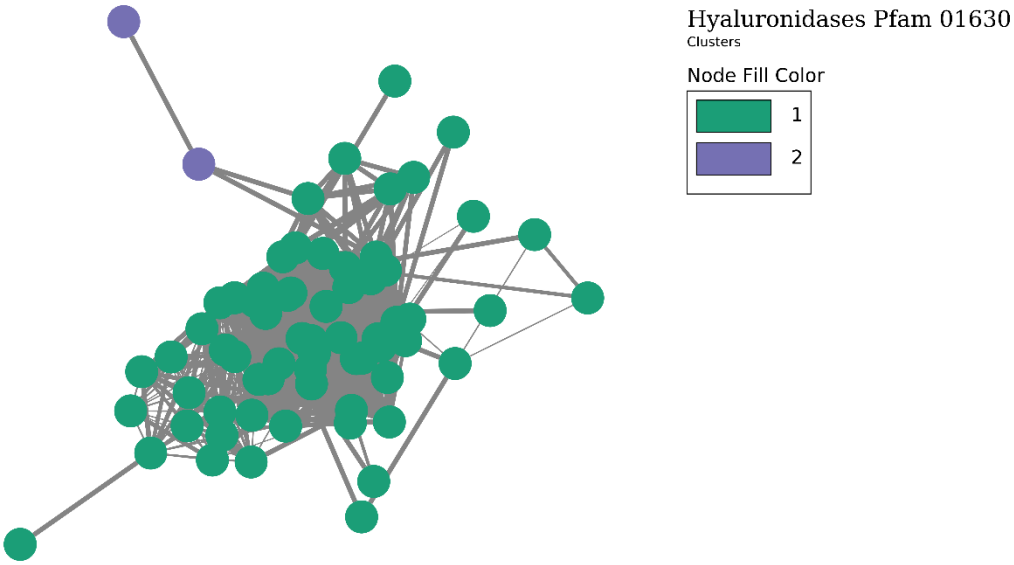

Heparanases
Clusters
Node Fill Color
1

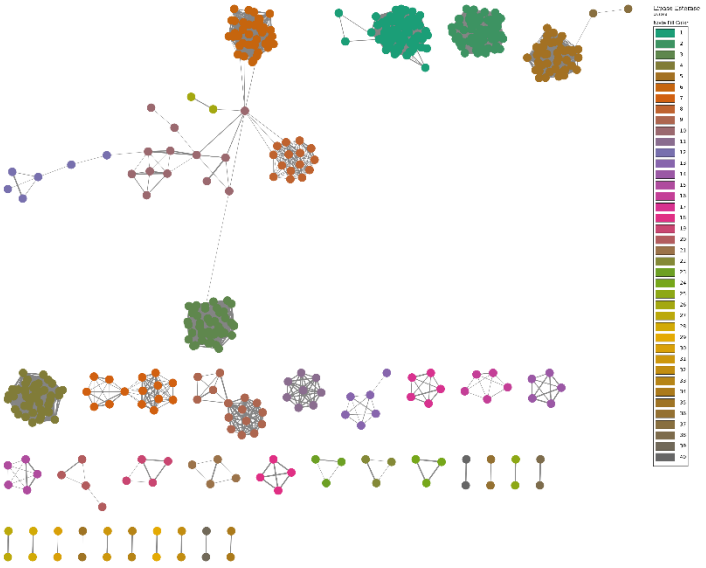


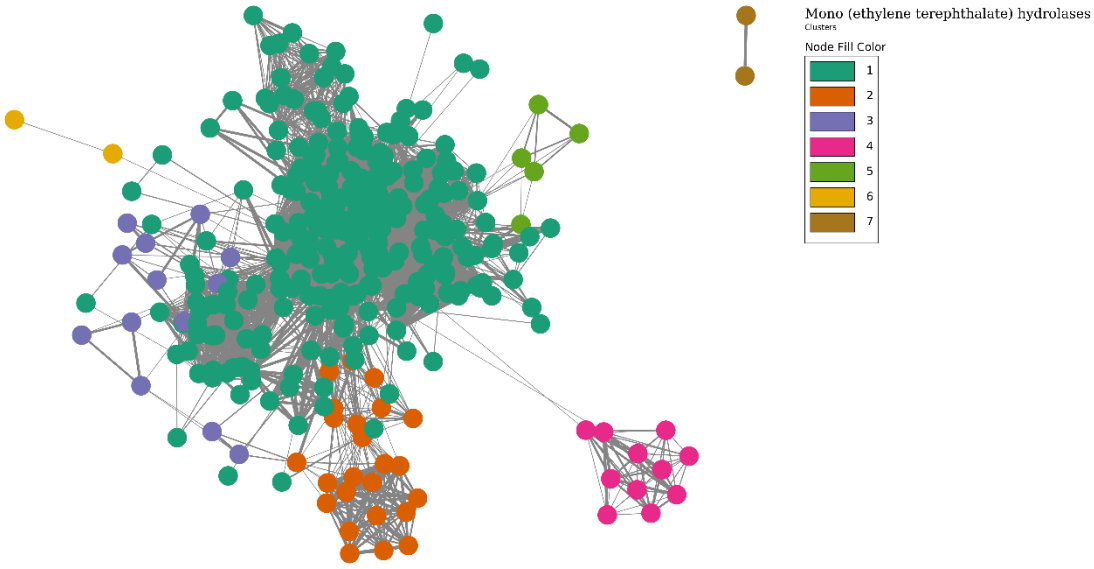

Hyaluronidases EC 3.2.1.35
Clusters
Node Fill Color
1
2
3
4
5
6
7
8
9

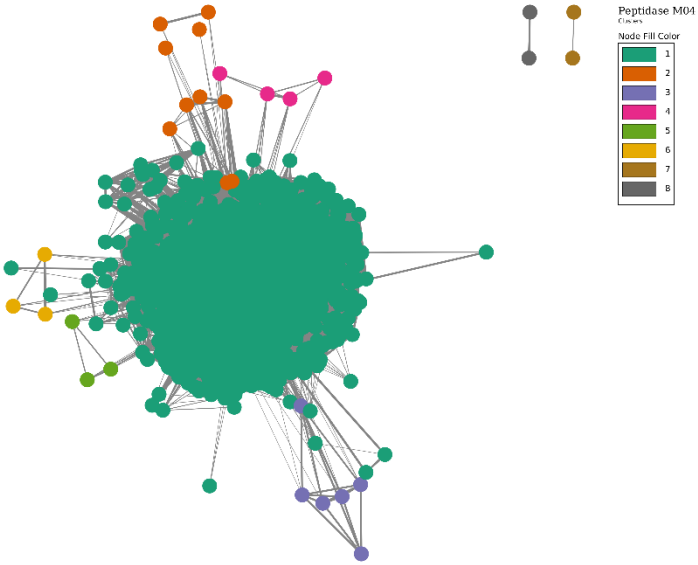

Hyaluronidases EC 3.2.1.36
Clusters
Node Fill Color
1
2
3
4
5
6
7
8
9
10
11
12
13

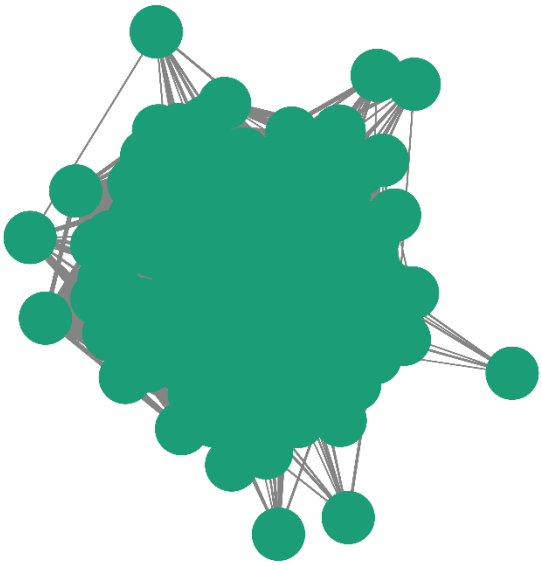

Hyaluronidase EC4.2.2.1
Clusters
Node Fill Color


Hyaluronidases Pfam 01630
Clusters
Node Fill Color
1
2




Mono (ethylene terephthalate) hydrolases
Clusters
Node Fill Color
1
2
3
4
5
6
7



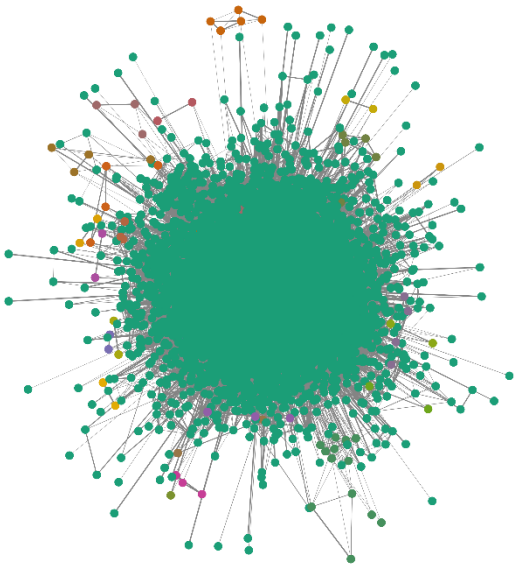


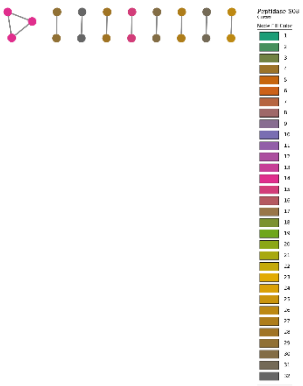
Peptidase type Papain

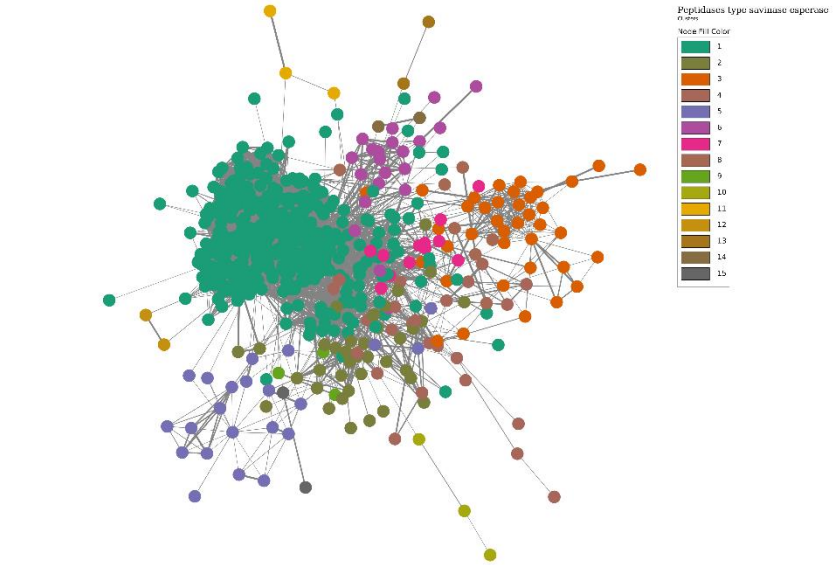
Clusters

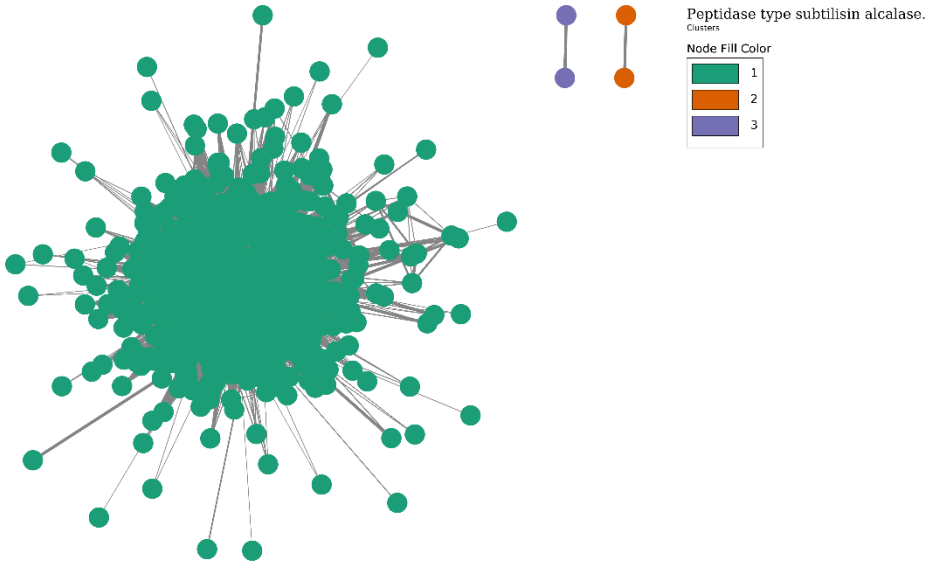
Node Fill Color

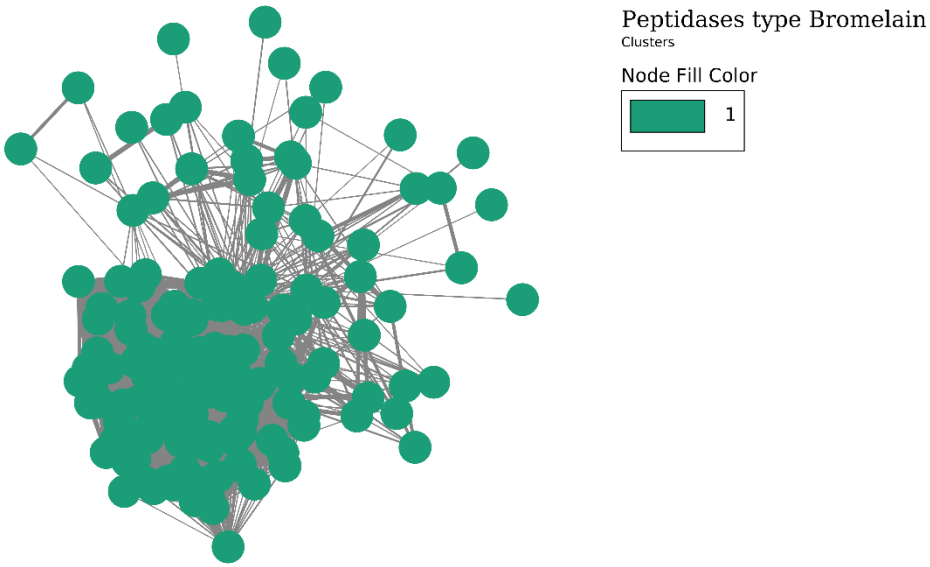
1

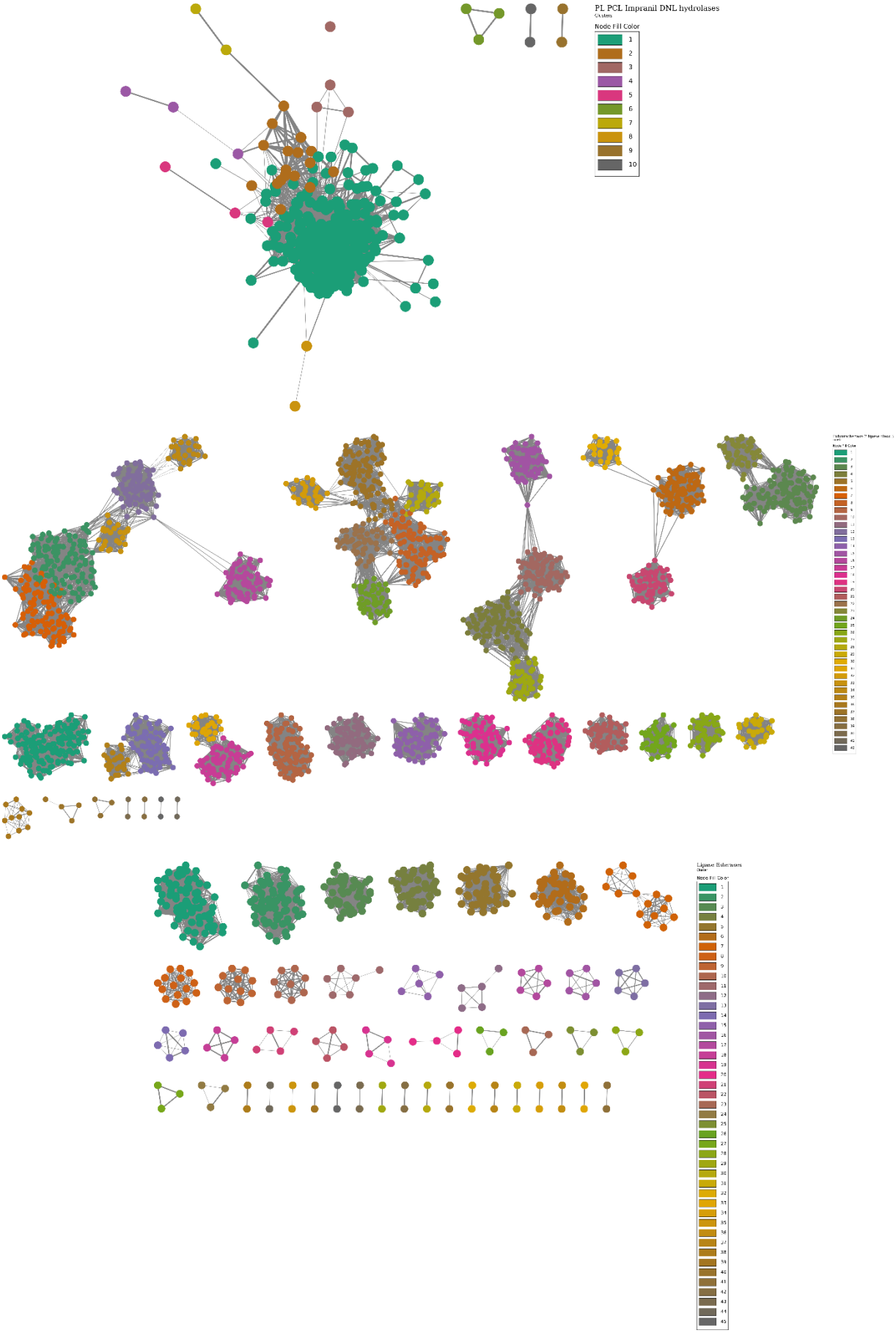





Peptidases type savinase esperase
Node Fill Color
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


Peptidase type subtilisin alcalase.
Clusters
Node Fill Color
1
2
3

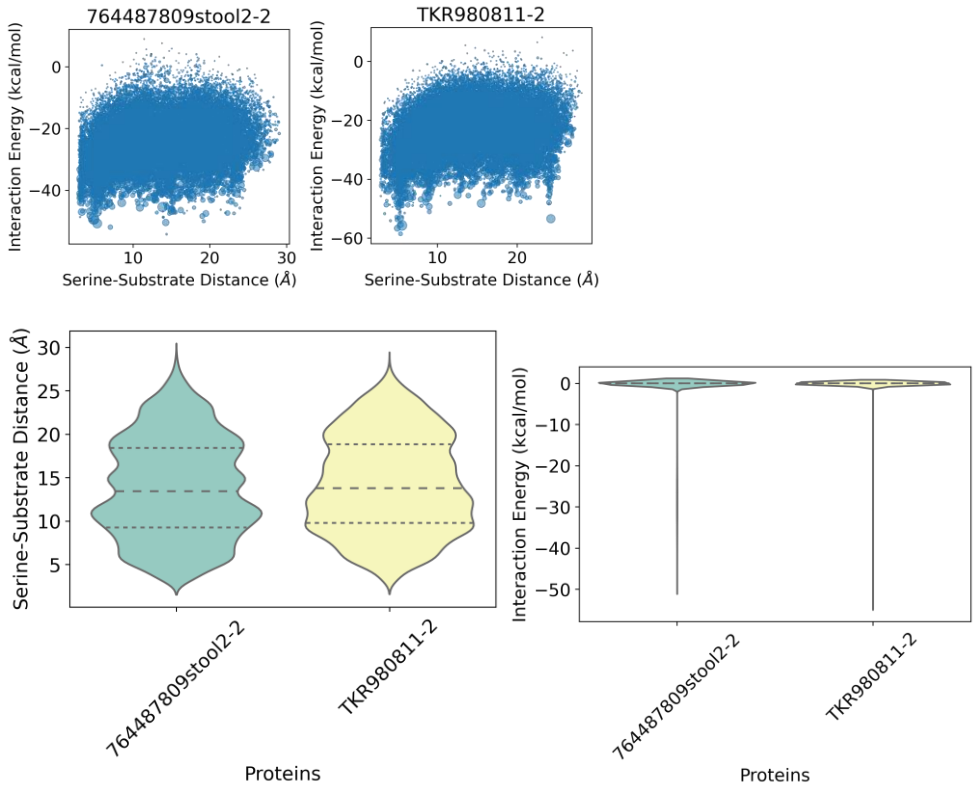

Peptidases type Bromelain
Clusters
Node Fill Color
1


PL PCL Impranil DNL hydrolases
Clusters
Node Fill Color
1
2
3
4
5
6
7
8
9
10

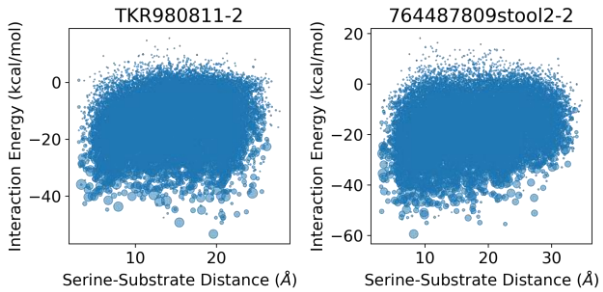
Anex Figure 2: Interaction Energy vs Catalytic Distance Serine-Substrate Plots and Violin Plots of the distribution of Interaction Energies and Serine-Substrate distance along PELE-Induced Fit Simulations for each family of enzymes

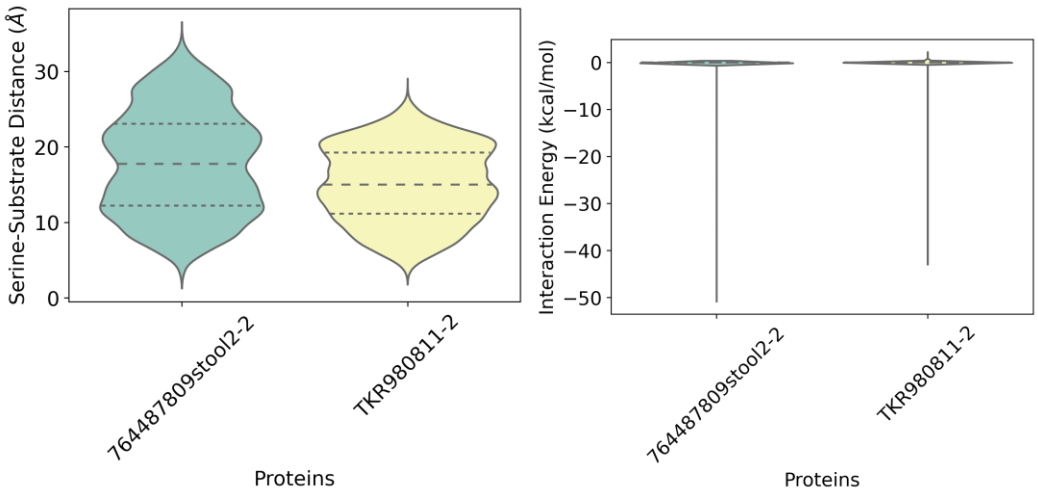
The size of the scatter dots represents the rejected pelesteps that we consider as a time of residence of substrate in the position.

Cysteine Proteases - NY6:

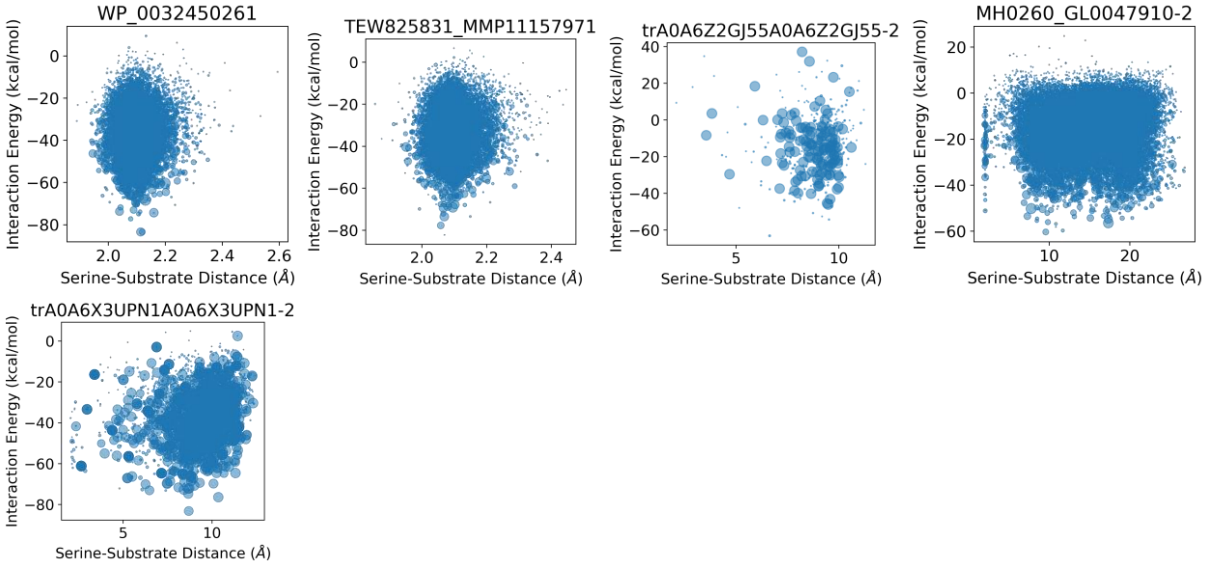


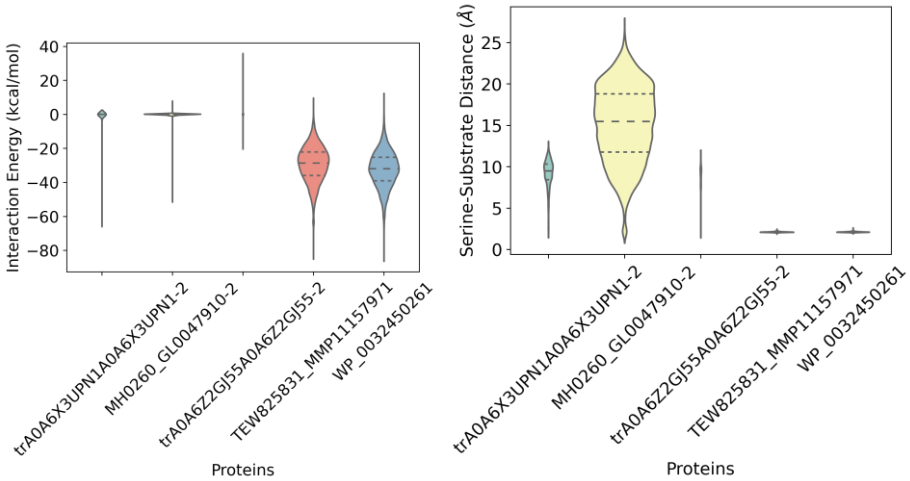
Cysteine Proteases - PRG:



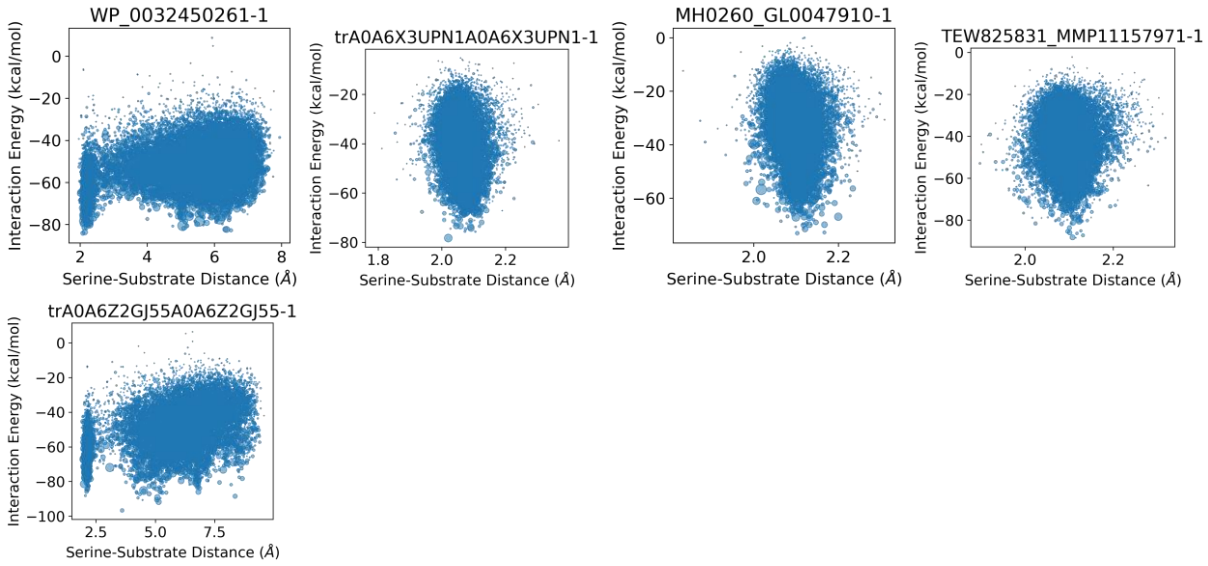


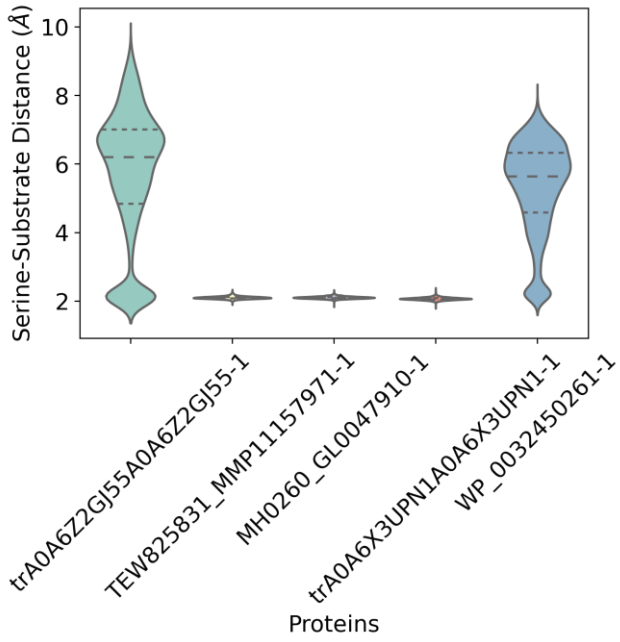
Zinc Proteases - NY6:

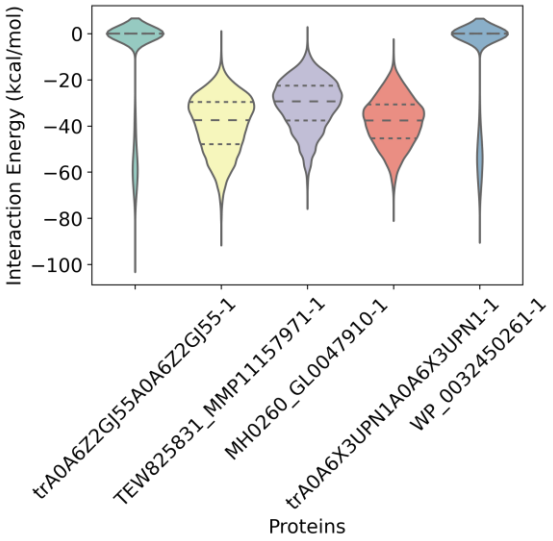




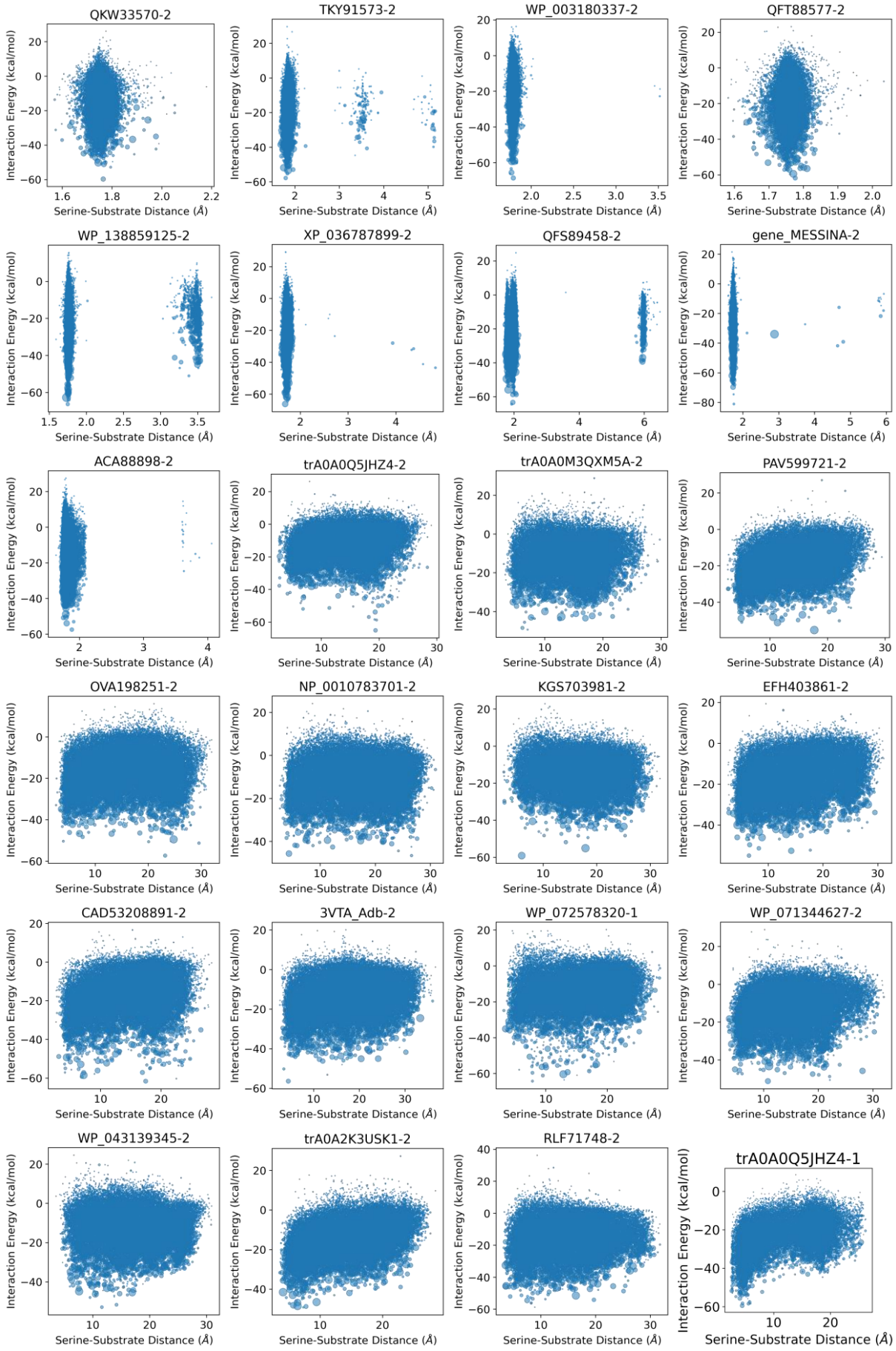
Zinc Proteases PRG:

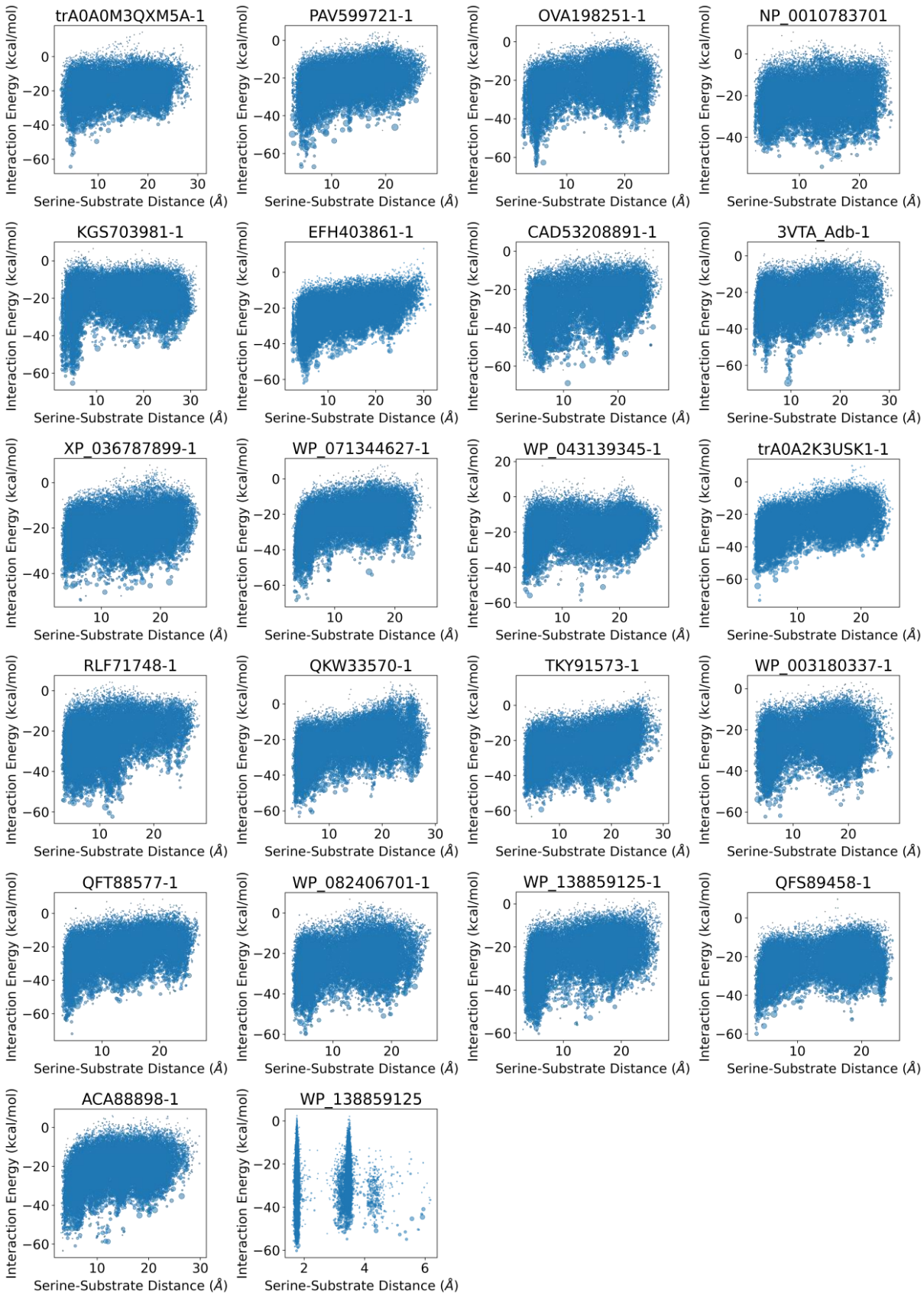


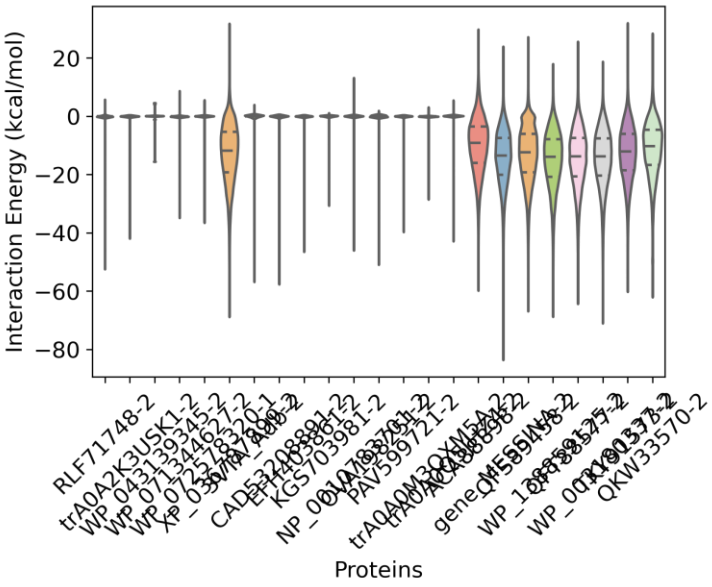


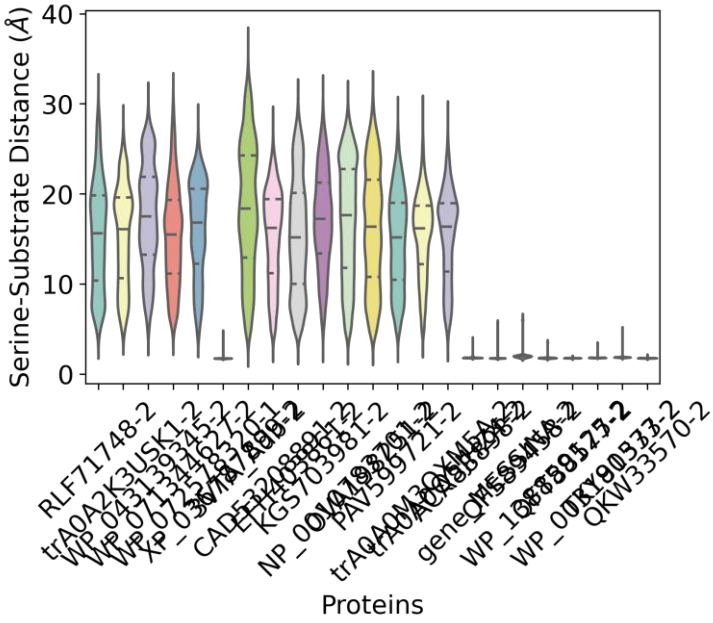


Serine Proteases:

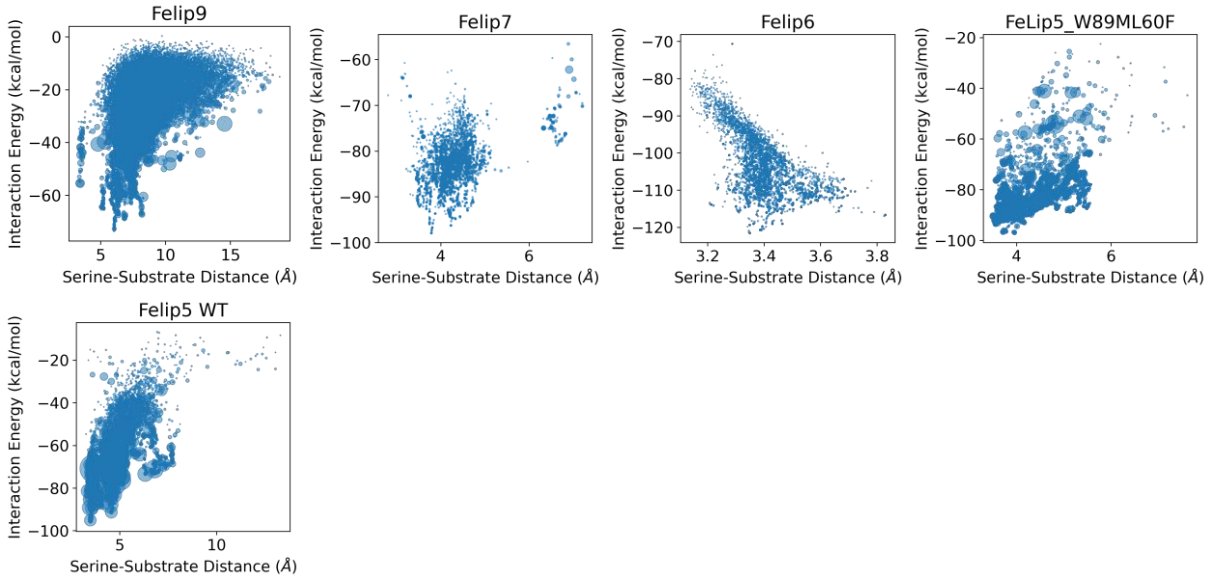


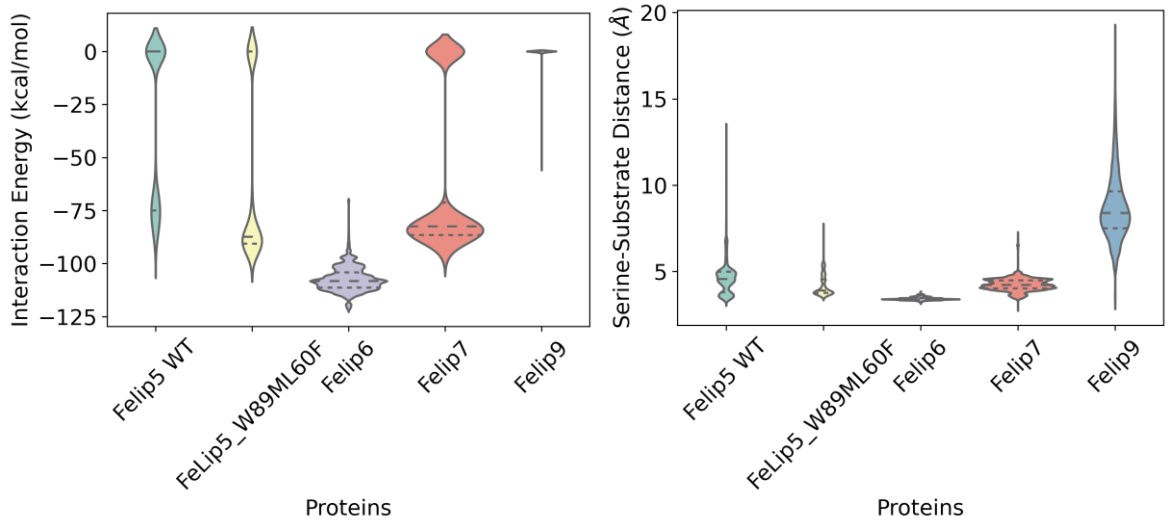

trA0A0M3QXM5A-1
PAV599721-1
OVA198251-1
NP_0010783701
KGS703981-1
EFH403861-1
CAD53208891-1
3VTA_Adb-1
XP_036787899-1
WP_071344627-1
WP_043139345-1
trA0A2K3USK1-1
RLF71748-1
QKW33570-1
TKY91573-1
WP_003180337-1
QFT88577-1
WP_082406701-1
WP_138859125-1
QFS89458-1
ACA88898-1
WP_138859125
Interaction Energy (kcal/mol)
Serine-Substrate Distance (Å)


Interaction Energy (kcal/mol)
20
0
−20
−40
−60
−80
RLF71748-2
trA0A2K3USK1-2
WP_043139345-2
WP_071344627-2
XP_072578320-1
CADF32088891-2
KGS703981-2
PAV59721-2
QKW33570-2
Proteins

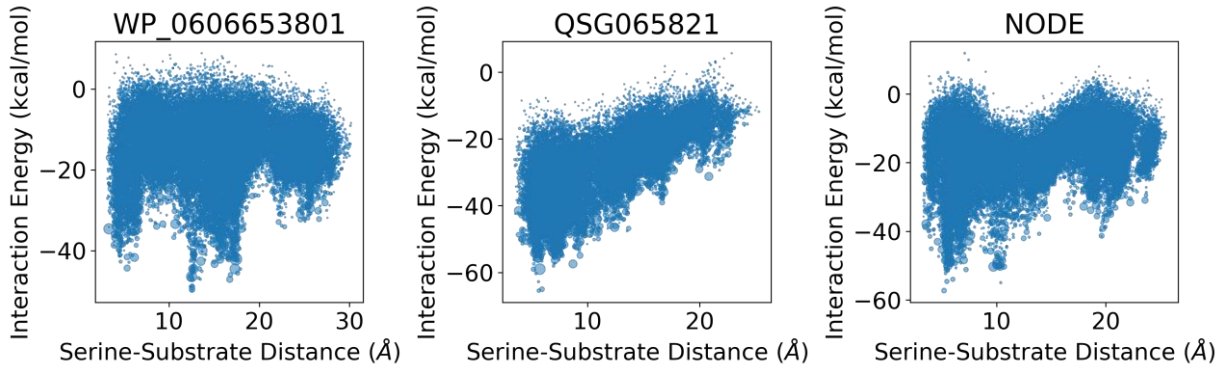

Serine-Substrate Distance (Å)
40
30
20
10
0
RLF71748-2
trA0A2K3USK1-2
WP_043139345-2
WP_071344627-2
XP_072578320-1
CADF32088891-2
KGS703981-2
PAV59721-2
QKW33570-2
Proteins

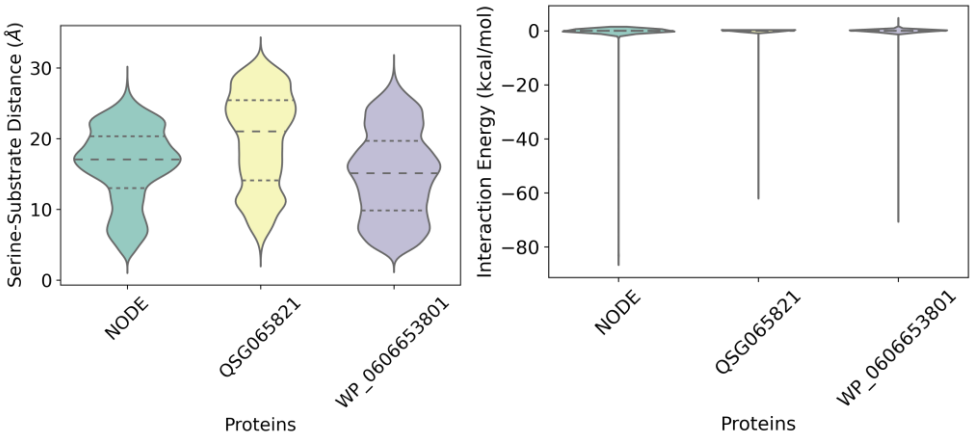
Lipases:



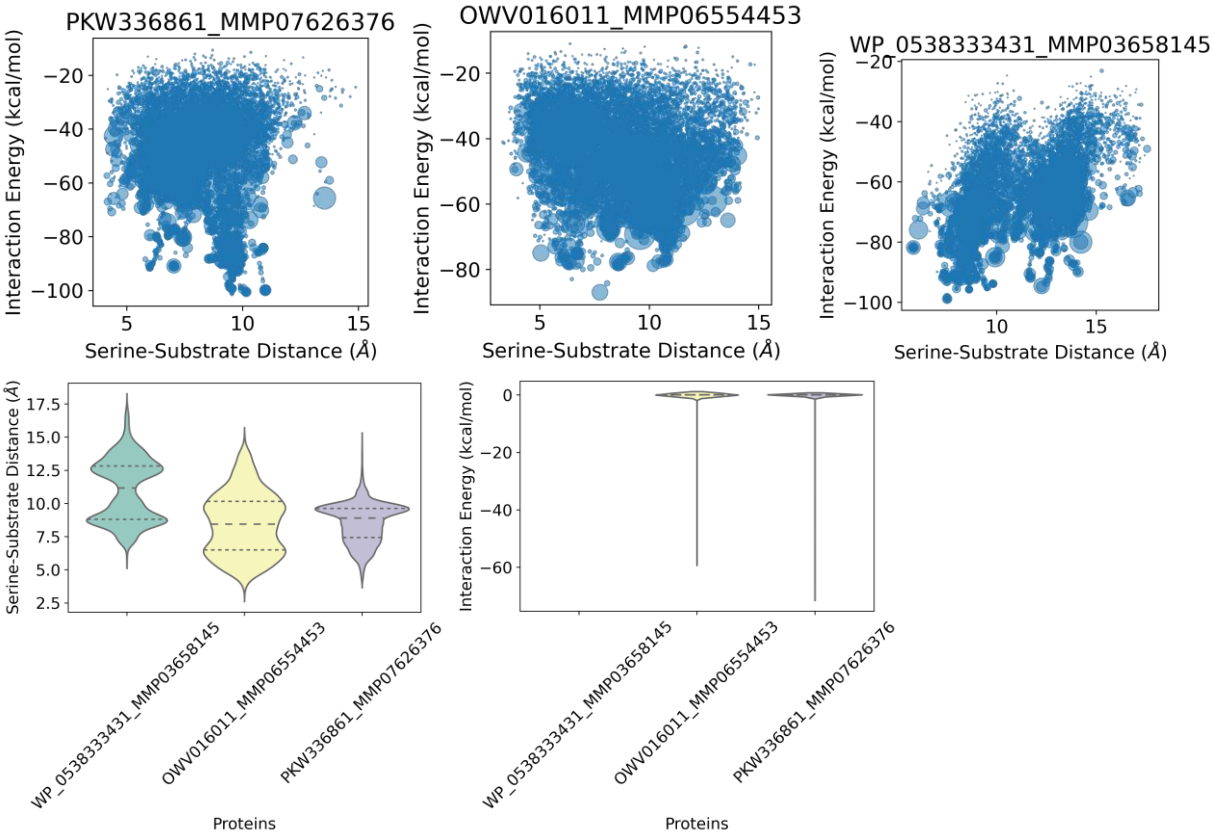


Amylases:

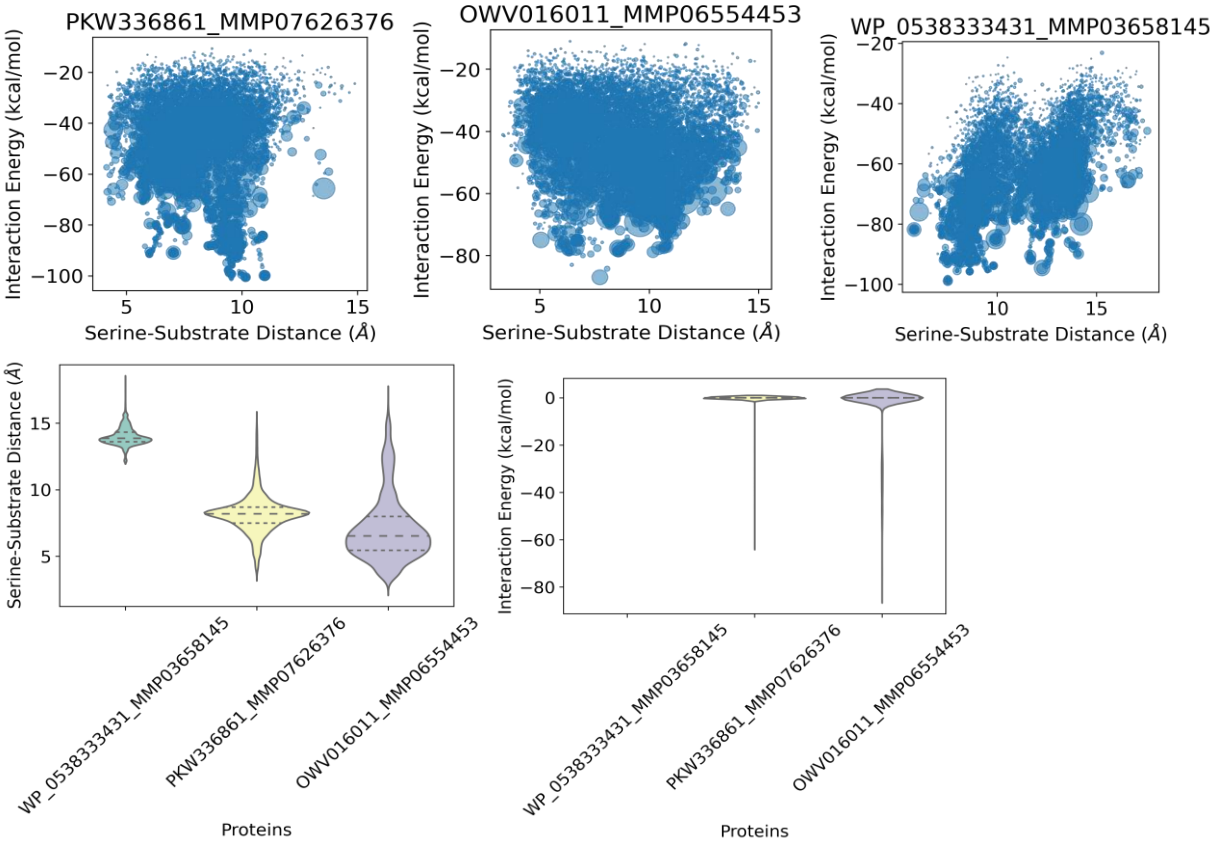




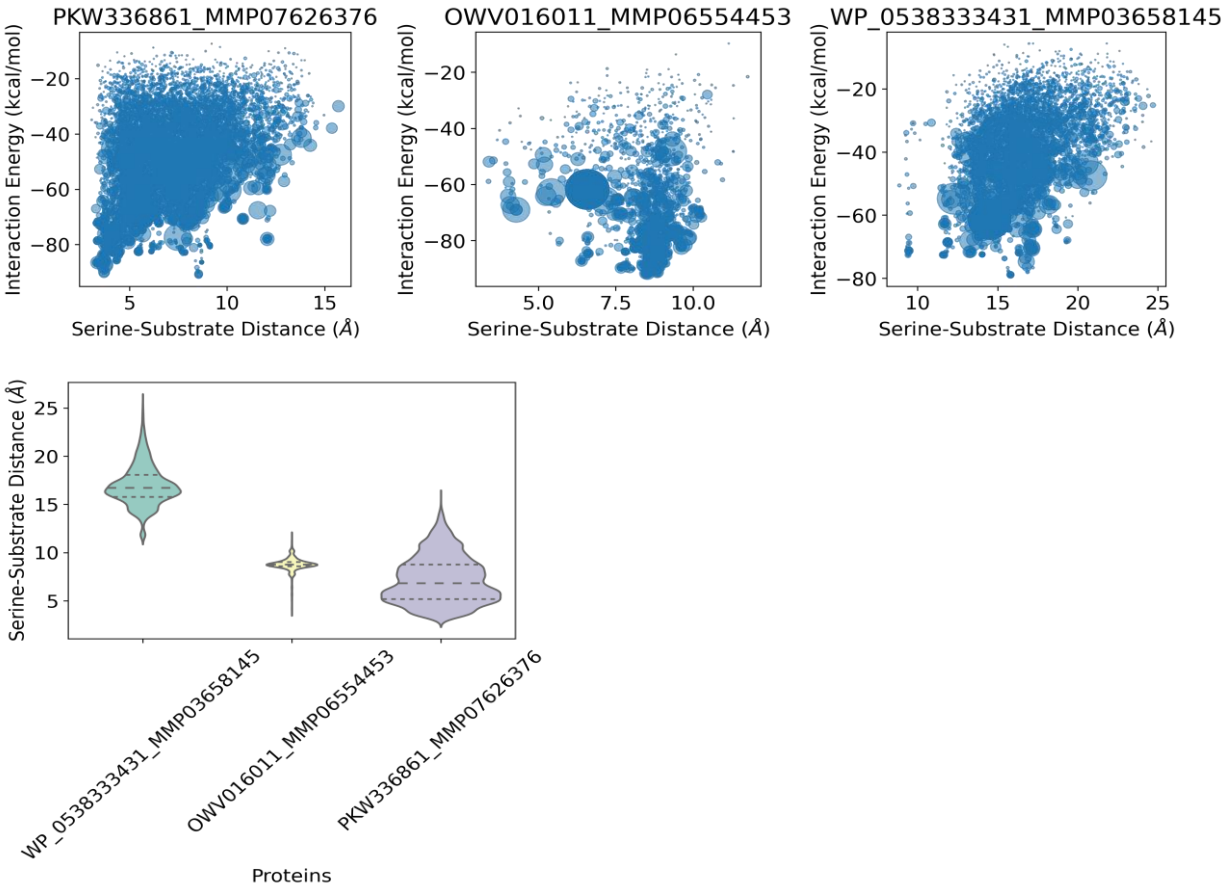
Polylactic acid:



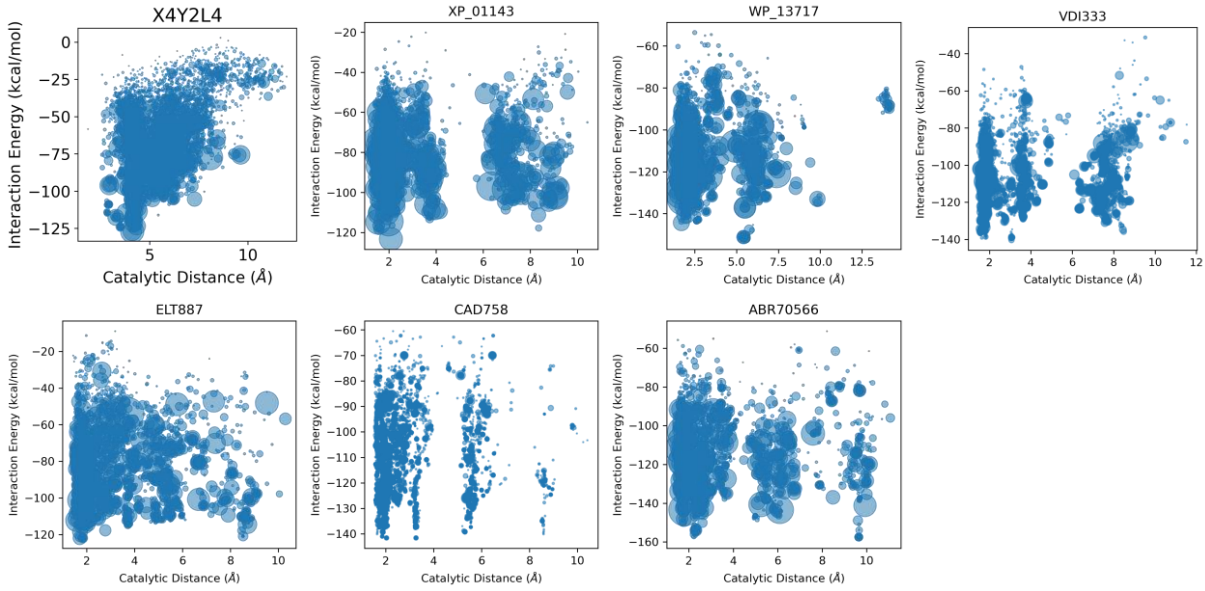
Aliphatic polyurethane:

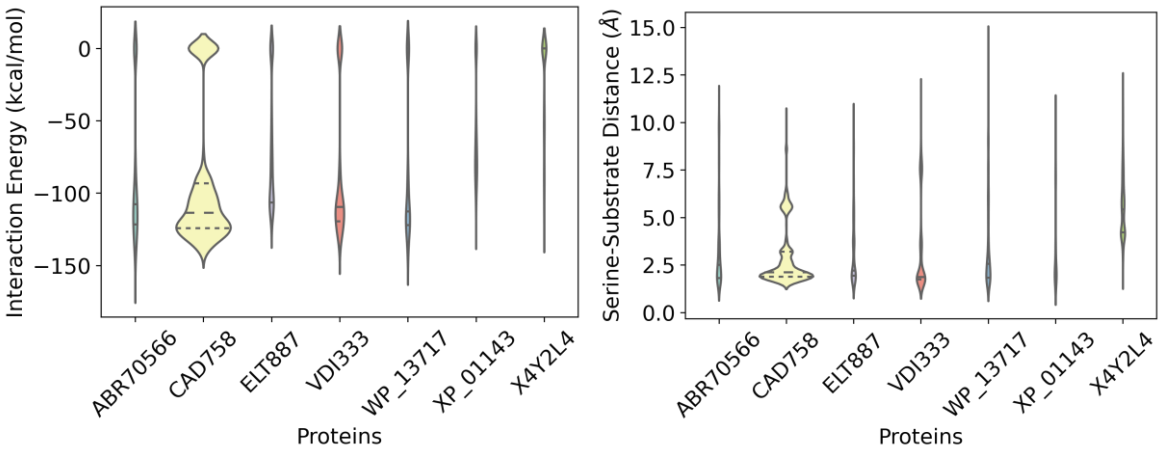


Polycarpolactone:

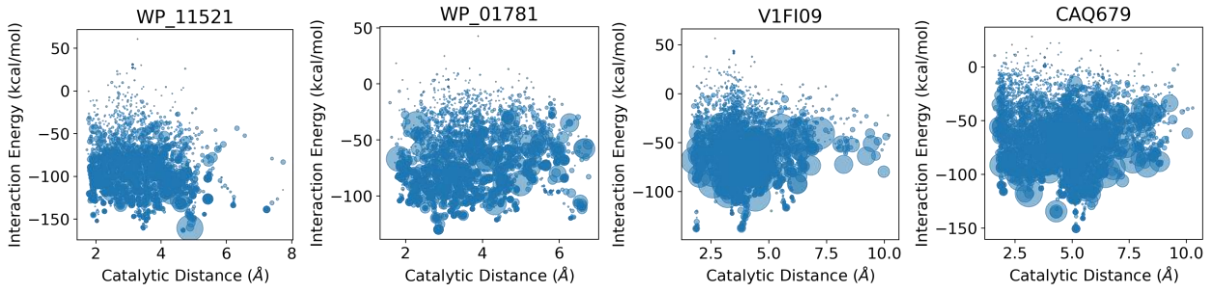


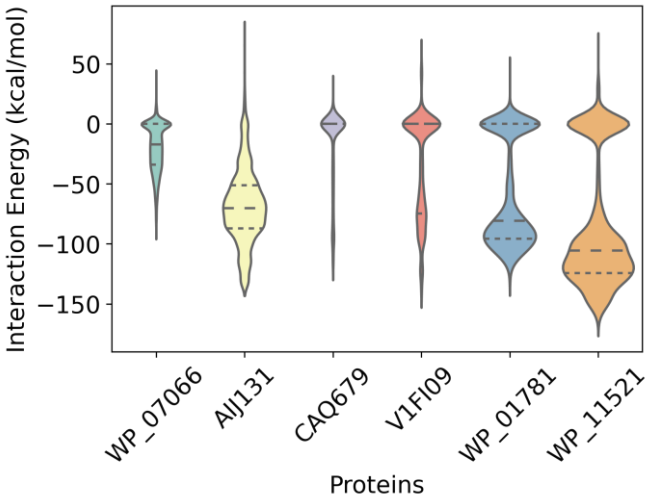
Hyaluronoglucuronidases:





Hyaluronate Lyase:




Interaction Energy (kcal/mol)
50
0
−50
−100
−150
WP_07066
AIJ131
CAQ679
V1FI09
WP_01781
WP_11521
Proteins