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

23/12/2022

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:	2
Date:	23/12/2022
Starting date:	01/06/2021
Duration:	48 months
Lead beneficiary:	BSC
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:	Víctor Guallar, victor.guallar@bsc.es; 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 list of at least 1,000 full-length candidate sequences encoding enzymes with high probability to fulfil manufacturers' specifications (according to the initial proposal). These sequences are selected in Task 2.3 by machine learning techniques applied to the at least 250,000 full-length candidate sequences delivered in deliverable D2.2 (according to the initial proposal). In brief, as a result of the activities done to achieve Deliverable D2.2 "Set of 250,000 sequences pre-selected" (November 2021, updated December 2022), a total of approx. 3.16 million sequences encoding target enzymes were retrieved and pre-selected. In Deliverable 2.3, a number of bioinformatics and computational tools were applied that allow the pre-selection of approx. 1355 sequences encoding enzymes relevant to FuturEnzyme, which were transferred to WP4. Note that these pre-selected sequences do not account those retrieved after functional screens in WP3. The list of sequences retrieved and pre-selected has been compiled in fasta or Excel tables deposited in the FuturEnzyme internal repository and also as a report that accompanies this deliverable (the present document).

2. Reasons for the update

The first version of the Deliverable D2.3 was submitted in May 2022. This update is due to the fact that since the submission, the partners were able to pre-select a new set of sequences by applying the same methods described in the previous version or new ones that were further developed. In November 2022, the Coordinator (Manuel Ferrer) contacted the Project Officer (Colombe Warin) to explain these circumstances and ask her to re-open the submission of this deliverable (amongst others), at which she agreed.

3. Origin of the deliverable

Along the already 18 months of project, one deliverable has been accomplished from which the present one nourishes. To be mentioned:

- D2.2: Set of 250,000 sequences pre-selected (November 2021, updated December 2022)
  In this deliverable, information about the approximately 3.16 million sequences encoding target enzymes that were retrieved and pre-selected by a number of in silico methods are detailed.

4. Introduction & Methodology

4.1. Source and profiling of enzymes

The source of sequences is detailed in Deliverable D2.2 "Set of 250,000 sequences pre-selected" (November 2021, updated December 2022). A total of about 3.16 million sequences encoding target enzymes were retrieved and pre-selected through the applications of a set of bioinformatics and computational tools to public sequence repositories and FuturEnzyme genomes and metagenomes sequences. The tools included: i) DIAMOND BLASTP, PSI-BLAST (EMBL-EBI) and Hidden Markov Model (HMM), high-throughput programs for aligning protein sequences against protein reference databases; ii) MCL algorithm (Markov Cluster Algorithm), an efficient algorithm for large-scale detection of protein families through network analyses; iii) PELE (Protein Energy Landscape Exploration), a protein-ligand Monte Carlo simulations software; and iv) the machine learning EP-Pred ensemble classifier.

4.2. Pre-selections by computationally-based rigidity screens

We applied the AlphaFold2-based workflow of ColabFold to generate 3D structural models of selected enzymes. A single model was generated for each enzyme with ten prediction cycles (--num_recycles) and structurally refined by running a relaxation with AMBER (--amber). For subsequent analyses, only enzymes with a sufficient quality of the 3D structural model, with a sequence length < 1000 residues, and without cofactors were considered. To test whether the catalytically active residues (CARs) of the 3D structures are accessible for substrates, we used the CAVER 3.0.3 PyMOL Plugin. Therefore, CARs were identified based on the minimal summed distances between all catalytic amino acids. Starting points for the computations were defined based on the Cartesian coordinates of the CARs' centre of mass (COM). Default values were used for the probe radius (0.9 Å), shell radius (3.0 Å), and shell depth (4.0 Å). We verified that CARs in all models are accessible for substrates, i.e. that all models are in an open conformation: CARs are either located on the protein surface or are buried and connected with the surface by tunnels.

The enzyme structures were pre-processed with pdb4amber, which is part of AmberTools21, and hydrogen atoms were added using the Reduce program. The prepared enzymes were solvated in a truncated octahedron of TIP3P water, leaving at least 20 Å between the enzyme structure and the edges of the solvent box, using the LeaP program of AmberTools21. All systems were neutralized by adding Na^+ or Cl^- ions as needed. We used the Amber ff14SB force field to parametrize the protein. Ion parameters were taken from Joung and Cheatham. Structural ensembles of enzymes were generated by all-atom MD simulations, with five replicas at 500 ns, yielding 2.5 µs of cumulative simulation time per enzyme. Minimization steps, thermalization, and production simulations were carried out using the GPU-accelerated CUDA version of PMEMD from the Amber21 suite of programs. The systems were heated to 298 K, and the pressure was adapted in NPT simulations such that a density of 1 g cm^{-3} was obtained. During thermalization and density adaptation, we kept the solute fixed by positional restraints of 1 kcal mol^{-1} $Å^{-2}$, which were gradually removed over five steps in short subsequent NVT simulations. Afterwards, five NVT production simulations of 500 ns length were performed using unbiased MD simulations. During production simulations, we set the time step to integrate Newton's equation of motion to 4 fs following the hydrogen mass repartitioning strategy. Coordinates were saved every 200 ps yielding 2500 conformations for each production run that were considered for subsequent analyses.

Rigidity analyses were performed using the Constraint Network Analysis (CNA) software package (version 3.0). In detail, we applied CNA on ensembles of network topologies generated from conformational ensembles obtained from MD simulations. Average stability characteristics were calculated by constraint counting on each topology in the ensemble. CNA functions as a front- and back-end to the graph theory-based software Floppy Inclusions and Rigid Substructure Topography (FIRST). Applying CNA to biomolecules aims to identify their composition of rigid clusters and flexible regions, which can aid in understanding the biomolecular structure, stability, and function. As the mechanical heterogeneity of biomolecular structures is intimately linked to their diverse biological functions, biomolecules generally show a hierarchy of rigid and flexible regions. To monitor this hierarchy, CNA performs thermal unfolding simulations by consecutively removing noncovalent constraints (hydrogen bonds and salt bridges) from a network in the order of their increasing strength. Therefore, a hydrogen bond energy E_{HB} is computed from an empirical energy function. For a given network state $\sigma = f(T)$, hydrogen bonds (including salt bridges) with an energy $E_{HB} > E_{cut}(\sigma)$ are removed from the network at temperature T. In the present study, thermal unfolding simulations were carried out by decreasing E_{cut} from -0.1 kcal mol^{-1} to –6.0 kcal mol^{-1} with a step size of 0.1 kcal mol^{-1}. As E_{cut} can be converted to a temperature T using the linear equation (see below equation), the range of E_{cut} is equivalent to increasing the temperature from 302 K to 420 K with a step size of 2 K.

$$T = -\frac{20\mathrm{K}}{(\mathrm{kcal\ mol^{-1}})} * E_{cut} + 300\mathrm{K}$$

The number of hydrophobic tethers was kept constant during the thermal unfolding simulations. From the thermal unfolding simulations, CNA computes a set of indices to quantify biologically relevant characteristics of the protein's stability at a global and local scale. We used the cluster configuration entropy H_{type2}, a measure for the global structural stability, to predict the phase transition temperature T_p (for details on H_{type2}). At T_p, the protein switches from a rigid (structurally stable) to a floppy (unfolded) state, and the largest rigid cluster stops to dominate the whole protein network. If the largest rigid cluster dominates the whole protein network, H_{type2} is low because of the limited number of possible ways to configure a system with a very large cluster. When the largest rigid cluster starts to decay or stops to dominate the network, H_{type2} jumps. There, the network is in a partially flexible state with many ways to configure a system consisting of many small clusters. The percolation behaviour of protein networks is usually complex, and multiple phase transitions can be observed. To identify T_p, a double sigmoid fit was applied to an H_{type2} *versus* $T(E_{\text{cut}})$ curve as done previously. In general, T_p was taken as that T value associated with the largest slope of the fit except for enzymes with $T_d > 50°C$, for which the second phase transition was chosen to focus on the decomposition of the core.

Figure 1 summarizes the pipeline used for enzyme pre-selection based on rigidity analysis.

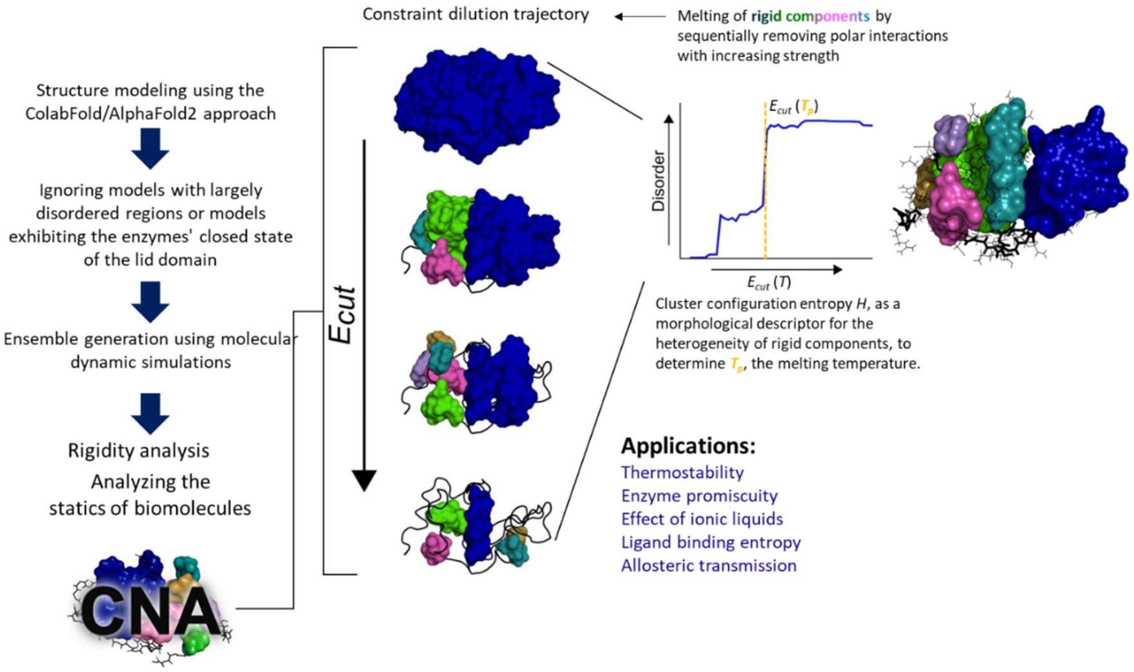

Constraint dilution trajectory
Melting of rigid components by sequentially removing polar interactions with increasing strength
Structure modeling using the ColabFold/AlphaFold2 approach
Ignoring models with largely disordered regions or models exhibiting the enzymes' closed state of the lid domain
Ensemble generation using molecular dynamic simulations
Rigidity analysis
Analyzing the statics of biomolecules
CNA
Ecut
Disorder
E_{cut} (T_p)
E_{cut} (T)
Cluster configuration entropy H, as a morphological descriptor for the heterogeneity of rigid components, to determine T_p, the melting temperature.
Applications:
Thermostability
Enzyme promiscuity
Effect of ionic liquids
Ligand binding entropy
Allosteric transmission

Figure 1. Illustration representing the pipeline used for enzyme pre-selection based on rigidity analysis.

4.3 Pre-selection by computationally-based Protein Energy Landscape Exploration (PELE)

We developed a pipeline to characterize 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 2**).

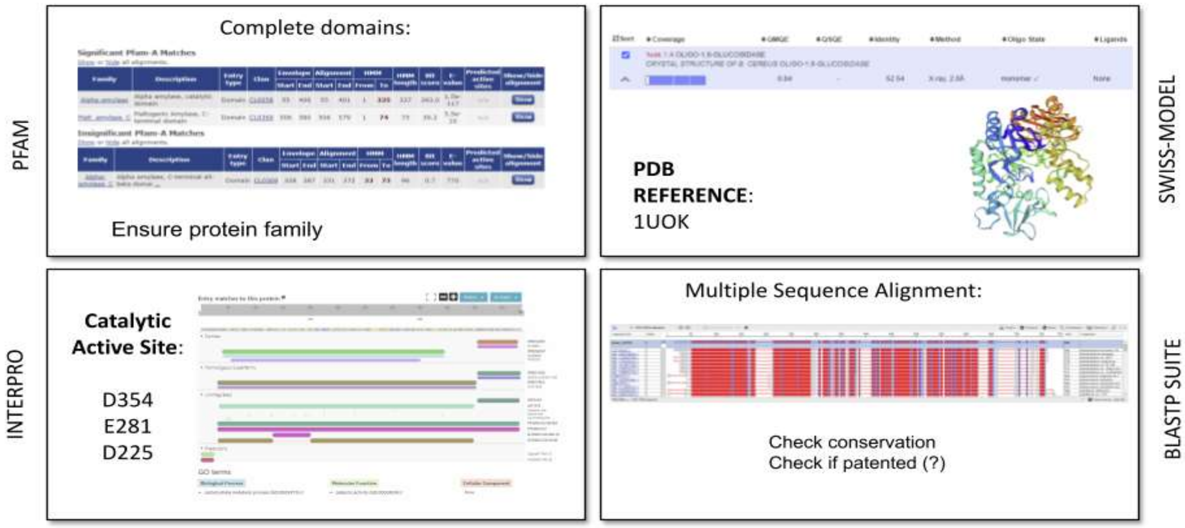


Figure 2. Illustration representing the software used to check the sequences with bioinformatic tools.

The sequences that passed this first filtering were modelled 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 partner BSC, Protein Energy Landscape Exploration (PELE). To account for the goodness of an enzyme-substrate interaction, the measure of the catalytic events was extracted (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 3**).

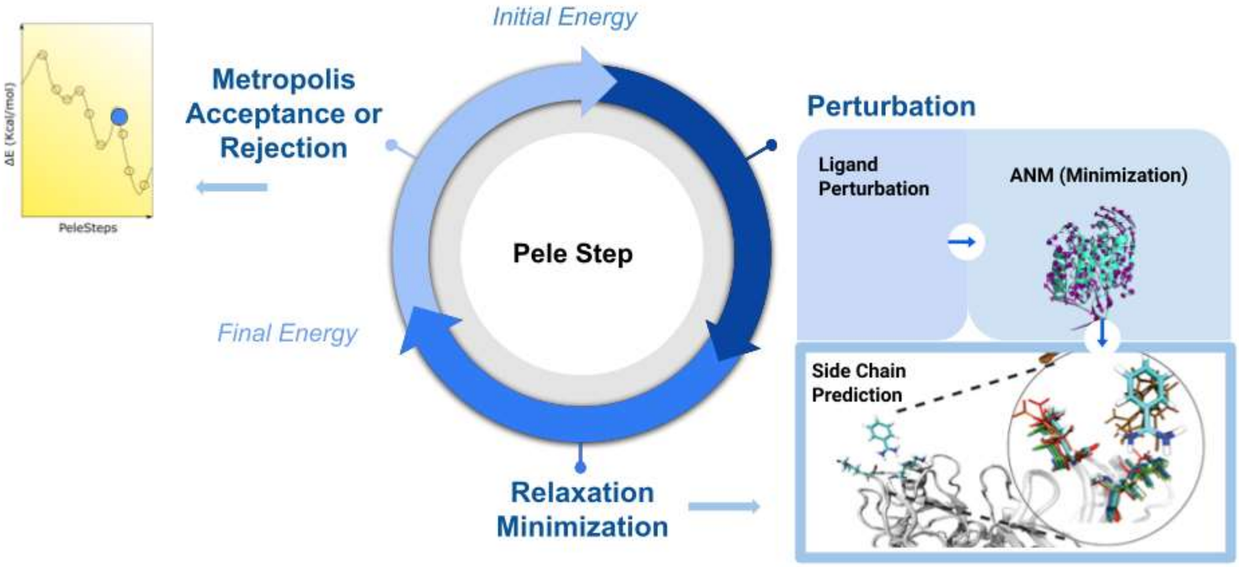
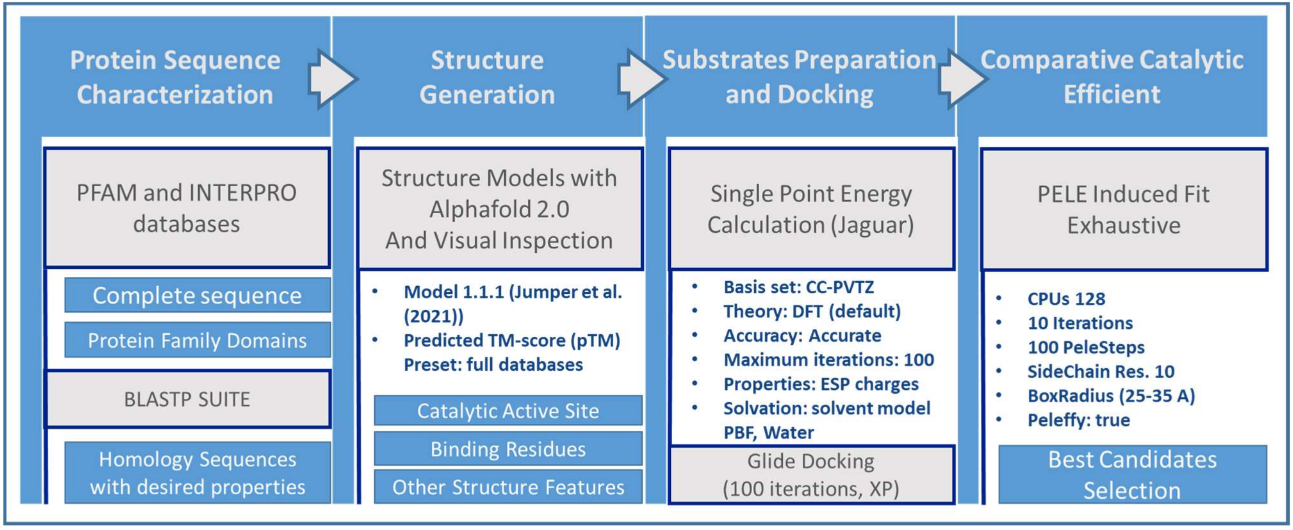


Figure 3. Scheme explaining the workflow of PELE's software.

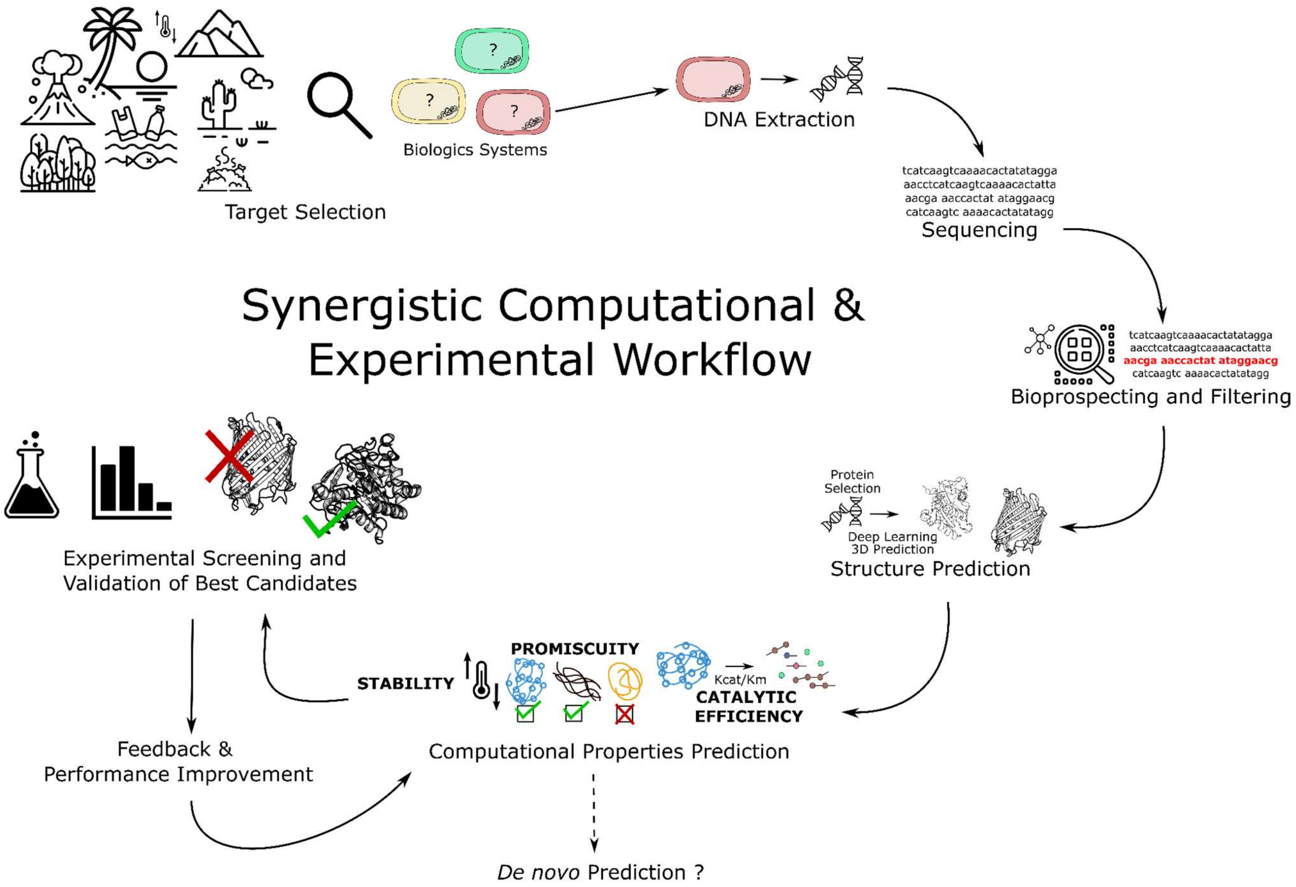
The substrates were downloaded from the PubChem database, and their electrostatic point (ESP) charges were calculated from a quantum mechanics single point energy calculation with the Jaguar software from the Schrödinger company. These ESP charges were used in the mentioned induced-fit PELE simulations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

Table 1 illustrates the summary workflow for pre-selecting sequences encoding enzymes with appropriated number of catalytic events towards target substrates. Those candidates are further transferred to WP4 for extensive characterization and validation of the pipeline (**Figure 4**).

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	• Model 1.1.1 (Jumper et al. (2021)) • Predicted TM-score (pTM) Preset: full databases	• Basis set: CC-PVTZ • Theory: DFT (default) • Accuracy: Accurate • Maximum iterations: 100 • Properties: ESP charges • Solvation: solvent model PBF, Water	• CPUs 128 • 10 Iterations • 100 PeleSteps • SideChain Res. 10 • BoxRadius (25-35 A) • Peleffy: true
BLASTP SUITE	Catalytic Active Site Binding Residues Other Structure Features	Glide Docking (100 iterations, XP)	Best Candidates Selection
Homology Sequences with desired properties			


Target Selection
Biologics Systems
DNA Extraction
tcatcaagtcaaaacactatatagga
aacctcatcaagtcaaaacactatta
aacga aaccactat ataggaacg
catcaagtc aaaacactatatagg
Sequencing
Synergistic Computational & Experimental Workflow
Bioprospecting and Filtering
Protein Selection
Deep Learning 3D Prediction
Structure Prediction
Experimental Screening and Validation of Best Candidates
STABILITY
PROMISCUITY
Kcat/Km
CATALYTIC EFFICIENCY
Computational Properties Prediction
Feedback & Performance Improvement
De novo Prediction ?

☐

Figure 4. Experimental and computational workflow to search for new enzymes.

4.4 Pre-selection through PSI-BLAST and PELE (Protein Energy Landscape Exploration)

In some cases, PSI-BLAST and PELE (Protein Energy Landscape Exploration) were used in combination to identify homologues to one enzyme that in the frame of the FuturEnzyme project was selected among the priority targets, such as the lipase Lip9. In brief, we applied PSI-BLAST to find a reasonable amount of similar enzymes in databases and test them using the protein-ligand Monte Carlo simulations software, PELE (Protein Energy Landscape Exploration). By doing so, one can ensure that the selected sequences are not overly similar to an original enzyme, but do retain similar or improved characteristics. More in details, for the bioprospecting of Lip9, hundreds of sequences were searched by means of PSI-BLAST, a tool designed to find distant homologs for a certain protein, using Lip9 as the seed. Then, the sequences were filtered through different parameters, including AlphaFold confidence level, the alignability of the catalytic residues to the Lip9 catalytic triad, the existence of a spatially well-designed triad in the AlphaFold models, and the low resemblance to a patented lipase (WP_106066877.1). A set of 15 different ligands were docked, all of them being triglycerides constituting different grease stains. Finally, protein-ligand simulations were run on the selected new sequences with the in-house all-atom Monte Carlo molecular modelling sampling technique PELE.

4.5 Preselection using the machine learning EP-Pred method

A machine learning tool for bioprospecting enzymes relevant to FuturEnzyme was also implemented. Briefly, we implemented a method called EP-Pred, an ensemble binary classifier built to predict the promiscuity of ester hydrolases. It combines 3 different machine learning algorithms: Support vector machines (SVC), K-nearest neighbours (KNN) and a lineal model (the RidgeClassifier implementation on Sckit-Learn. It was trained on a dataset containing 147 phylogenetically diverse esterases and their activity on 96 distinct ester substrates. The labelling of the classes was based on the number of substrates catalyzed, where 20 or more substrates were considered promiscuous and less than 20, non-promiscuous. The program can be dowonloaded in GitHub etiur/EP-pred: A machine learning program to predict promiscuity of esterases (github.com). For its use, it is required to install 3 external programs: Ifeature, Possum and Blast+ NCBI. It is also needed a protein database, in this case, the Uniref50. The main.py script will then perform the rest of the operations if provided with the input esterases and the appropriate flags. It will transform the uniref50 into a Blast database and use it to extract the PSSM profiles. It will generate the features used by the models using Ifeature and Possum and finally it will predict the promiscuity of the sequences. EP-Pred has been evaluated against the Lipase Engineering Database (http://www.led.uni-stuttgart.de/) together with a HMMs approach leading to select sequences encoding esterases and lipases. For extensive details see the recent reference (https://www.mdpi.com/2218-273X/12/10/1529). **Figure 5** summarizes the EP-Pred pipeline for enzyme pre-selection.

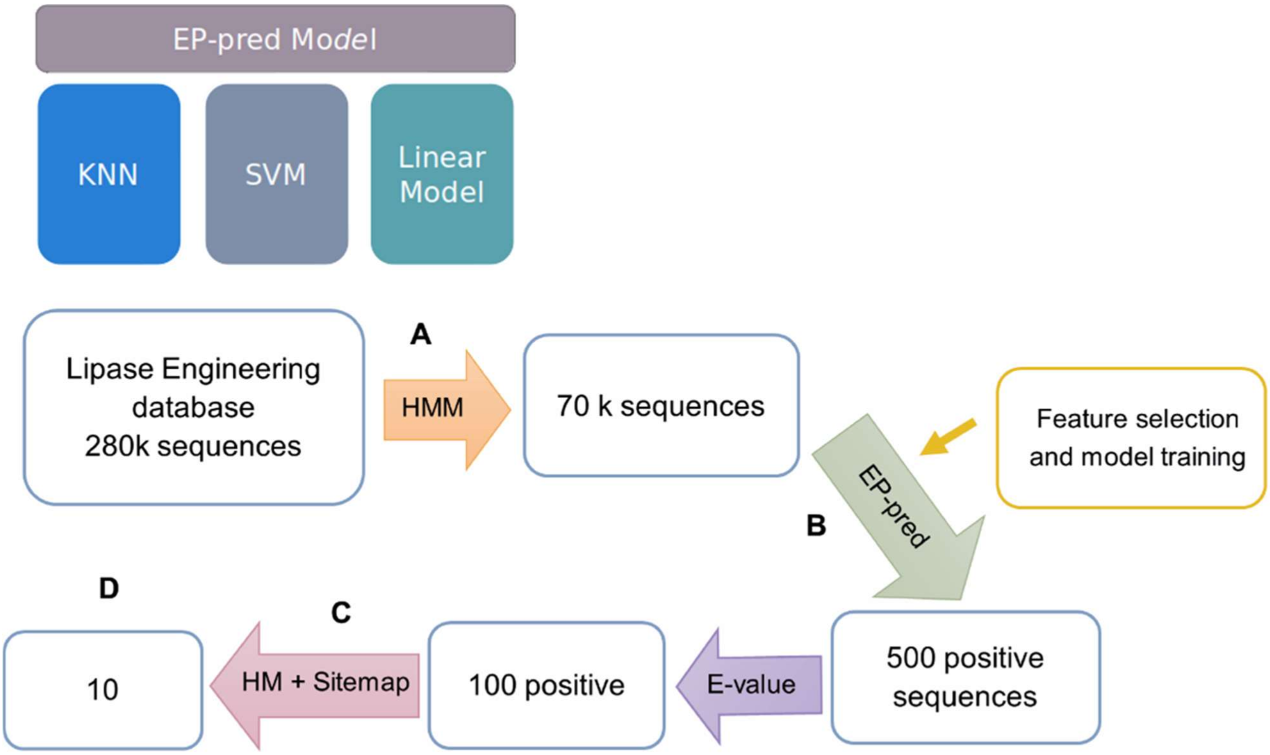


Figure 5. EP-Pred pipeline for enzyme pre-selection. A, Since there was a mix of different families in LED, first we applied an HMM profile created from the esterase dataset to clean the database and keep only esterases. B, EP-pred evaluated the remaining sequences and predicted around 500 positive hits. C, The top 100 sequences according to E-values returned by HMM in step A were isolated and analyzed according to molecular descriptors from homology modeling (HM) and Sitemap calculations. D, A final set of 10 sequences with the highest hydrophobicity and enclosure/exposure scores were gathered and sent to be validated experimentally.

4.5 Pre-selection by Hidden Markov Model (HMM)

As detailed in Deliverable D2.2, HMMs was performed to identify homologues to one enzyme that in the frame of the FuturEnzyme project was selected among the priority targets, such as the lipase Lip9. In this case, the sequence of Lip9 was compared against NCBI's database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) and in-house collections of genomes and shotgun sequences of metagenomes obtained by shotgun sequencing. Alignment was performed with diamond 2.0.15.153 (https://doi.org/10.1038/s41592-021-01101-x) and alignments within 25% range of top alignment score were reported.

4.6 Pre-selection by Stability Consensus Metapredictor (SCOT)

SCOT is a Random Forest based Machine Learning metapredictor that combines the estimations of 8 already published protein stability predictors and a molecular filter to produce a more reliable result. Predictors: MAESTRO, CUPSAT, AUTOMUTE-SVM and AUTOMUTE-TR, FOLDX, INPS3D, MUPRO and I-MUTANT. More in details, Delta Delta G (DDG) is a metric for predicting how a single point mutation will affect protein stability. DDG, often referred to as **ΔΔ**G, is the change in the change in Gibbs free energy (double changes intended). DDG is a measure of the change in energy between the folded and unfolded states (**Δ**Gfolding) and the change in **Δ**Gfolding when a point mutation is present. This has been found to be an excellent predictor of whether a point mutation will be favourable in terms of protein stability. There are other protein stability predictors that can be apply such as: PROSS, FIREPROT 2.0, or Cyrus DDG tool from Rosetta. The Cyrus DDG tool uses a new version Rosetta DDG calculation, Cartesian DDG, which has a variety of improvements over the previous method (Kellogg). The new method is more complex, but fully automated in Cyrus Bench, and features improvements in energy function, mutations with change in net charge, proline free energies, solvation models, side chain optimization, and structure preparation. The new version has improved accuracy, better consistency across similar input protein structures, and most notably fewer outliers with highly inaccurate calculated results.

5. Results

5.1 Pre-selection by PELE of sequences selected by BLASTP-DIAMOND and MCL algorithm

As detailed in Deliverable D2.2 “Set of 250,000 sequences pre-selected” (November 2021, updated December 2022), a total of 3,152,857 sequences were pre-selected by applying BLASTP and DIAMOND BLASTP against public sequence repositories and internal FuturEnzyme sequences detailed in Section 4.2, using also the manually curated and customized database contain 37,403 taxonomically diverse protein sequences featuring the key enzyme families relevant to FuturEnzyme (see Section 4.1). Network analysis by MCL algorithm further revealed that they grouped into 457 clusters, each containing enzymes that most likely do show similar properties. A total of 457 sequences of reference conforming each of the clusters were pre-selected (Annex **Figure 1**) and further checked whether the sequence contained the proper catalytic domains and catalytic residues, and the presence of signal peptides, by performing the computational pipeline summarized in **Table 1** for the different needs in the project.

As detailed in the Deliverable D2.1 “Manufacturers' needs and specifications, protocol”, in the case of Evonik’s needs, the objective is to identify hyaluronidase-like enzymes capable of degrading hyaluronic acid to small hyaluronic acid products with 1-2 kDa molecular weight, at <37˚C, no solvents, and high viscosity solutions. Priority targets will be enzymes degrading hyaluronic acid:

- Heparanase (EC 3.2.1.166)
- Hyaluronate lyase (cd01083 - EC 4.2.2.1)
- Hyaluronidase (EC 3.2.1.35, EC3.2.1.36, pfam03662, pfam01630).

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 6** lists the number of EC3.2.1.36/166 hyaluronidases pre-selected as having appropriated catalytic events.

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 7** lists the number of 4.2.2.1 hyaluronate lyases pre-selected as having appropriated catalytic events.

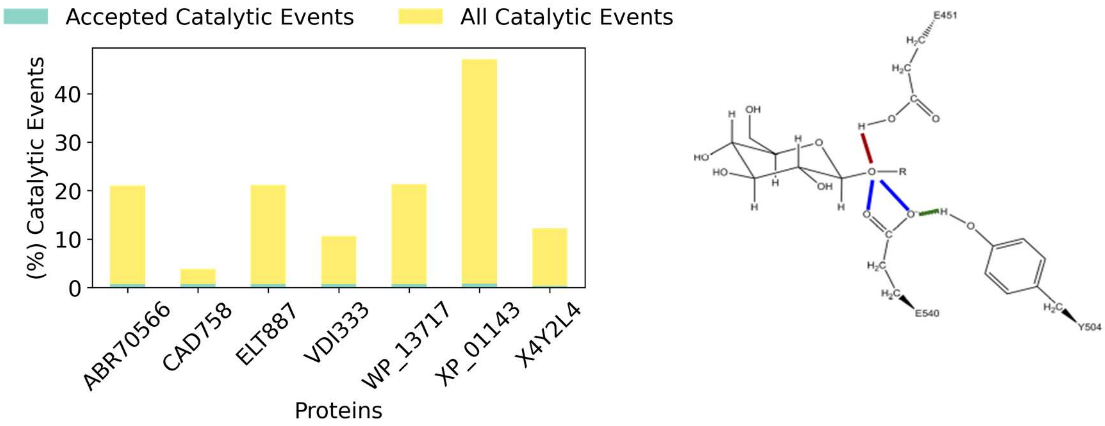


Figure 6. Plot showing the number of catalytic events in the 3.2.1.36/166 hyaluronidases compared to a control from UniProt entry; X4Y2L4 (left). Catalytic residues and the catalytic distances of 3.2.1.36/166 hyaluronidases highlighted (right).

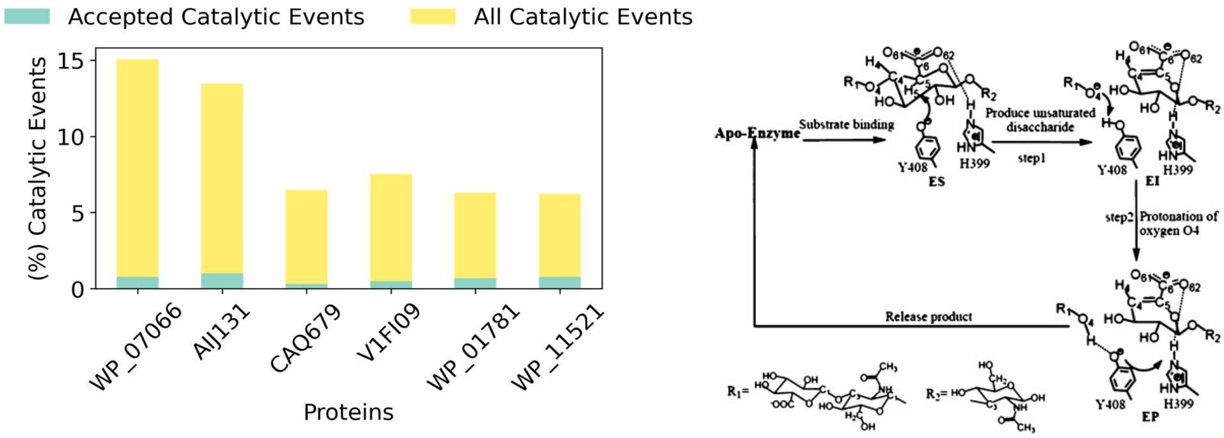


Figure 7. Plot showing the number of catalytic events in the 4.2.2.1 hyaluronate lyases. Catalytic residues and the catalytic mechanism of 4.2.2.1 hyaluronidases (right). Image taken from https://pubs.acs.org/doi/10.1021/jp406206s.

As detailed in the Deliverable D2.1 “Manufacturers' needs and specifications, protocol”, in the case of Schoeller’s needs and specifications, the following are considered priority:

- Priority 1: Lipases for removing residual spinning oils/sizing products that, if not eliminated, will otherwise generate emissions during the drying and fixation steps; priority textiles are those made of polyester (PES). Schoeller requested enzymes working in water, and temperatures below 80˚C.
- Priority 2: Oxidoreductases (laccase or peroxidase-like) for supporting in the decolorization of dyes. Schoeller requested enzymes working in water, and temperatures below 80˚C.
- Priority 3: Polyesterases that can be applied in the biodegradation of the current textile materials in such a way that they can even be reused to produce new recycled textiles. Schoeller did not request any working conditions.

Other enzymes involved in different processes are requested at lower priority level. 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 oxidoreductases.

The substrates used for the computational simulations of plastic degrading enzymes with PELE were, where those related to the intermediates during the degradation of polyurethane and polyester (polyethylene terephthalate, PET): polyurethane dimer, mono-(2-hydroxyethyl) terephthalic acid (MHET) (**Figure 8**), several ester polymers like polylactic acid (PLA), polycaprolactone (PCL), and aliphatic polyurethane, and two types of proteins: 6-units of nylon and 7-units of polyglycine). **Figure 8** lists the number of polyester degrading hydrolases as having appropriated catalytic events with MHET.

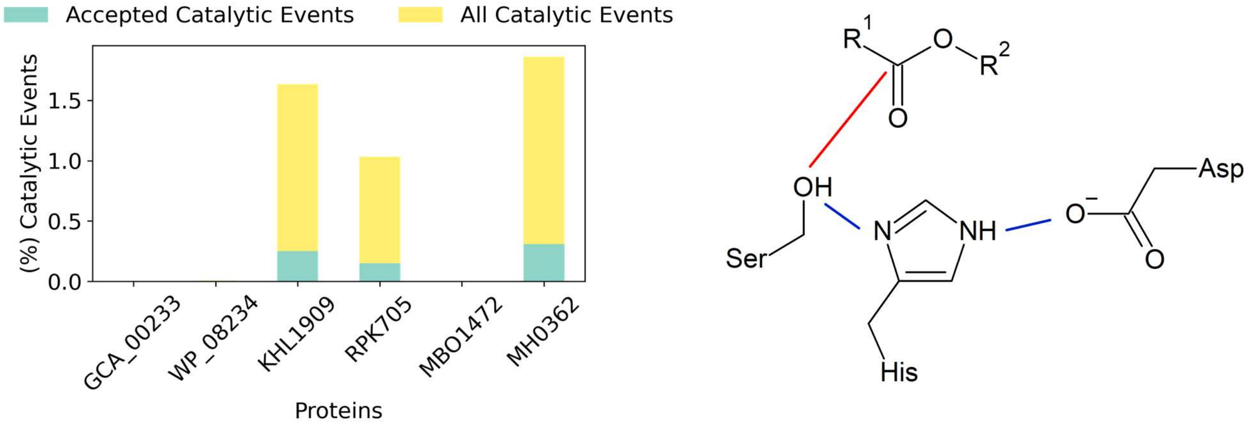


Figure 8. Accepted and total catalytic events for the 6 selected PETases or MHETases (left). In 3 of them the ligand never reaches catalytic positions. Catalytic mechanism for esterases (right).

As detailed in Deliverable D2.1 “Manufacturers' needs and specifications, protocol”, in the case of Henkel’s needs for 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. In the case of lipases, it is to highlight that there are two types of lipases: with and without lid domain (**Figure 9**). Study of lid domain 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. This is why the analysis of the lid domain presence and movement was considered during the analysis (**Figures 10-15**).

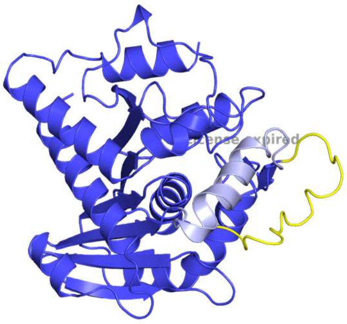


Figure 9. Lipase structure. The lid domain enclosing the catalytic active site is shown in light blue, and the same lid domain in an open conformation is shown in yellow.

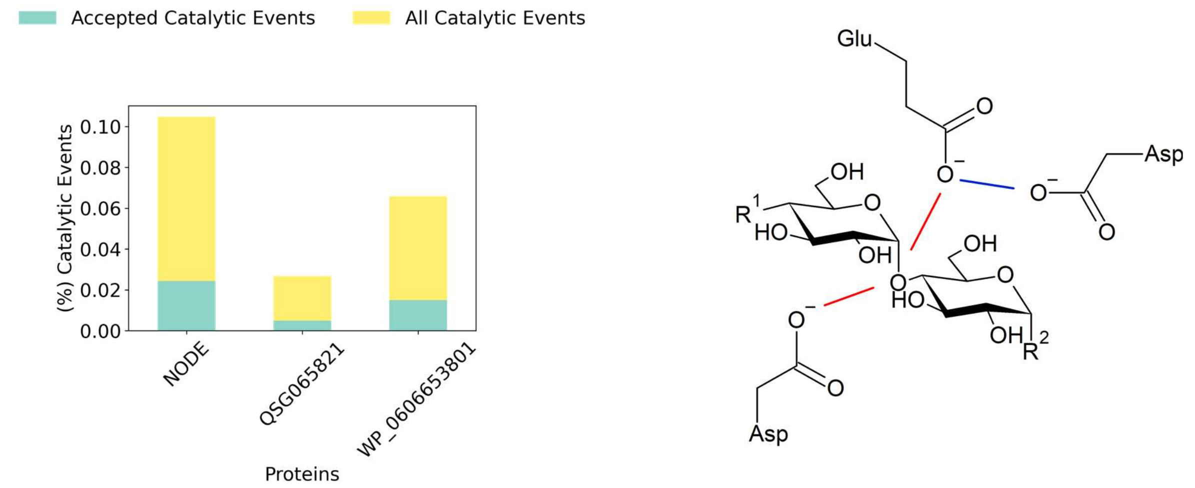
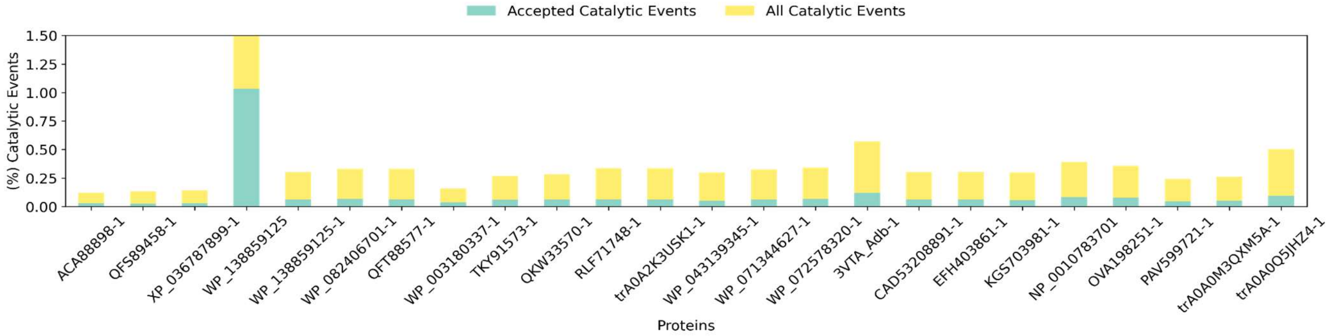
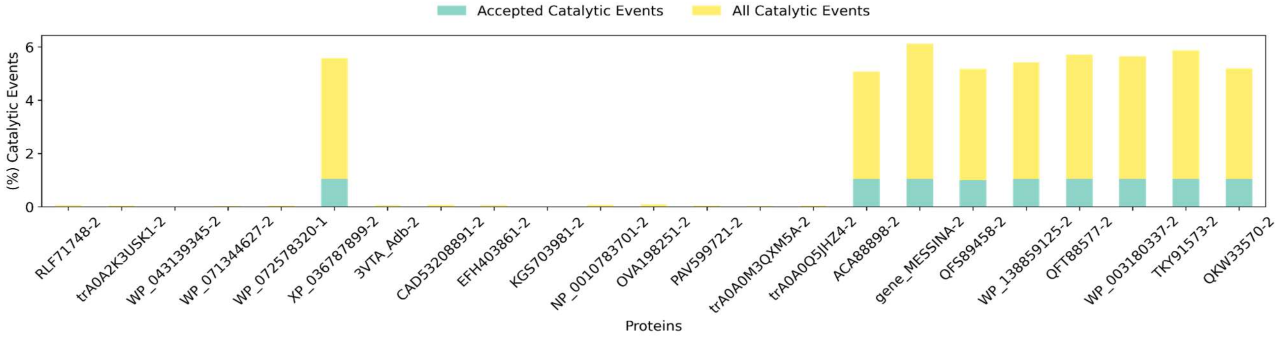


Figure 10. Accepted and all catalytic events for the selected amylases (left). Catalytic mechanism for these enzymes (right).



Nylon -6



PRG

Figure 11. Accepted and all catalytic events for the selected serine proteases against a 6-units nylon ligand (top) and a 7-units polyglycine (bottom).

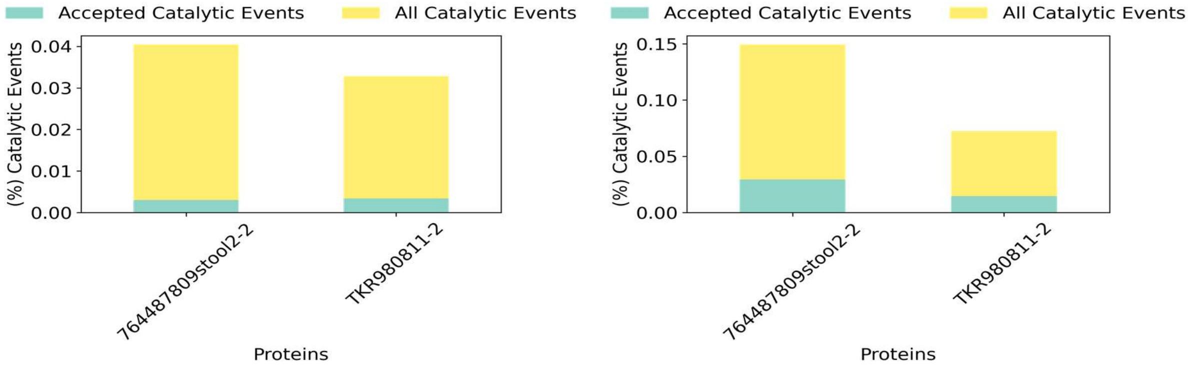


Figure 12. Accepted and all catalytic events for the selected cysteine proteases. From left to right, PELE simulations using 6-units nylon and 7-units polyglycine.

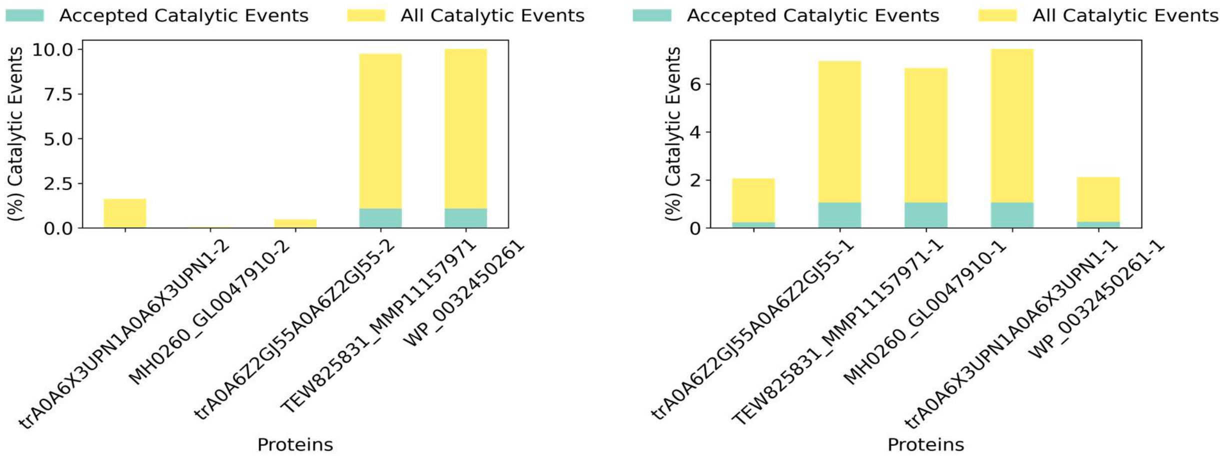


Figure 13. Accepted and all catalytic events for the zinc proteases. From left to right, PELE simulations using 6-units nylon and 7-units polyglycine.

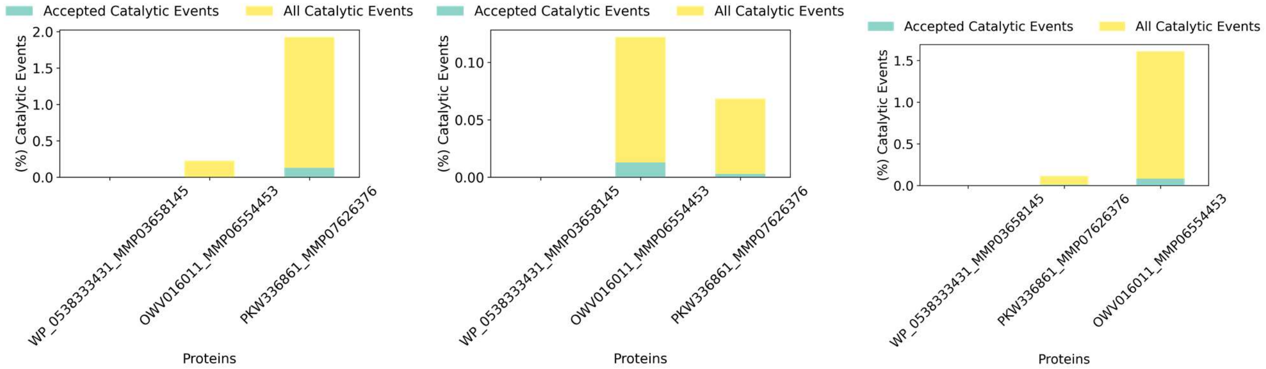


Figure 14. Accepted and all catalytic events for the polymer degrading enzymes. From left to right, the same proteins against polycaprolactone, polylactic acid, and aliphatic polyurethane.

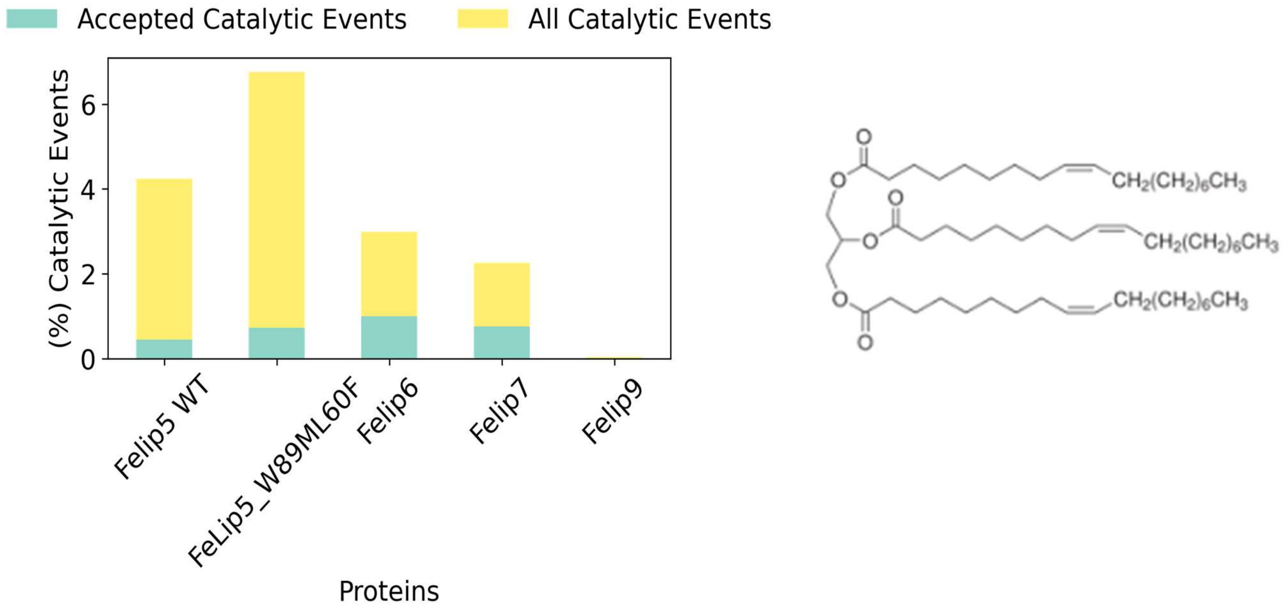


Figure 15. Accepted and all catalytic events for the selected lipases (left). Ligand used for the simulations: triolein (right).

Detailed interaction energy vs catalytic distance serine-substrate plots and violin plots of the distribution of Interaction Energies and catalytic residues-substrate distance along PELE-Induced Fit Simulations for each family of enzymes is given in Annex, **File 1**.

5.2 Pre-selection by PELE of sequences selected by BLASTP

As detailed in the Deliverable D2.2 “Set of 250,000 sequences pre-selected” (November 2021, updated December 2022), a total of 349 sequences (Annex **File 2**) were annotated using BLAST searches of UniProt and the non-redundant GenBank databases using default parameters and they were directly presented as having all proper catalytic domains and catalytic residues, and the presence of signal peptides, by performing the computational pipeline summarized in **Table 1** for the different needs in the project. Among them, hyaluronidases were subjected to PELE simulations, as detailed above. Two hyaluronidases, whose sequences are detailed below, were particularly interested as they did show appropriated number of catalytic events and all catalytic machinery.

Extracellular exopolygalacturonate lyase (PL9).
ID: VD_PL9
MKKHTLALCLAAILAPVAHAAEIKVEDLTWKAITFGQSTDMNFGSTILPEKVGVNQVTVNGEAVAAGKLASTFTIESRGGKL
ANSHEGLTFYYTELPTDVNFTLSADVVLEQLGPETGATPNRQEGAGLMVRDILGAERLVPQPEGHEEFPSASNMVMNLM
RSHTRTNDGMTNINASFREGVYQPWGTPGNRLSRVDYVEGVPYGTAETYRMTLTRTNDGFKVSYRNGEKFIEQAVKGAN
ANIVEMQNSDSQYVGFFASRNAKMTVSNVDLQLAAADTVDAPKYAVKQGELVFKIASSPRSATKEYPVQARANYSGEFEV
LHNDKVVAKQTVTAGDLFSQWLTLDSGANQMEVRFTAIDGPNKETQAHRYSVDVVSLPDPMTLYVAPNGSDKGNGSQA
QPLDLATAVELLPTGGTIILKDGDYQGMEIPLTASGSADKLKHLRAEGDNVRFVSELRHEANYWHYQGIEVAGAQFIVHGS
HNTFEKMVTHGAPDTGFVITSPENIGRALWASYNQVIESESYNNMDPSQINADGFAAKMRIGDGNTFIRCLSHHNIDDG
WDLFNKVEDGANGAVTILDSISFSNGRTLDVANKGGTIGNGFKLGGEGIPVPHVVKNSLSFNNNMDGFTDNFNPGALVLS
DNVSIDNKRFNYLFRKSPYSGEIEQGTFTNNRSYRFHVSSKYDDVINSAKSTGNALVENGTTYTSDGKAVDSKMLAPLKQAS
VIDTQQAIPGKQEAMQLKQLIH
Signal peptide underlined (as seen by http://www.cbs.dtu.dk/services/SignalP/)
Intracellular pectin lyase

ID: VD_PL
MAKGDVITLNFETFVDSDTQVKVTRLTPTDVICHRNYFYQKCFTQDGKKLLFAGDFDGNRNYYLLNLETQQAVQLTEGKGD
NTFGGFISTDERAFFYVKNELNLMKVDLETLEEQVIYTVDEEWKGYGTWVANSDCTKLVGIEILKRDWQPLTSWEKFAEFY
HTNPTCRLIKVDIETGELEVIHQDTAWLGHPIYRPFDDSTVGFCHEGPHDLVDARMWLVNEDGSNVRKIKEHAEGESCTHE
FWIPDGSAMAYVSYFKGQTDRVIYKANPETLENEEVMVMPPCSHLMSNFDGSLMVGDGCDAPVDVADADSYNIENDPF
LYVLNTKAKSAQKLCKHSTSWDVLDGDRQITHPHPSFTPNDDGVLFTSDFEGVPAIYIADVPESYKH
No signal peptide (as seen by http://www.cbs.dtu.dk/services/SignalP/)

Both enzymes belong to a bacterium, *Vibrio diabolicus*, which is able to synthesize a polymer with properties very similar to hyaluronic acid (HE800) [https://link.springer.com/article/10.1007/s00253-014-6086-8, https://www.sciencedirect.com/science/article/pii/S0369811403002839]. Thus, if it generates a polymer similar to hyaluronic acid, it could be that it has a hyaluronic acid degradation system. The two pre-selected genes are part of a gene cluster (01552, 01553, 01554, 01555, 01556). Gene 01552 is a "Sodium/glucose cotransporter" and 01553 is an "Oligogalacturonate-specific porin protein". Thus, these first two proteins would be involved in absorbing the fragments released by the hyaluronate lyase that degrades the polymer (01554). 01555 is another "Sodium/glucose cotransporter" and finally we have 01556 which is the oligogalacturonate lyase, which will be involved in the degradation of the oligomers to the polymer starting materials (probably to re-synthesize it).

Figure 16 illustrates the poses of extracellular hyaluronate lyase and can clearly bind hyaluronan (Glide Scores of ~ -10 to -7 kcal/mol), plus the substrate carboxylate is interacting with the Ca^{2+} of the active centre, as it should. **Figure 17** illustrates the poses of are the poses of intracellular hyaluronate lyase and can clearly bind hyaluronan (Glide Scores of ~ -15 to -10 kcal/mol), plus the carboxylic substrate is interacting with the Mn^{2+} of the active centre, as it should.

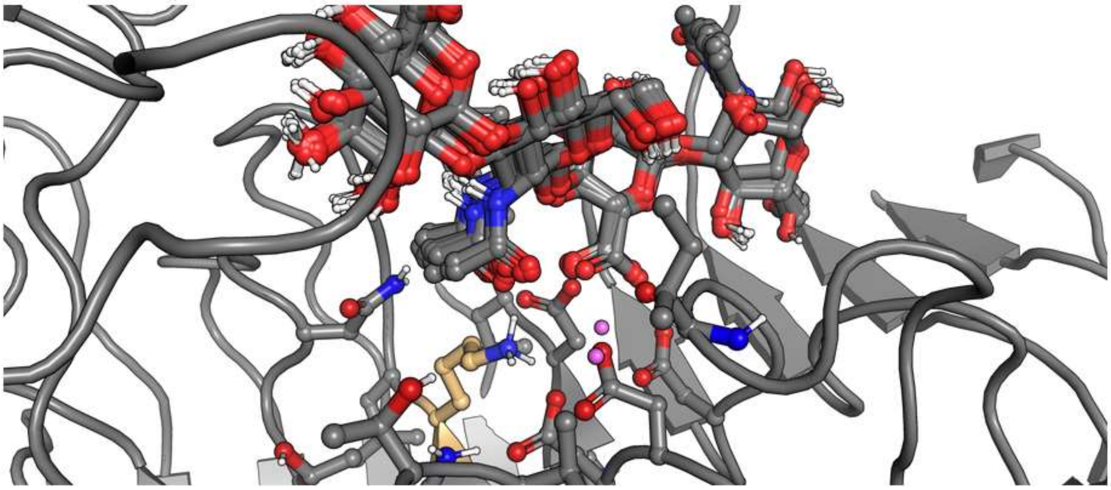


Figure 16. Poses of extracellular hyaluronate lyase with hyaluronan (Glide Scores of ~ -10 to -7 kcal/mol), plus the substrate carboxylate is interacting with the Ca^{2+} of the active centre. The catalytic Lys is coloured yellow.

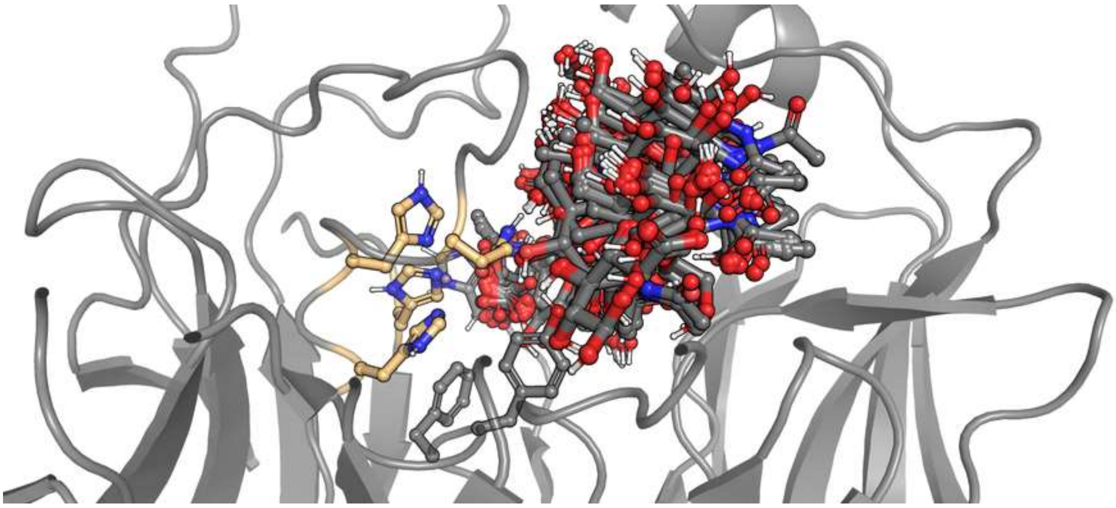


Figure 17 Poses of intracellular hyaluronate lyase with hyaluronan (Glide Scores of ~ -15 to -10 kcal/mol), plus the carboxylic substrate is interacting with the Mn^{2+} of the active centre. The coordination sphere of the catalytic Mn2+ is coloured yellow.

5.3 Pre-selection by the machine learning EP-Pred method

As detailed in the Deliverable D2.2 “Set of 250,000 sequences pre-selected” (November 2021, updated December 2022), a total of 506 esterases and lipases were pre-selected by applying EP-Pred from the Lipase Engineering Database (http://www.led.uni-stuttgart.de/). By applying the methods detailed in Deliverable D2.2, several filters were applied to the model EP-Pred to decrease the number of hits to a final set of about 10 sequences for the experimental validation. Briefly, the top 100 sequences according to E-values returned by HMM were selected to be modelled and their active site cavity analyzed in search of the catalytic triad and geometric descriptors. Only 73 sequences passed this second filter and were forwarded to the subsequent analysis by SiteMap, a widely used binding site analysis tool, which then generated various binding cavity descriptors. As seen in our previous engineering studies, two metrics: hydrophobicity, and the ratio of enclosure/exposure, were useful in ranking promiscuity; thus, we used these to rank the final set of ten proteins for experimental validation picking those that intersected at the top in both metrics (**Figure 5**). A set of sequences encoding esterases or lipases with presumptive substrate promiscuous character of interest for the textile and detergent sectors were pre-selected for experimental validations in WP4 (accession numbers, AJP48854.1, ART39858.1, PHR82761.1, WP_014900537.1, WP_026140314.1, WP_042877612.1, WP_059541090.1, WP_069226497.1, WP_089515094.1).

5.4 Pre-selection by HMMs

As detailed in Deliverable D2.2 “Set of 250,000 sequences pre-selected” (November 2021, updated December 2022), a total of 409 different sequences encoding lipases of relevance for detergent and textile sectors were pre-selected by applying HMMs. After a careful checking for the presence of proper catalytic domains and catalytic residues, and the presence of signal peptides, which is indicative of higher lipase character, and similarity lower than 40% identity to known enzymes (see Deliverable D2.2 for details), two lipases, related to the top priority lipase Lip9 found after experimental tests in WP4, were pre-selected.

> k127_15135326_1
MAGWVAGLACAIAVVSAVDAVAAPKQYPVVMDFATAIAKSAQNPAASPAGVNVPCTLTAEHPRPVVLINGTHASMMM
NWAGLGPTLANQGFCVYSTALGASASDQIQTCGPVADSIAQIASFVDDVLNRTGAQKVDLIGHSQGGLIAESYTKFYGRDK
VANVALLSPSTHGSDQSGTSVHPTDLGAQIASIGCPAVLDQLQSSDVVRELNTGPITVPGVNYTVIETRYEFIITPTPSAAFIQ
EPGVRNLIVQDYCPQDLSDHLSLAYSEPAWNLLIDAISARTGEISC
Metagenome: AGWS_m_17
Source: Wilhelmsburg soil_oil contaminated
35.2% identity Lip9
Triad: Ser141 His262 Glu229
Signal peptide: underlined

> k127_129897_3
MRRCVTVSVILFLAFVMWSGVASAAPTYPVPDSFLAGVPLELGNPGGSAPGSNDWSCVPSDAHPEPVVLVHGTGGARQT
NWAVYAPLLANEGYCVYSLTYGNFPELPWPLDAIGGMTPIDTGTAQIATFVDQVLSSTGASKVDLVGHSQGTLQANNYVK
FFGGADKVSKIVSLAPPWHGTYGNDQISVGRSMRALGIDDEVAAGFPVCGACPEMFQGSAFIDRMRADGVYVPGIEYANI
ATRYDELVVPYTSGIEPGPNTTNIVVQDDCEQDYSDHVAVAGSARAAGFVLNALDPAHPRDVPCRFVAPVAG
Metagenome: AGWS_m_58
Source: Elbe river_enrichment
32.1 % identity Lip9
Triad: Ser148 His276 Asp244
Signal peptide: underlined

In addition, as detailed in the Deliverable D2.2 “Set of 250,000 sequences pre-selected” (November 2021, updated December 2022), a total of 9115 different sequences encoding predicted polyester degrading hydrolases (PETases) were selected. After a careful checking for the presence of proper catalytic domains and catalytic residues, as detailed above, a set of 21 were further selected for in vitro expression and activity assays (see details in Deliverable D4.3 “Cell-free expression reported system developed”).

5.5 Preselection by PSI-BLAST and PELE

As detailed in Deliverable D2.2 “Set of 250,000 sequences pre-selected” (November 2021, updated December 2022), a total of 525 different sequences were pre-selected by applying PSI-BLAST and PELE. By applying stricter filtering protocols, namely, removing redundant sequences from which 3D models by AlphaFold could not be obtained, and similarity lower than 75% identity to known enzymes (see Deliverable D2.2 for details), a final set of 288 new sequences (Annex **File 3**) were pre-selected for in deep analysis in WP4.

5.6 Pre-selection by rigidity analysis

Since changes in protein flexibility may play a role in the thermal adaptation to different temperatures without altering the global structure and the active site, we analysed the flexibility of the esterases and lipases pre-selected. The idea was to preselect those having computed phase transition temperature (Tp) fitting to the thermal conditions requested by manufactures (see D2.1 “Manufacturers' needs and specifications, protocol”). A total of 228 esterases and lipases, relevant to textile and detergent sectors, were pre-selected (Annex **File 4**) after calculation of Tp. In details, biomolecular thermostability can have a thermodynamic or kinetic origin. Thus far, rigidity analysis has been used to investigate structural effects on the folded state only, and it has been estimated that increased structural rigidity is responsible for increased thermostability in 60% of cases. Furthermore, rigidity analysis cannot account for the time-dependency of processes. Constraint network analysis (CNA)-based analyses of the relationship between structural rigidity and flexibility versus thermostability have been applied on pairs and a series of homologous proteins from psychrophilic to hyperthermophilic organisms, as well as on a series of variants from one protein retro- and prospectively. As a result of the analysis, we observed significant correlations between the computed phase transition temperature (Tp), a measure for global structural rigidity, and mean annual temperature (MAT) from where the enzyme was retrieved (Irish Sea–Red Sea transect: R2 = 0.33, p < 0.001, **Figure 18**, left panel. 2e; Tara Ocean: significant regression only after MAT breakpoint of 21.6˚C, R2 = 0.1, p < 0.05, **Figure 18**, right panel) for the two esterase datasets (Irish Sea–Red Sea transect and Tara Ocean). Overall, these findings indicate that esterases and lipases from microorganisms found in environments with higher MAT have evolved so that their esterases and lipases are more rigid (less flexible). This might mirror the principle of corresponding states, according to which homologs from mesophilic and thermophilic organisms have similar flexibility and rigidity characteristics at their respective growth temperatures. The lack of regression observed in the phase transition analysis of esterase at environmental temperatures below 22˚C could represent the onset of evolutionary trade-offs that may occur during biochemical adaptation to lower temperatures where enzymes have to keep a minimum of rigidity for correct functioning while those adapted to higher temperatures can increase it to cope with the higher metabolic requirements of the organisms. Taken together, a total of 228 esterases and lipases were selected based on the proximity of the thermal conditions requested by manufactures and the MAT of the site from where the enzyme was retrieved.

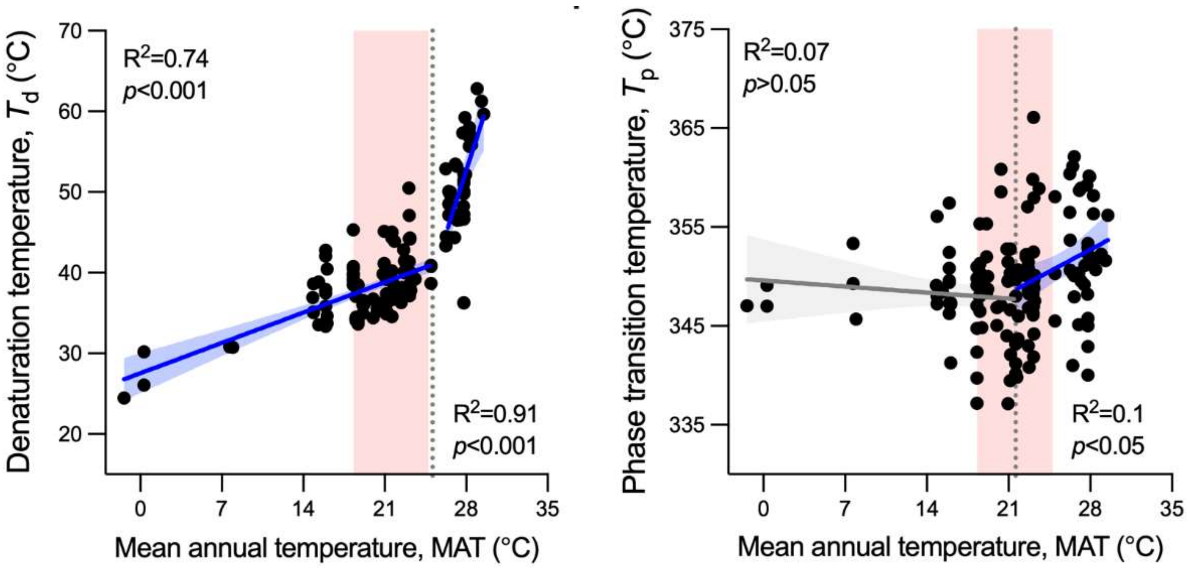


Figure 18. Thermal adaptation of pre-selected enzymes. Left panel, phase transition temperature (Tp) patterns as a function of the MAT at the site from which the esterases and lipases originated along the north–south longitudinal transect. Right panel, Tp patterns of 150 pre-selected esterases and lipases from Tara ocean locations as a function of the MAT at the site. R2 and p-values of regressions are reported in each graph and the blue zone represents the confidence value of 95%. In the case of esterases from the Tara ocean dataset, piecewise regressions were run and the breakpoints (flexus) where the slope of the regressions significantly changed are indicated with dashed lines on the MAT axis. Non-significant regression is reported in grey.

5.7 Preselection by DDG value

As mentioned above, it is important to pre-select sequences encoding enzymes with thermal characteristics similar to the thermal conditions requested by manufactures (see D2.1 "Manufacturers' needs and specifications, protocol"). In this line the objective was to develop a predictive tool that allow such a pre-selection. Note that the aplication of a computational method to guide the screening of stabilizing amino acids or mutations greatly reduces the experimental time and cost required for experimental effort. Following on from that we applied DDG calculations to all esterases and lipases pre-selected in Deliverable D2.2 "Set of 250,000 sequences pre-selected" (November 2021, updated December 2022), from which one esterase/lipase, namely, EH$_{37}$ (Protein data Bank acc. nr. 5JD5), was pre-selected for in deep characterization in WP5. In brief, a total of 100 mutations were selected on the basis of DDG values, from which ten were selected as priority to find correlations between DDG values and melting temperature or denaturing temperature (T_d) of the mutants (**Table 2**). As will be detailed in WP5, the predicted changes in stability for a prospective esterase system (PDB ID: 5JD5) correlates with the observed denaturation temperature, T_d (˚C), as can be seen in the **Figure 19**. Thus, the DDG values could be for pre-selecting sequences encoding enzymes with different thermal behaviour.

Table 2. List of the esterase / lipase preselected by DDG calculations, and the mutants selected.

Predictor	Mutation	DDG value	Td (˚C)
5JD5	None	None	35.99 ± 0.19
AL	36F	-1.593600001	39.36 ± 0.33
AL	36H	-0.902499993	37.99 ± 0.29
AL	62I	-0.844500003	31.77 ± 1.12
AL	87G	-0.791399995	35.97 ± 0.44
AL	242G	-0.769099992	32.11 ± 0.85
AL	41W	-0.8463	34.59 ± 0.23
AL	214I	-0.885600004	34.45 ± 0.31
AL	36M	-0.680699998	38.16 ± 0.45
AL	294S	-0.426099996	37.39 ± 0.39
AL	36E	-0.378099997	38.26 ± 0.52

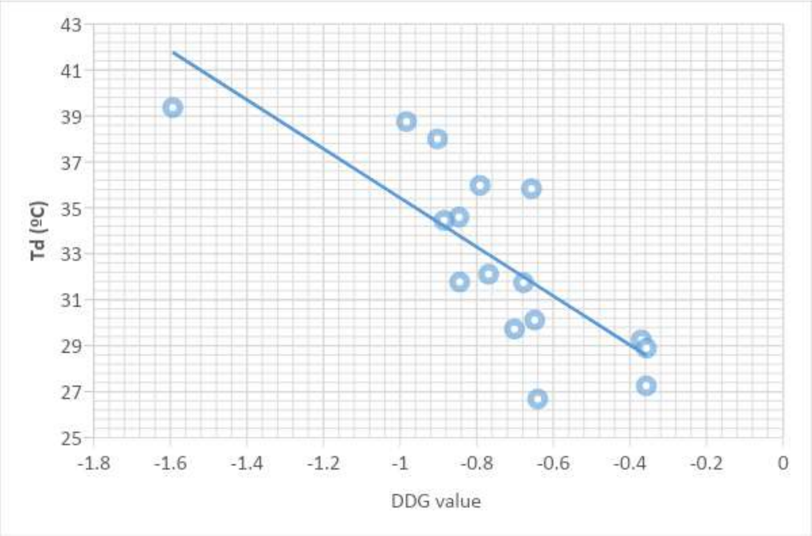


Figure 18. Correlation between DDG predicted changes and experimental Td.

6. Conclusions and outlook

As a result of the activities done to achieve Deliverable D2.2 "Set of 250,000 sequences pre-selected" (November 2021, updated December 2022), a total of approx. 3.16 million sequences encoding target enzymes were retrieved and pre-selected. In Deliverable 2.3, a number of bioinformatics and computational tools were applied that allow the pre-selection of approx. 1355 sequences encoding enzymes relevant to FuturEnzyme, which were transferred to WP4. Note that these pre-selected sequences do not account those retrieved after functional screens in WP3.

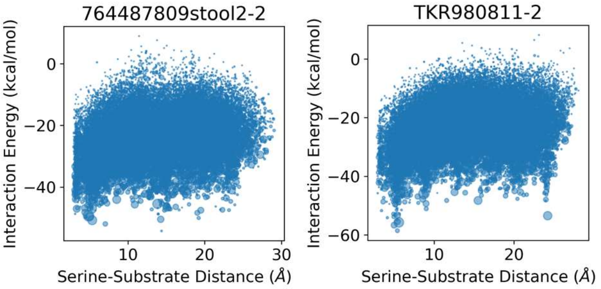
Annex

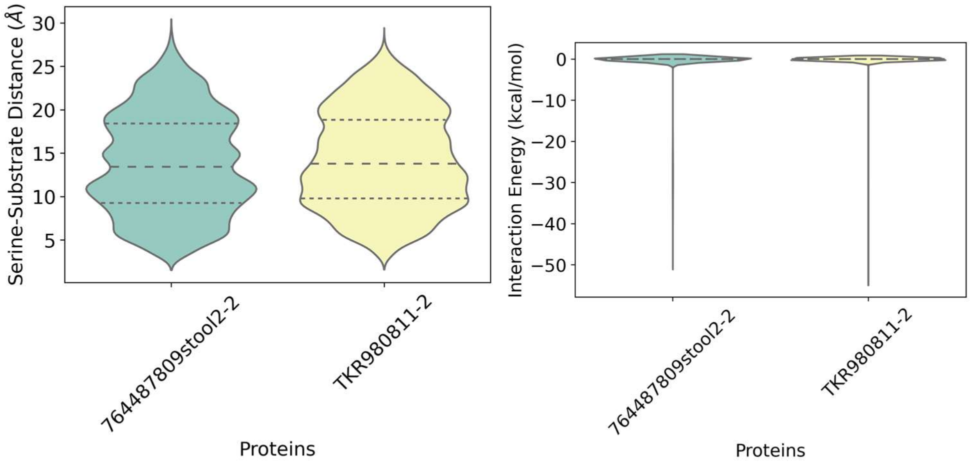
Because of their extensive size, the following Annex files are provided in a separate ZIP file:
See intranet's project website File 1 (D2_3) in www.futurenzyme.eu -> login -> private-area -> shared-data.

- **Annex File 1_Network Analysis Enzymes_Preselected**
  List of pre-selected sequences encoding enzymes constituting each of the networks identified per enzyme family using network analysis followed by computational (PELE-based) tools. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin. SAME AS IN DELIVERABLE D2.2 (see Annex File 3_Network Analysis Enzymes).
- **Annex File 2_ BLASTP_PELE _Preselected**
  List of pre-selected sequences encoding enzymes retrieved by BLASTP analysis followed by computational (PELE-based) tools. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin. SAME AS IN DELIVERABLE D2.2 (see Annex File 4_ BLAST_Results).
- **Annex File 3_HMMs_Preselected**
  List of pre-selected sequences encoding enzymes most similar to lipase Lip9, through PSI-BLAST followed by computational (PELE-based) tools. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin. SAME AS IN DELIVERABLE D2.2 (see Annex File 7_ PSI-BLAST and PELE_Results_Lip9).
- **Annex File 4_Rigidity_Preselected**
  List of 228 sequences encoding esterases and lipases pre-selected by rigidity analysis.

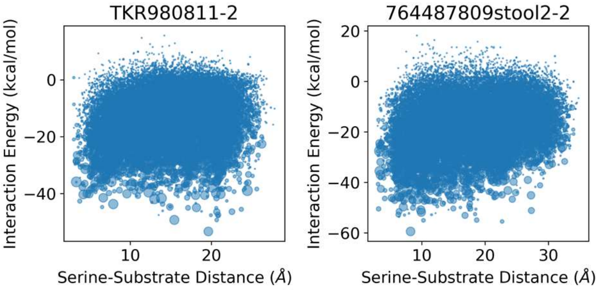
Annex Figure 1. Interaction Energy vs Catalytic Distance Serine-Substrate Plots and Violin Plots of the distribution of Interaction Energies and catalytic residue-substrate distance along PELE-Induced Fit Simulations for each family of enzymes. The size of the scatter dots represents the rejected PELE steps that were considered as a time of residence of substrate in the position.

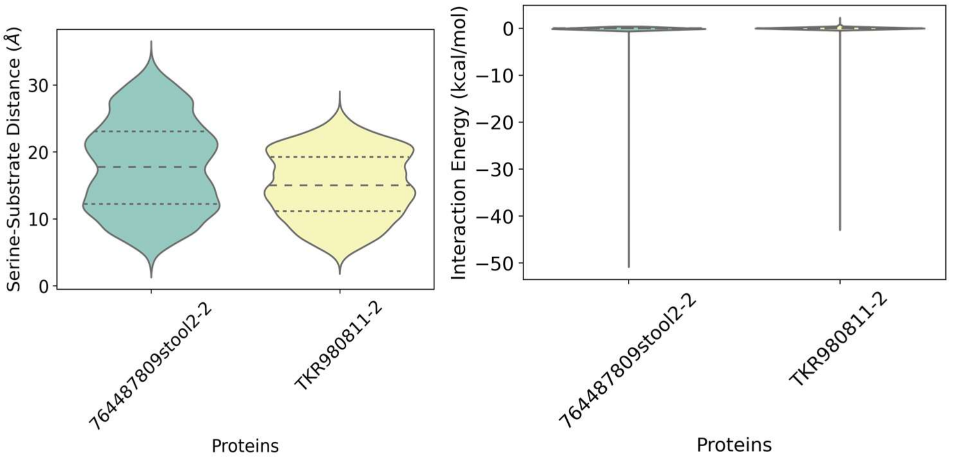
Cysteine Proteases - NY6:



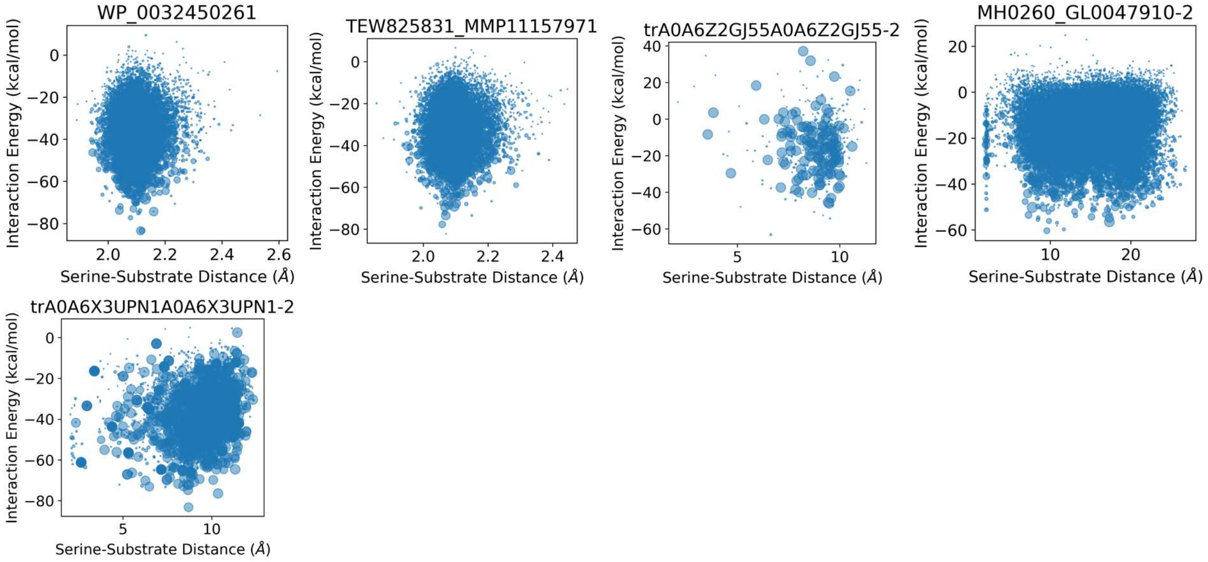


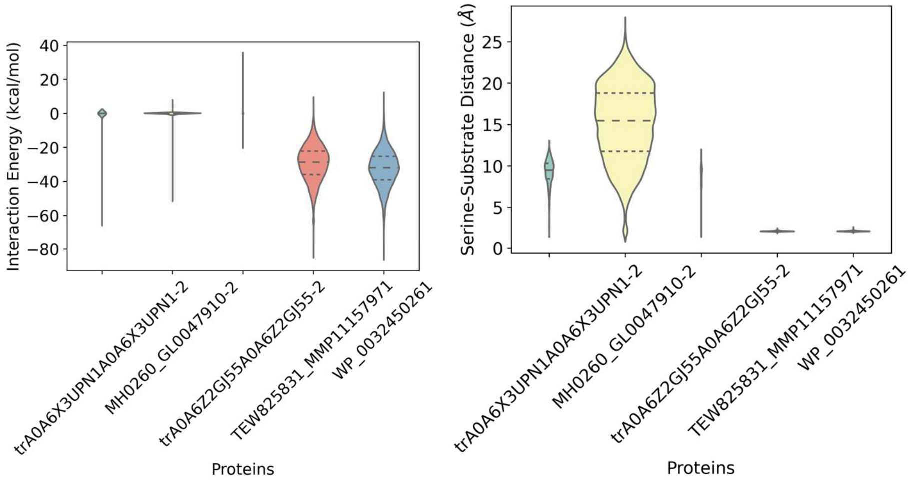
Cysteine Proteases - PRG:



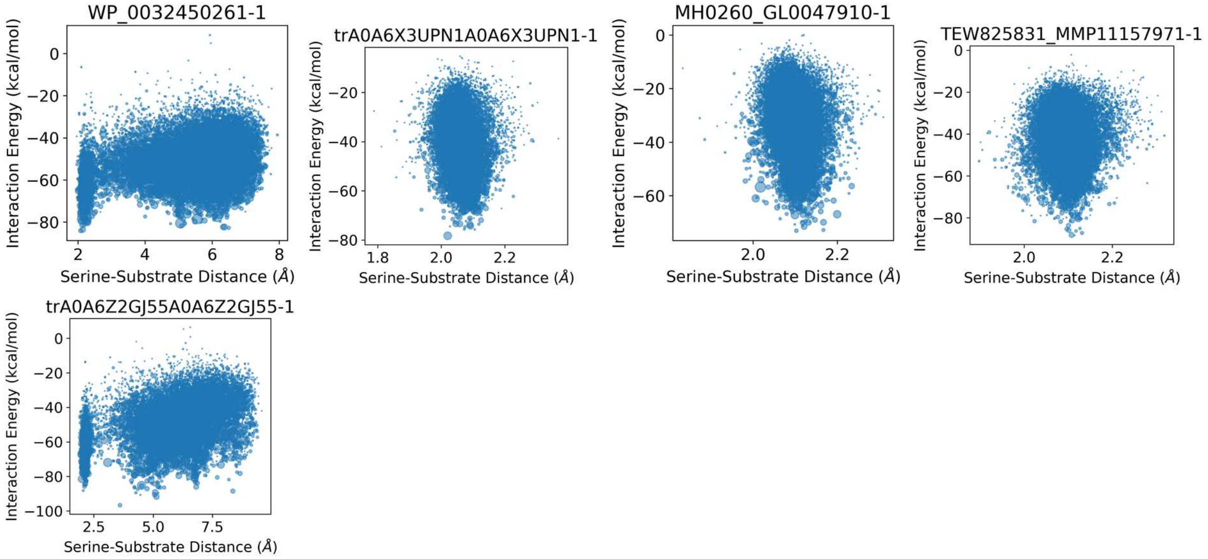


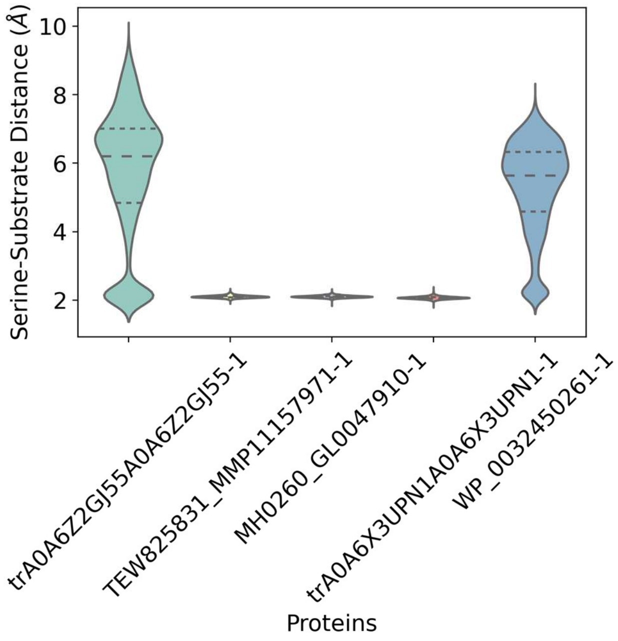
Zinc Proteases - NY6:

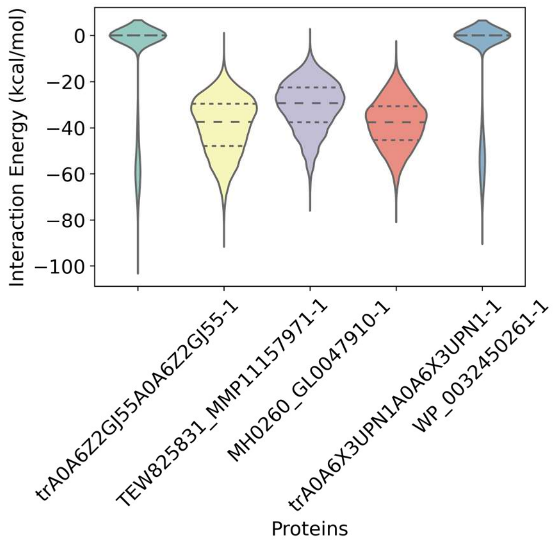




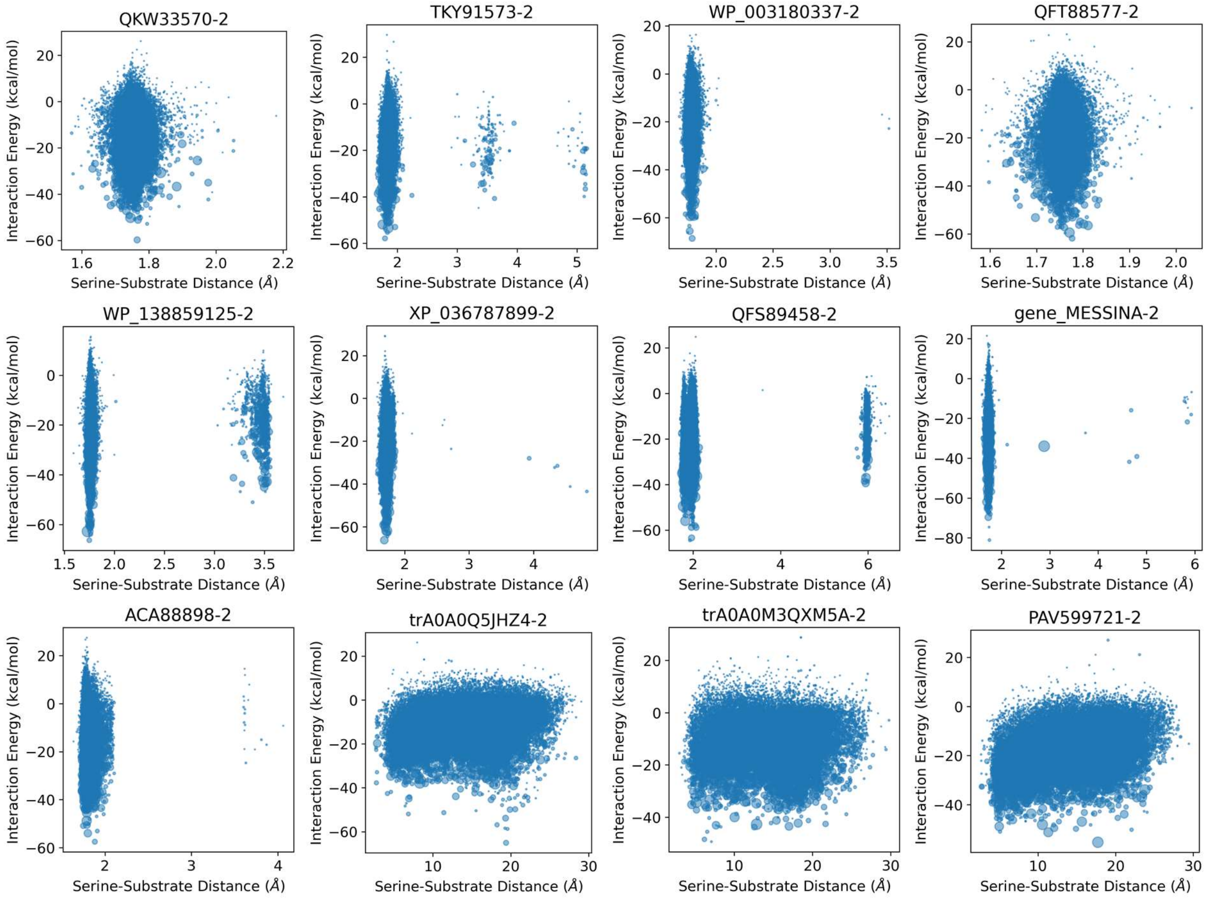
Zinc Proteases PRG:

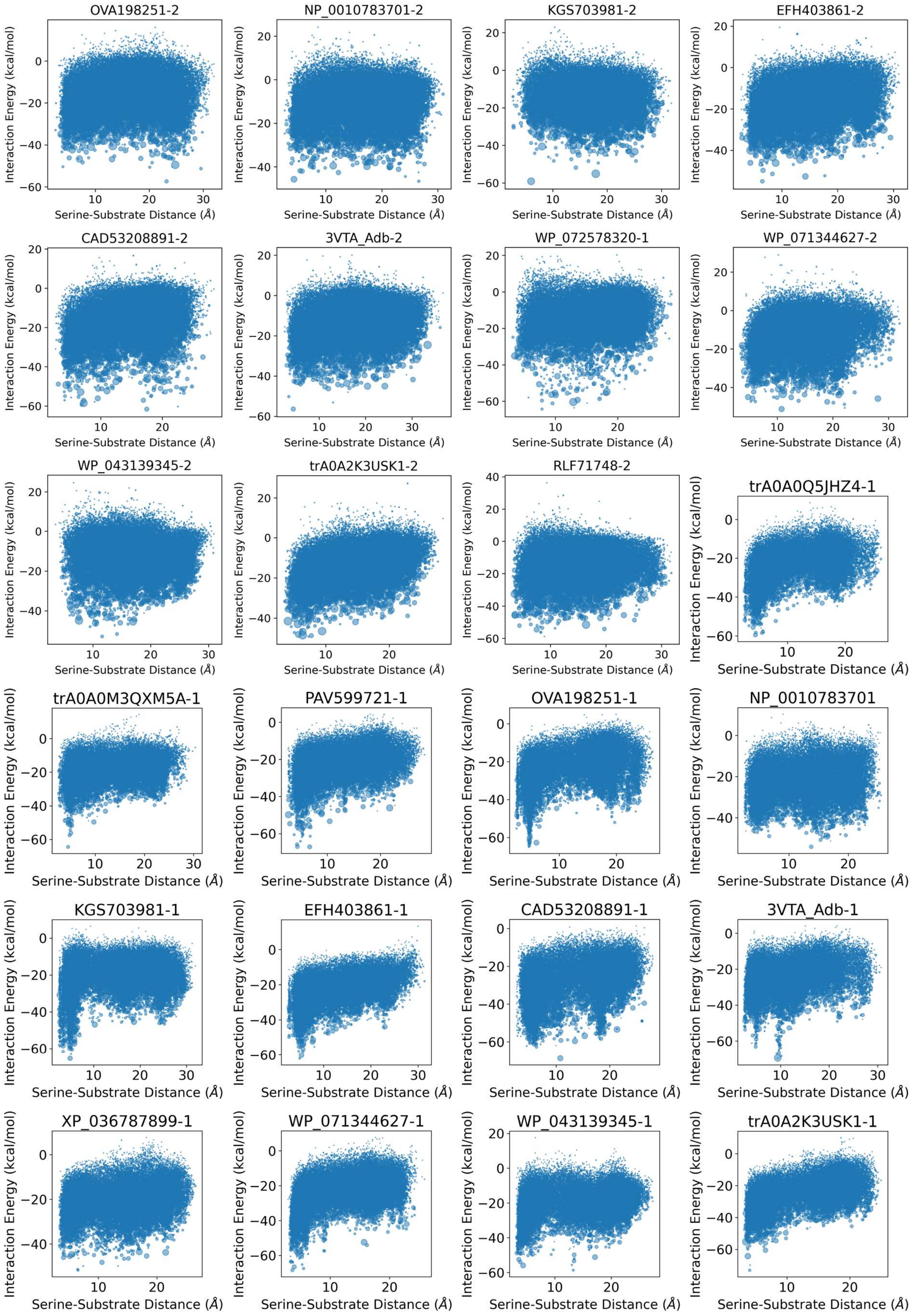


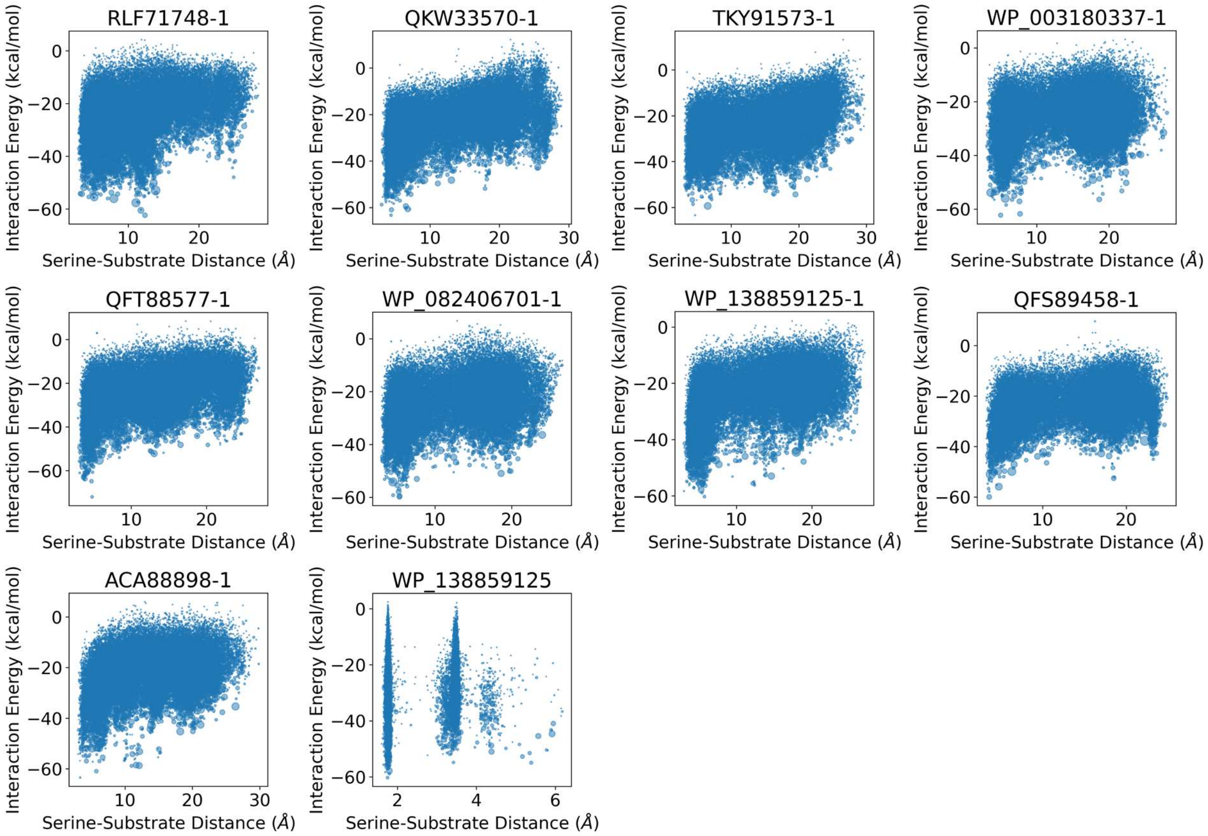


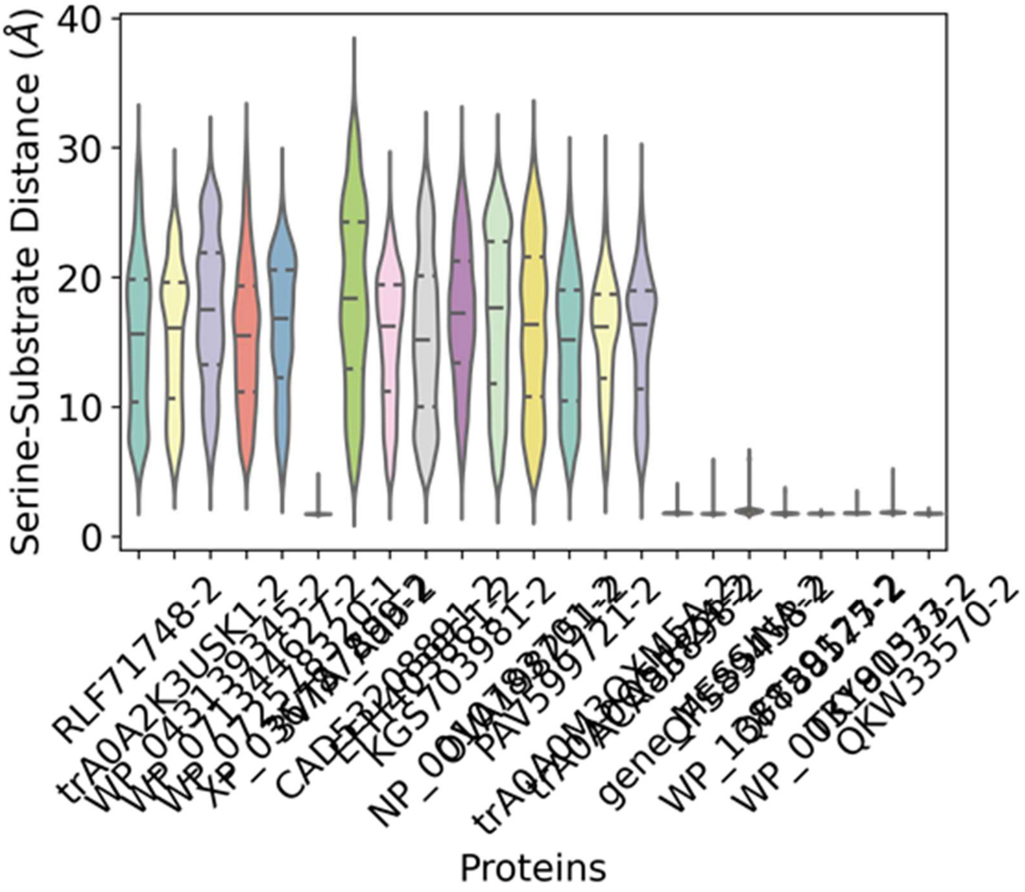


Serine Proteases:

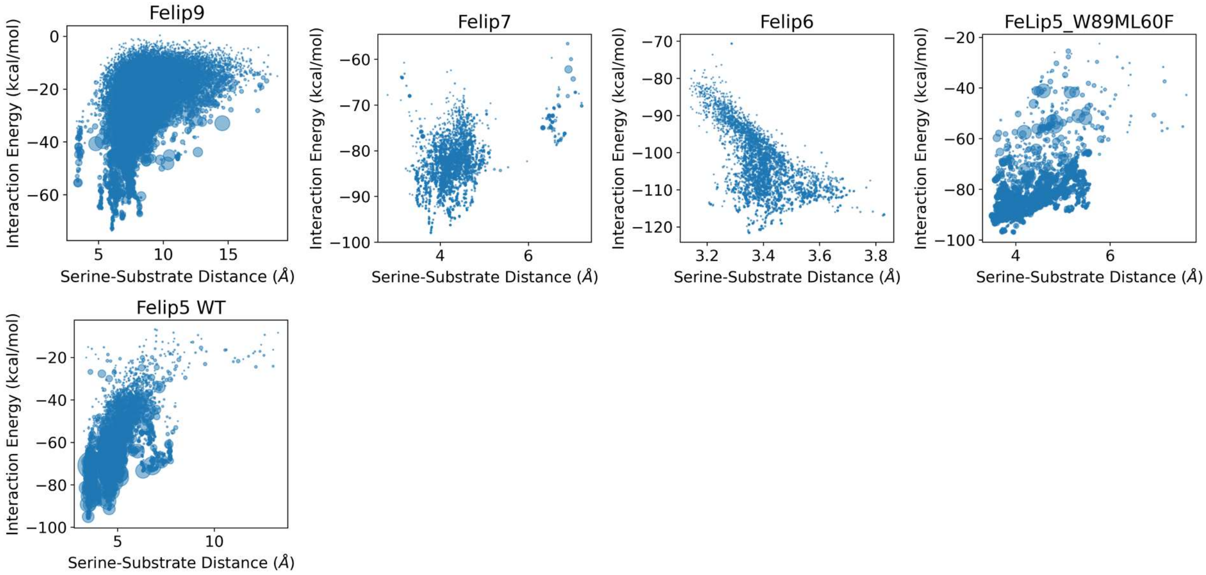


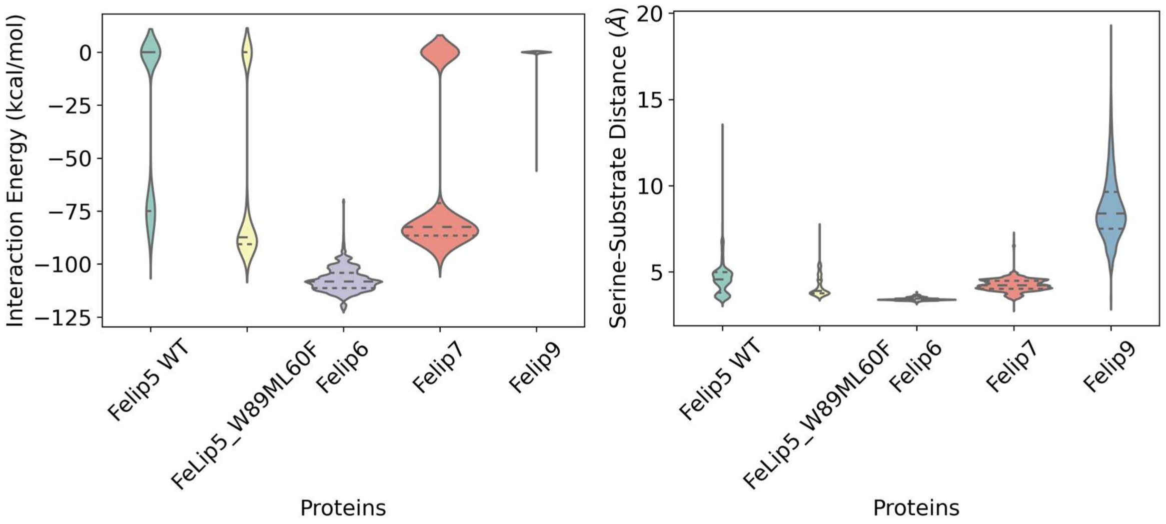

OVA198251-2
NP_0010783701-2
KGS703981-2
EFH403861-2
CAD53208891-2
3VTA_Adb-2
WP_072578320-1
WP_071344627-2
WP_043139345-2
trA0A2K3USK1-2
RLF71748-2
trA0A0Q5JHZ4-1
trA0A0M3QXM5A-1
PAV599721-1
OVA198251-1
NP_0010783701
KGS703981-1
EFH403861-1
CAD53208891-1
3VTA_Adb-1
XP_036787899-1
WP_071344627-1
WP_043139345-1
trA0A2K3USK1-1
Interaction Energy (kcal/mol)
Serine-Substrate Distance (Å)


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)
0
−20
−40
−60
2
4
6
10
20
30

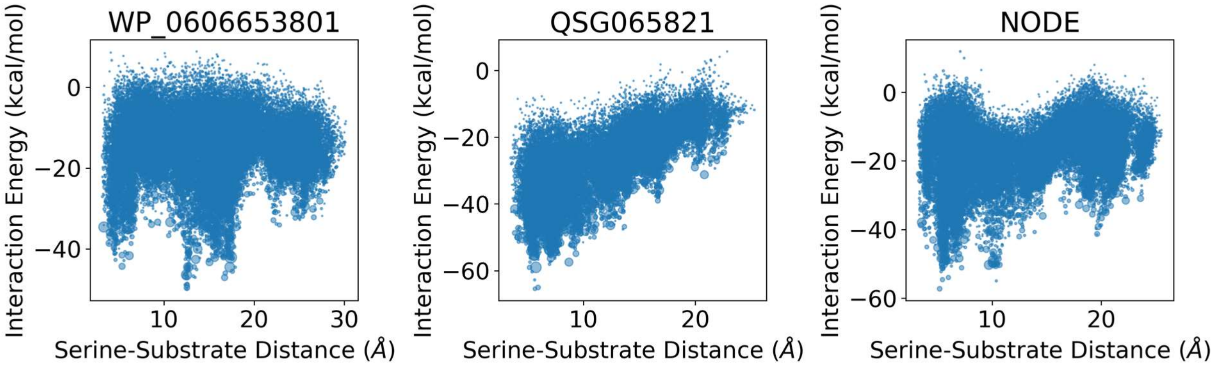

Serine-Substrate Distance (Å)
0
10
20
30
40
RLF71748-2
trA0A2K3USK1-2
KGS703981-2
QKW33570-2
Proteins

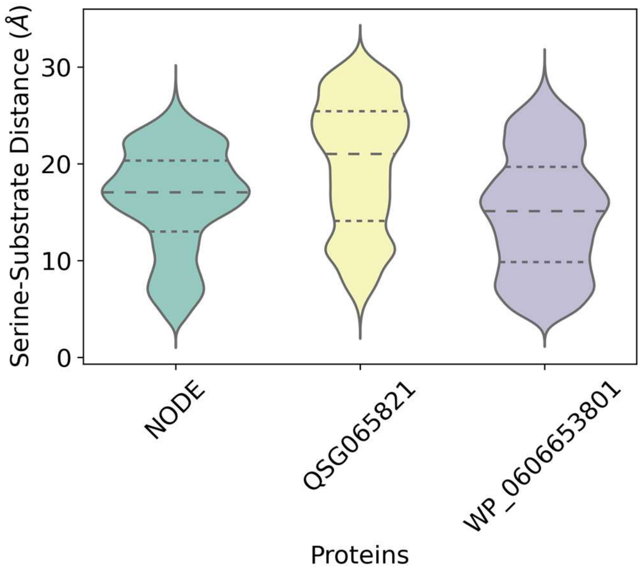
Lipases:

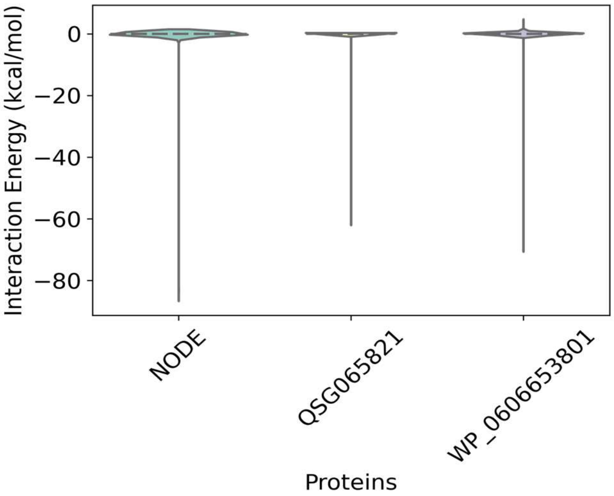




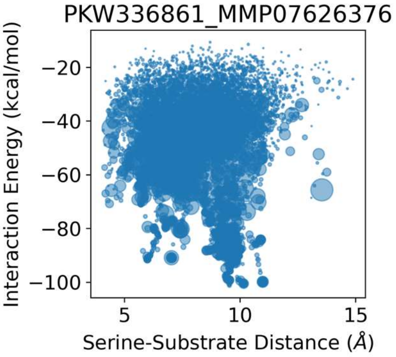
Amylases:

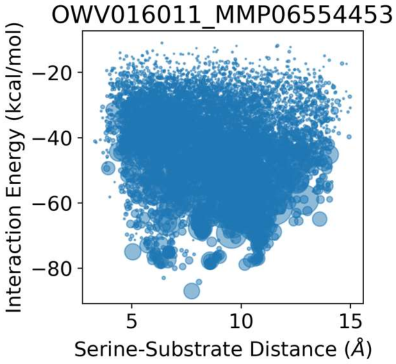


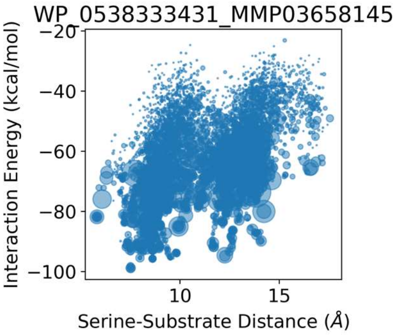


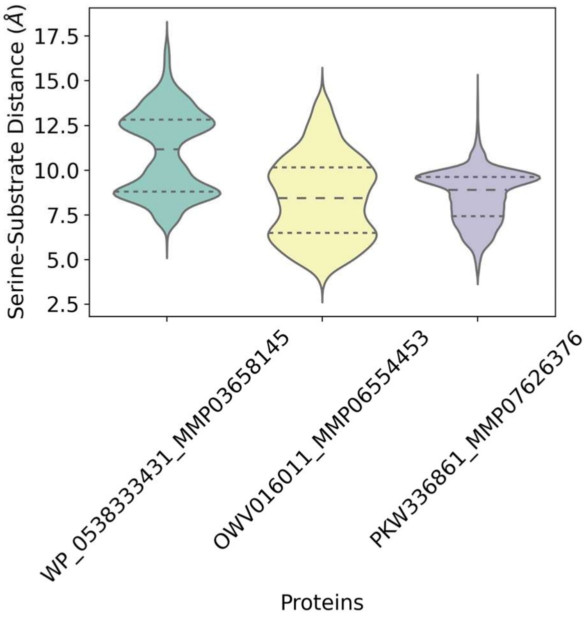


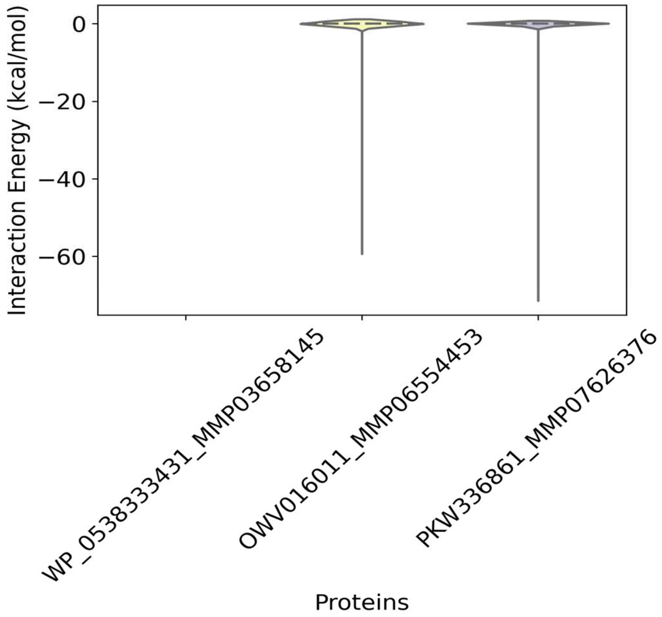
Polylactic acid:



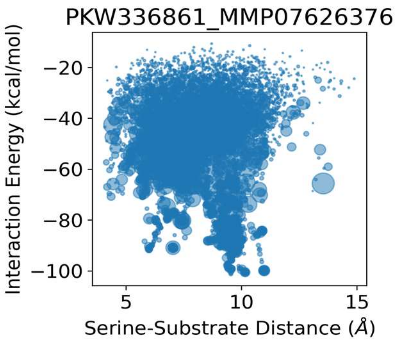


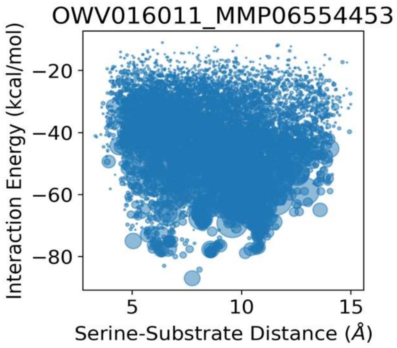


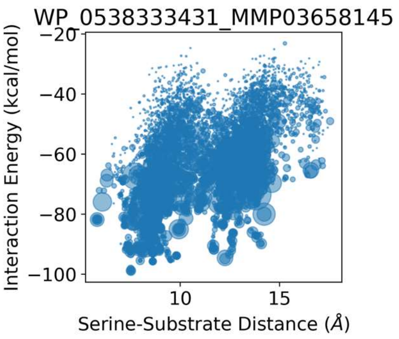


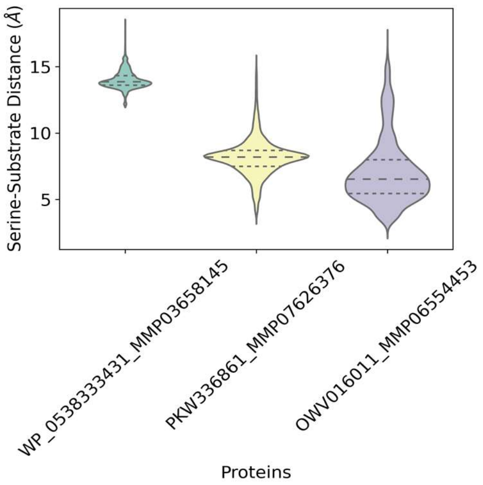


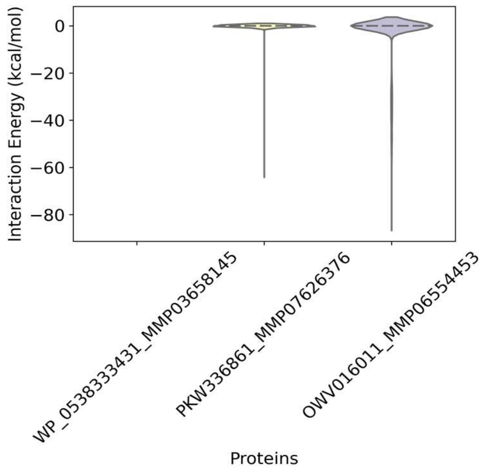
Aliphatic polyurethane:



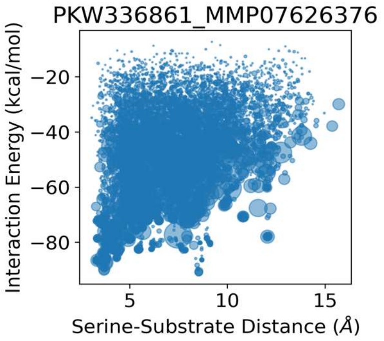


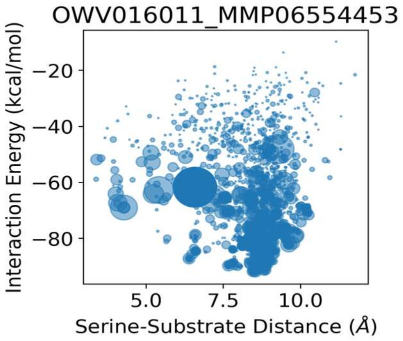


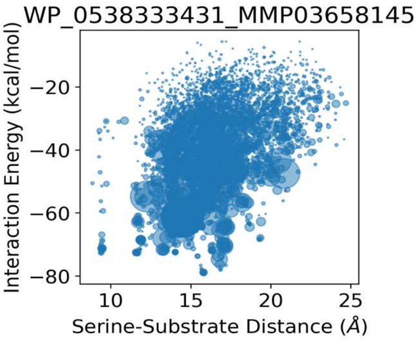


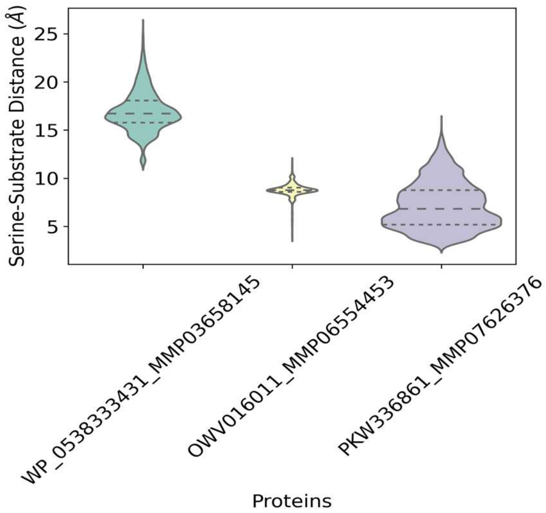


Polycarpolactone:

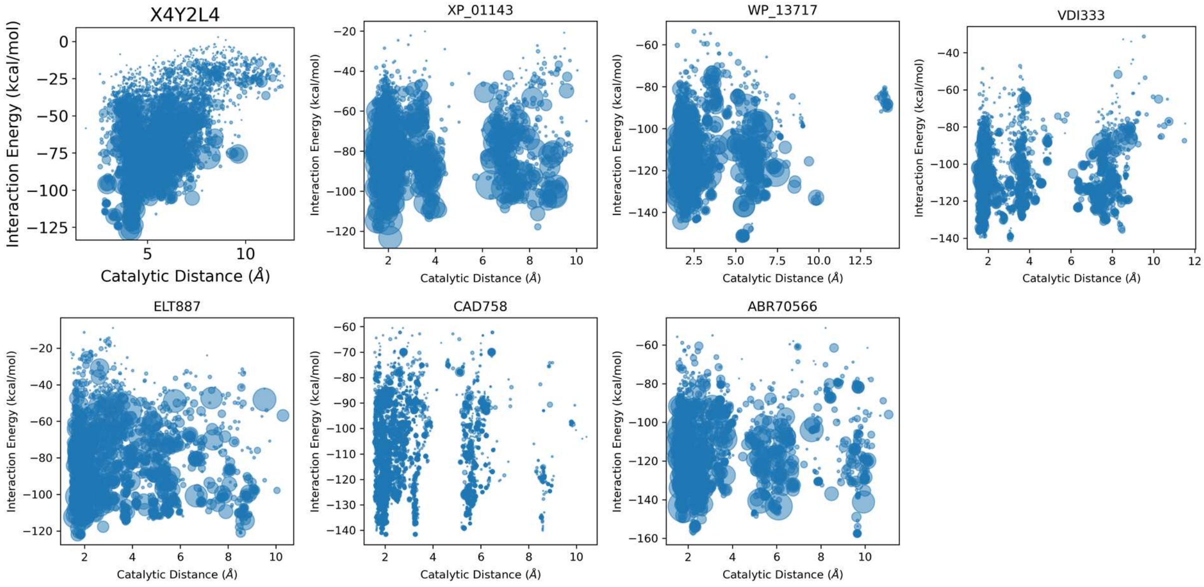


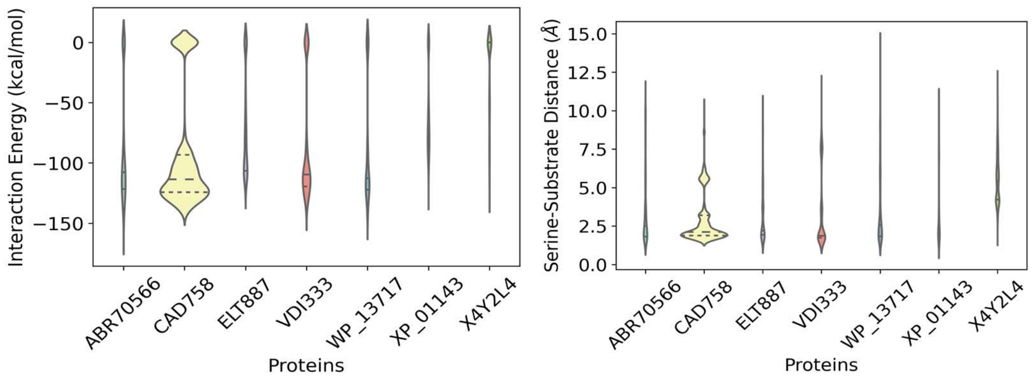




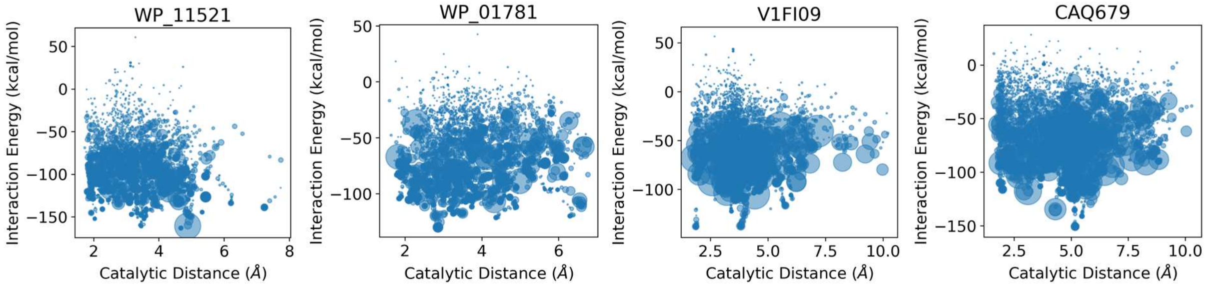


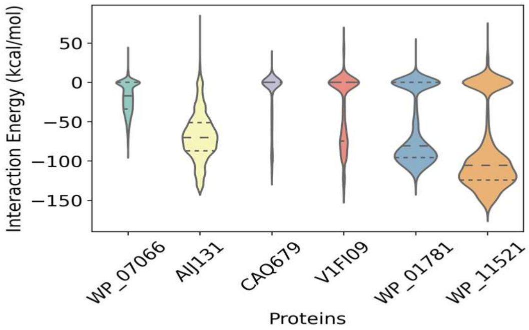
Hyaluronoglucuronidases:





Hyaluronate Lyase:


WP_11521
WP_01781
V1FI09
CAQ679
Interaction Energy (kcal/mol)
Catalytic Distance (Å)


Interaction Energy (kcal/mol)
WP_07066
AIJ131
CAQ679
V1FI09
WP_01781
WP_11521
Proteins