WP2 – Machine learning enzyme bio-prospecting integrated into an industrial context

FuturEnzyme Technologies of the FUTURe for low-cost ENZYMEs for environment-friendly products

FuturEnzyme: 2nd annual meeting
Start date: 1 June 2021 - End date: 31 May 2025
Proposal number: 101000327 - Consortium: 16 partners
Requested EU Contribution: 5,995,035.13 €

Project funded by the European Union’s Horizon 2020
Research and Innovation Programme under grant agreement No [101000327]

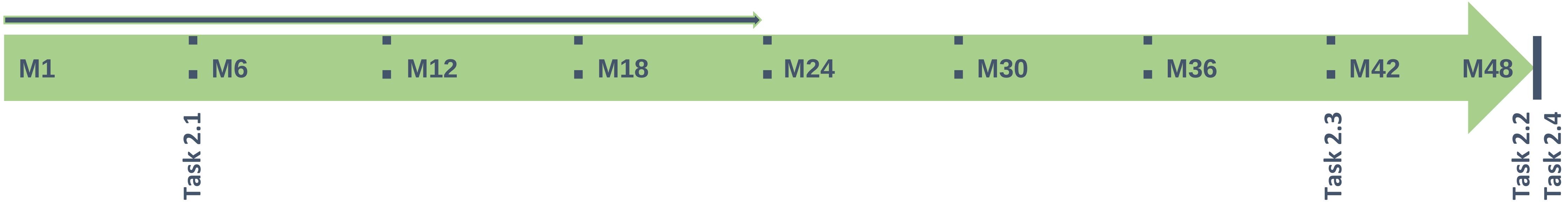
WP2 - Machine learning enzyme bio-prospecting integrated into an industrial context

- OBJECTIVE

To pre-select **enzymes meeting products' requirements by bioinformatics and supercomputing pipelines:**

- Public and consortium **sequence repositories**
- Knowledge of the **needs and requirements** of manufacturing companies
- **Meta-data analysis**

- TASKS
  - Compile the on-demand manufacturers' needs and specifications (M1 - M6) (TASK 2.1)
  - Pre-selecting candidate sequences through extensive homology search (M1 - M48) (TASK 2.2)
  - Motif buildup for massive and smart search of enzymes fitting manufacturers' needs (M1 - M42) (TASK 2.3)
  - Iterative and decision-making hierarchical procedure for speed up enzyme discovery (M3 - M48) (TASK 2.4)



M1 M6 M12 M18 M24 M30 M36 M42 M48

Task 2.1

Task 2.3

Task 2.2

Task 2.4

WP2 - Partners involved

WP2 lead

BSC, Barcelona Supercomputing Center (11.45/32 PM)

WP2 contributing partners

CSIC, Agencia Estatal Consejo Superior de Investigaciones Científicas (1.61/3 PM)

UDUS, Heinrich-Heine Universitaet Duesseldorf (0.68/1 PM)

UHAM, Universitaet Hamburg (0/6 PM)

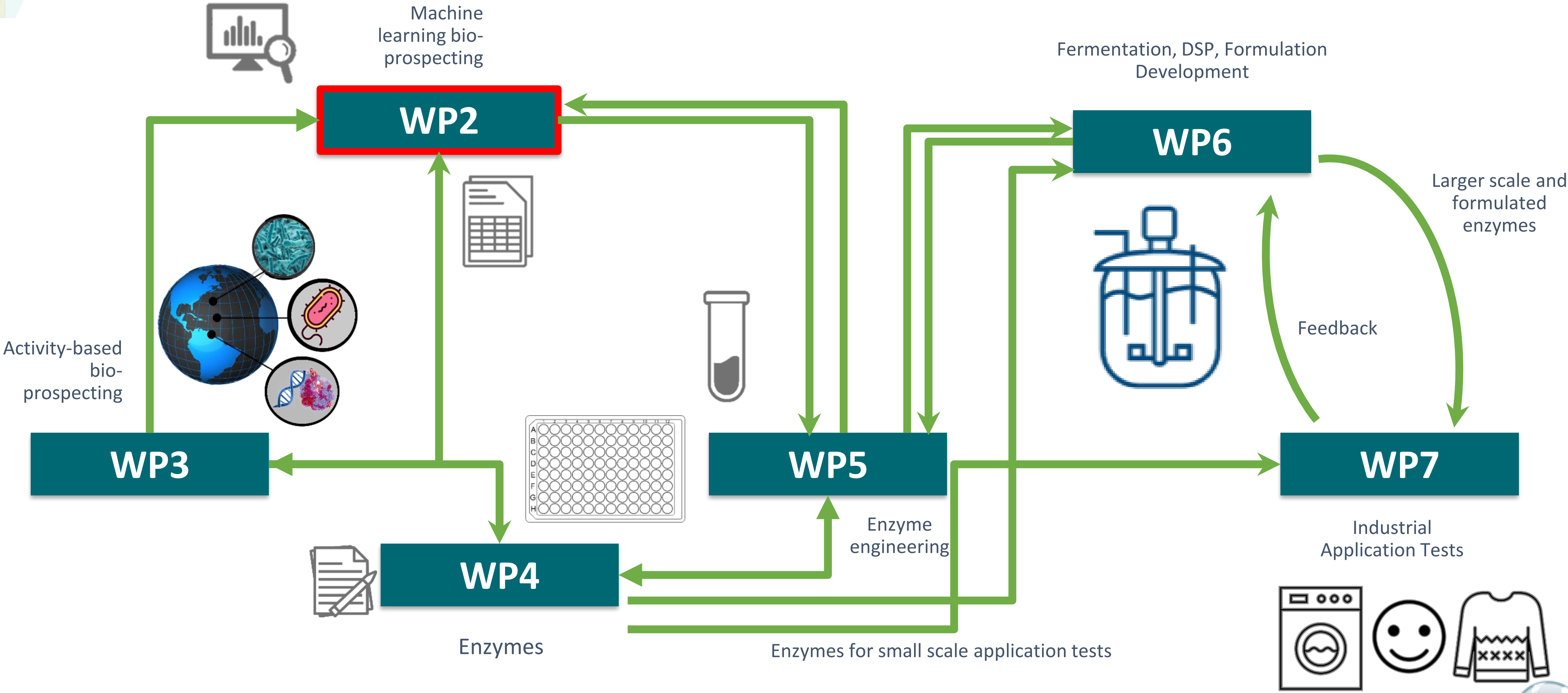
BANGOR, Bangor University (0/2 PM)

EVO, Evonik Operations GMBH (0.2/1 PM)

HENKEL, Henkel AG & Co KGaA (1.94/2 PM)

SCHOELLER, Schoeller Textil AG (0.74/4 PM)

WP2 - Interactions


Machine learning bio-prospecting
Fermentation, DSP, Formulation Development
WP2
WP6
Larger scale and formulated enzymes
Feedback
Activity-based bio-prospecting
WP3
WP5
WP7
Enzyme engineering
Industrial Application Tests
WP4
Enzymes
Enzymes for small scale application tests
A
B
C
D
E
F
G
H
xxxx

WP2 - Machine learning enzyme bio-prospecting

- WORK DONE M1-18
  - Task 2.1: Detailing the manufacturers' needs, specifications and priorities and a state-of-art analysis:
    1) Products, requests and innovations
    2) Priority enzymes to be targeted
    3) Specifications that enzymes should meet
    4) Decision taken strategies

COMPLETED

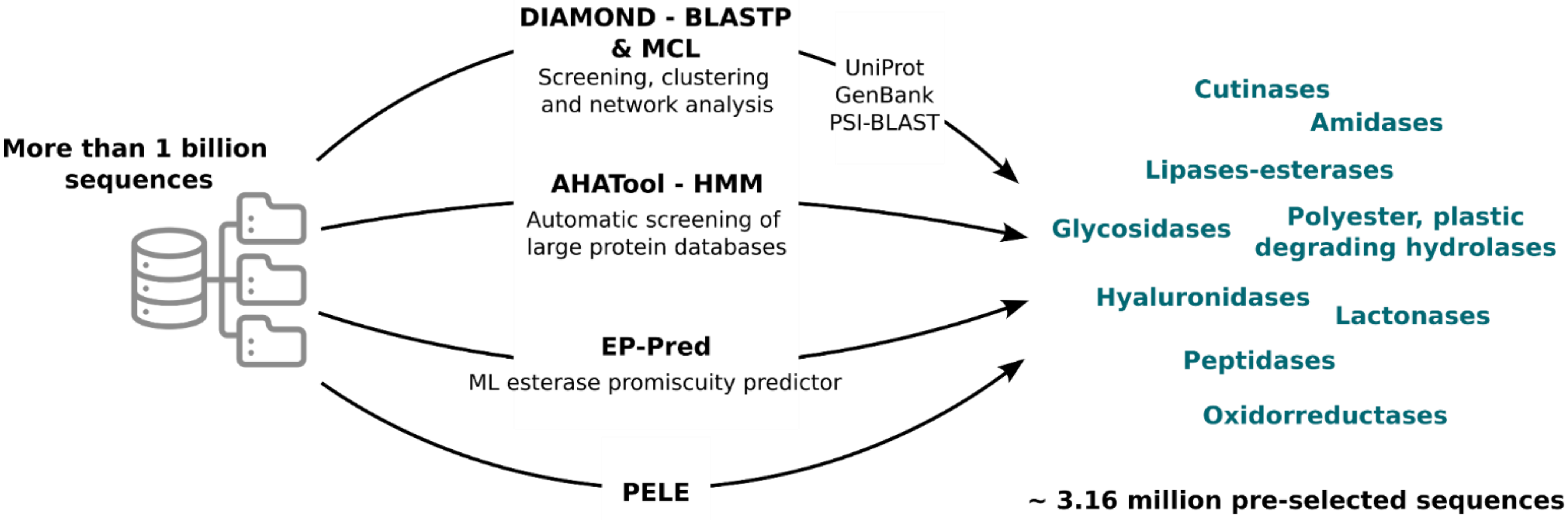
  - Task 2.2: Implementing and using at least five bioinformatics and computational methods to bio-prospect for (+250,000), and pre-select the target enzymes (+1,000) from:
    1) +900 Giga-bytes and +400 million sequences generated in the project
    2) +1 Billion sequences in public repositories

  - Tasks 2.3-2.4: Development of novel algorithms and biocontainers for enzyme bio-prospecting through:
    1) Integrating experimental meta-data (WP4 & WP5) and motif buildup to search for enzymes fitting manufacturers' needs
    2) Establishing novel consensus machine learning predictors and core software

Task 2.2: Explanation of the work carried

Progress undertaken and outputs achieved M1-M18

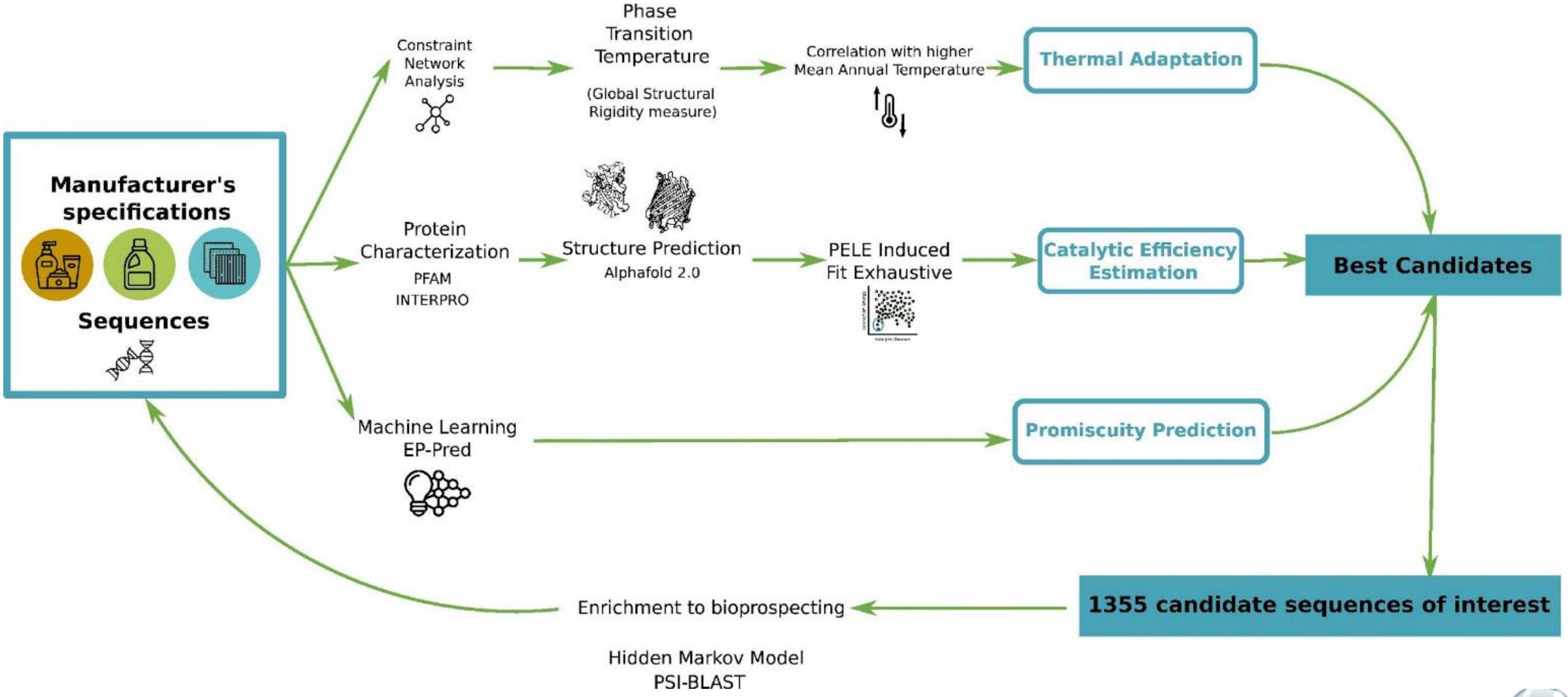
- Pre-selecting candidate sequences through extensive homology search
  - After screening more than 1 billion sequences, about 3.16 million sequences encoding target enzymes were retrieved and pre-selected.



More than 1 billion sequences
DIAMOND - BLASTP & MCL
Screening, clustering and network analysis
UniProt
GenBank
PSI-BLAST
AHATool - HMM
Automatic screening of large protein databases
EP-Pred
ML esterase promiscuity predictor
PELE
Cutinases
Amidases
Lipases-esterases
Glycosidases
Polyester, plastic degrading hydrolases
Hyaluronidases
Lactonases
Peptidases
Oxidorreductases
~ 3.16 million pre-selected sequences

Task 2.3: Explanation of the work carried

Progress undertaken and outputs achieved in M1-M18



Manufacturer's specifications

Sequences

Constraint Network Analysis

Phase Transition Temperature

(Global Structural Rigidity measure)

Correlation with higher Mean Annual Temperature

Thermal Adaptation

Protein Characterization

PFAM

INTERPRO

Structure Prediction

Alphafold 2.0

PELE Induced Fit Exhaustive

Catalytic Efficiency Estimation

Best Candidates

Machine Learning EP-Pred

Promiscuity Prediction

1355 candidate sequences of interest

Enrichment to bioprospecting

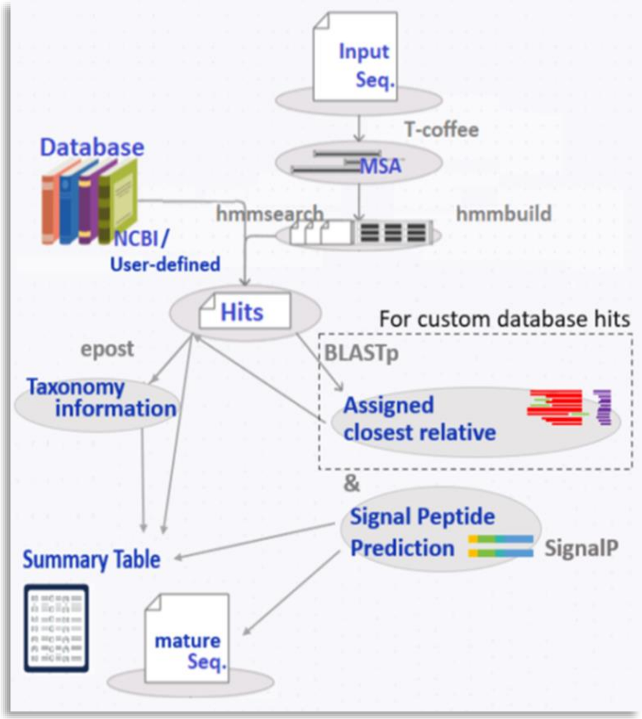
Hidden Markov Model

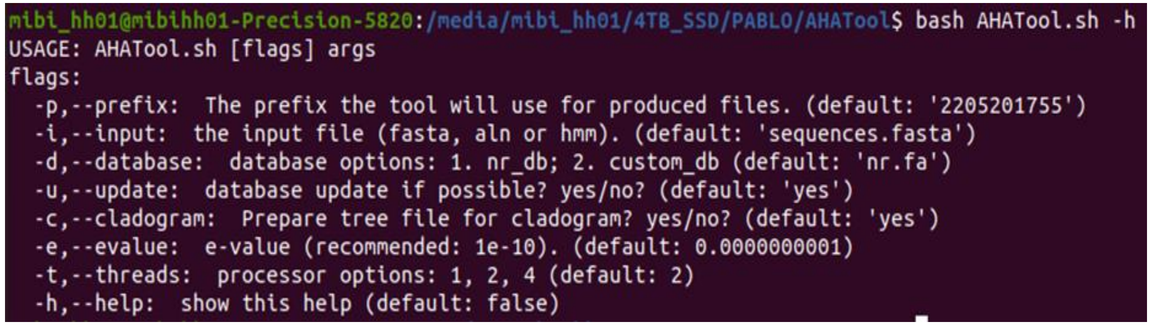
PSI-BLAST

Task 2.3: Explanation of the work carried

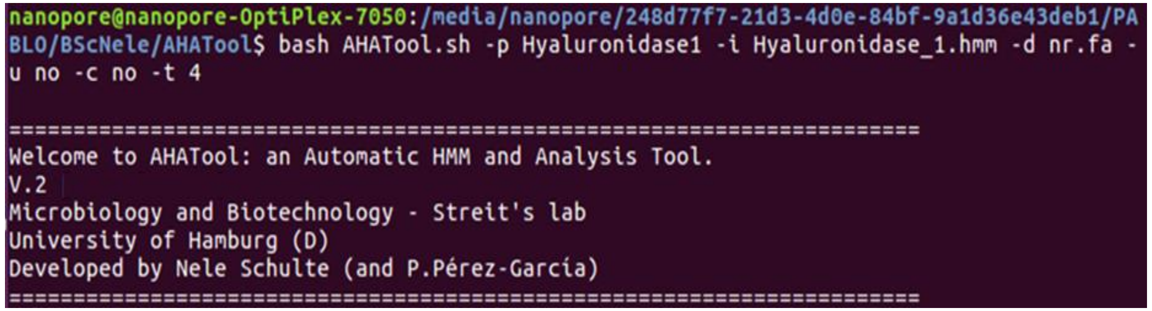
Progress undertaken and outputs achieved in M1-M18

- Development of AHA-tool, an HMM tool to find new enzymes



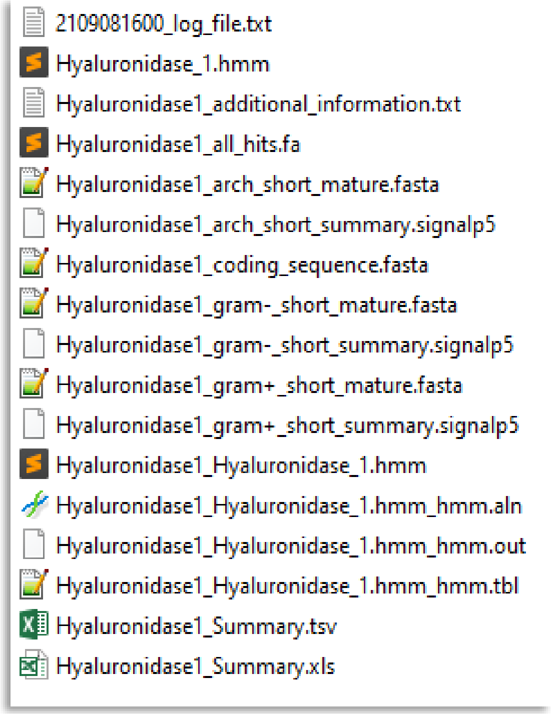


```
mibi_hh01@mibihh01-Precision-5820:/media/mibi_hh01/4TB_SSD/PABLO/AHATool$ bash AHATool.sh -h
USAGE: AHATool.sh [flags] args
flags:
  -p,--prefix:  The prefix the tool will use for produced files. (default: '2205201755')
  -i,--input:  the input file (fasta, aln or hmm). (default: 'sequences.fasta')
  -d,--database:  database options: 1. nr_db; 2. custom_db (default: 'nr.fa')
  -u,--update:  database update if possible? yes/no? (default: 'yes')
  -c,--cladogram:  Prepare tree file for cladogram? yes/no? (default: 'yes')
  -e,--evalue:  e-value (recommended: 1e-10). (default: 0.0000000001)
  -t,--threads:  processor options: 1, 2, 4 (default: 2)
  -h,--help:  show this help (default: false)
```



```
nanopore@nanopore-OptiPlex-7050:/media/nanopore/248d77f7-21d3-4d0e-84bf-9a1d36e43deb1/PA
BLO/BScNele/AHATool$ bash AHATool.sh -p Hyaluronidase1 -i Hyaluronidase_1.hmm -d nr.fa -
u no -c no -t 4

=========================================================================
Welcome to AHATool: an Automatic HMM and Analysis Tool.
V.2
Microbiology and Biotechnology - Streit's lab
University of Hamburg (D)
Developed by Nele Schulte (and P.Pérez-García)
=========================================================================
```



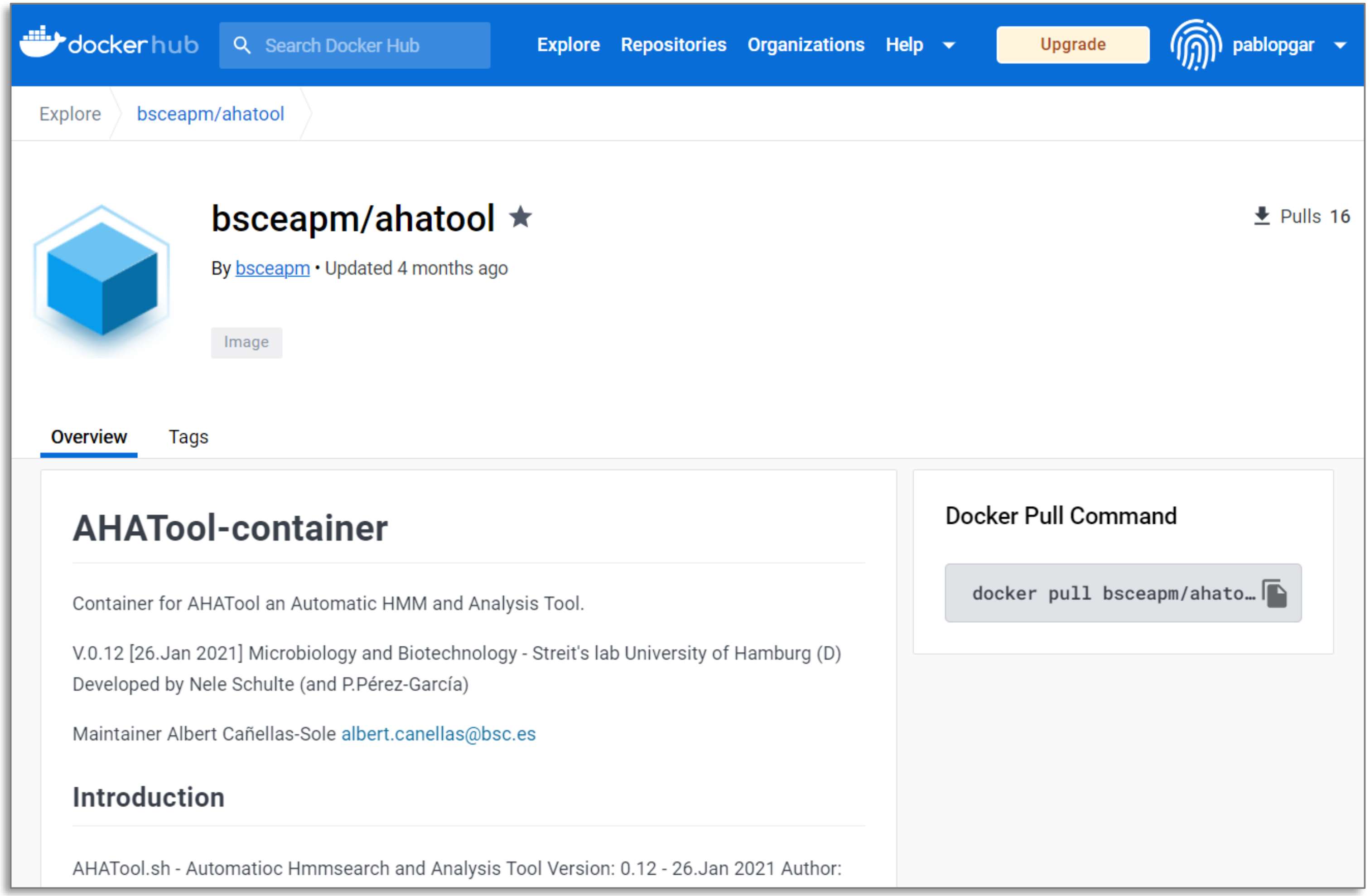
2109081600_log_file.txt
Hyaluronidase_1.hmm
Hyaluronidase1_additional_information.txt
Hyaluronidase1_all_hits.fa
Hyaluronidase1_arch_short_mature.fasta
Hyaluronidase1_arch_short_summary.signalp5
Hyaluronidase1_coding_sequence.fasta
Hyaluronidase1_gram-_short_mature.fasta
Hyaluronidase1_gram-_short_summary.signalp5
Hyaluronidase1_gram+_short_mature.fasta
Hyaluronidase1_gram+_short_summary.signalp5
Hyaluronidase1_Hyaluronidase_1.hmm
Hyaluronidase1_Hyaluronidase_1.hmm_hmm.aln
Hyaluronidase1_Hyaluronidase_1.hmm_hmm.out
Hyaluronidase1_Hyaluronidase_1.hmm_hmm.tbl
Hyaluronidase1_Summary.tsv
Hyaluronidase1_Summary.xls

https://hub.docker.com/r/bsceapm/ahatool

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- Development of AHA-tool, an HMM tool to find new enzymes



bsceapm/ahatool

By bsceapm • Updated 4 months ago

AHATool-container

Container for AHATool an Automatic HMM and Analysis Tool.

V.0.12 [26.Jan 2021] Microbiology and Biotechnology - Streit's lab University of Hamburg (D) Developed by Nele Schulte (and P.Pérez-García)

Maintainer Albert Cañellas-Sole albert.canellas@bsc.es

Introduction

AHATool.sh - Automatioc Hmmsearch and Analysis Tool Version: 0.12 - 26.Jan 2021 Author:

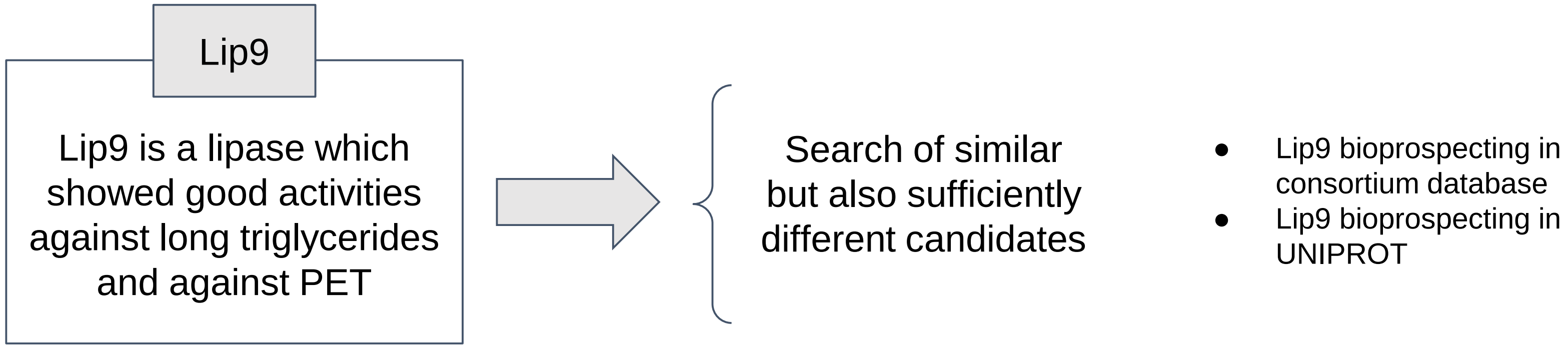
Docker Pull Command

```
docker pull bsceapm/ahato…
```

Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- Pre-selecting candidate sequences through extensive homology search
  - Second round of the iterative bioprospecting
  - 2 Lipases from the MarDB database (WP_054709477.1_MMP00000377; MTH54922.1_MMP13326190)
  - 2 Lipases from UHAM metagenomes (k127_15135326_1; k127_129897_3)
  - 13 Lipases from UNIPROT database (A0A1Q5DFC1, A0A5J6FBP2, A0A7X0G2Z7, A0A4R4W4R8, A0A7K3AES8, A0A7W0VIT6, A0A1K1R0A5, A0A2S6PTK3, A0A810NRQ6, A0A4P7DGE7, A0A1S2R3C1, A0A117RE37, and A0A4R6SGI7)



Lip9

Lip9 is a lipase which showed good activities against long triglycerides and against PET

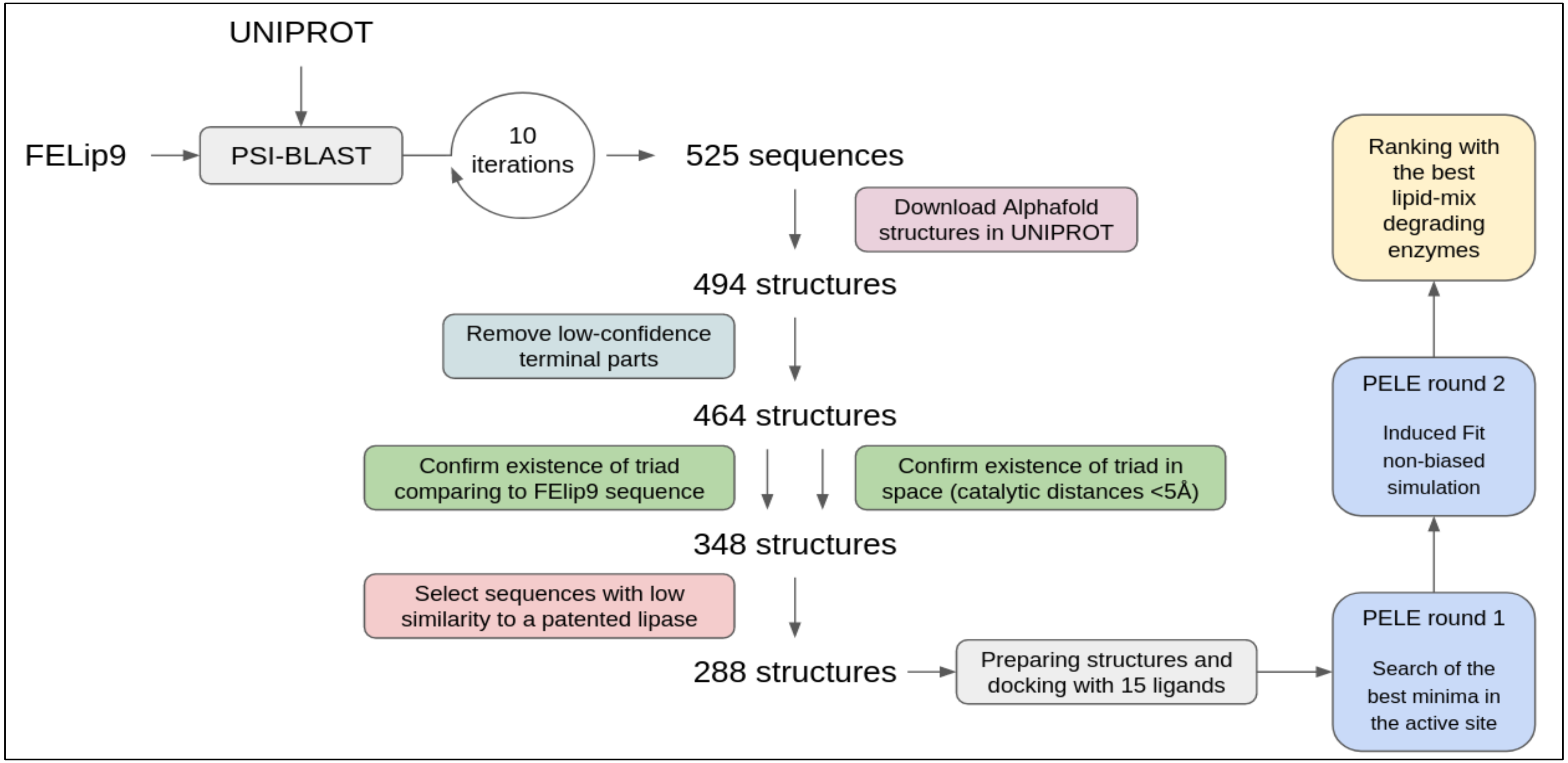
Search of similar but also sufficiently different candidates

- Lip9 bioprospecting in consortium database
- Lip9 bioprospecting in UNIPROT

Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- Lip9 bioprospecting in UNIPROT database



UNIPROT

FELip9

PSI-BLAST

10 iterations

525 sequences

Download Alphafold structures in UNIPROT

494 structures

Remove low-confidence terminal parts

464 structures

Confirm existence of triad comparing to FElip9 sequence

Confirm existence of triad in space (catalytic distances <5Å)

348 structures

Select sequences with low similarity to a patented lipase

288 structures

Preparing structures and docking with 15 ligands

PELE round 1

Search of the best minima in the active site

PELE round 2

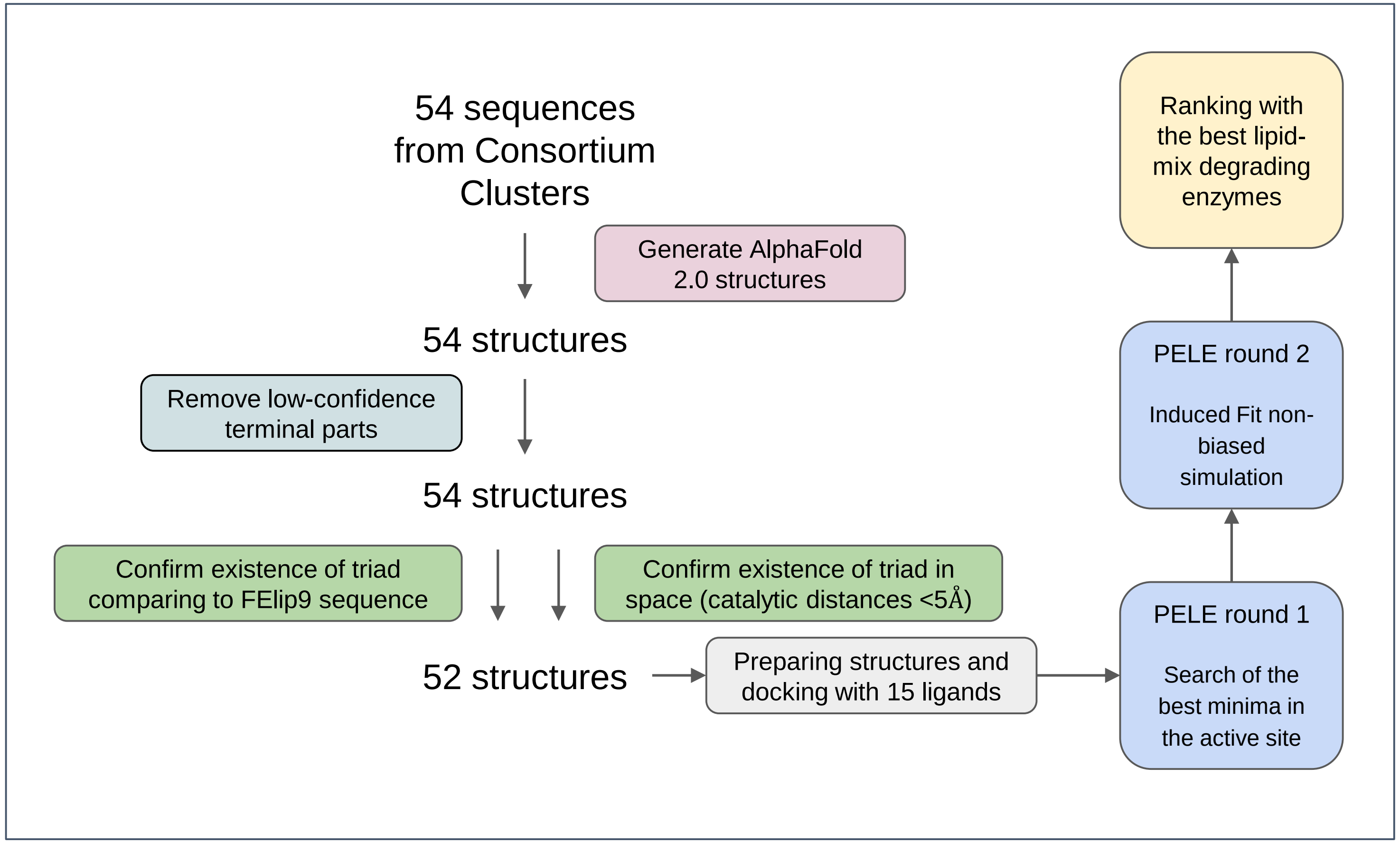
Induced Fit non-biased simulation

Ranking with the best lipid-mix degrading enzymes

Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- Lip9 bioprospecting in the consortium database



54 sequences from Consortium Clusters

Generate AlphaFold 2.0 structures

54 structures

Remove low-confidence terminal parts

54 structures

Confirm existence of triad comparing to FElip9 sequence

Confirm existence of triad in space (catalytic distances <5Å)

52 structures

Preparing structures and docking with 15 ligands

PELE round 1

Search of the best minima in the active site

PELE round 2

Induced Fit non-biased simulation

Ranking with the best lipid-mix degrading enzymes

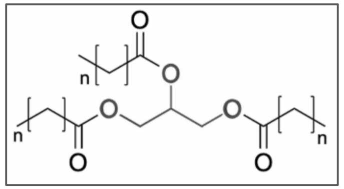
Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

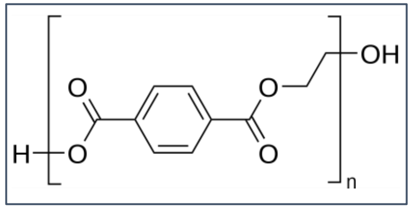
Ligands used in simulations

Ligands

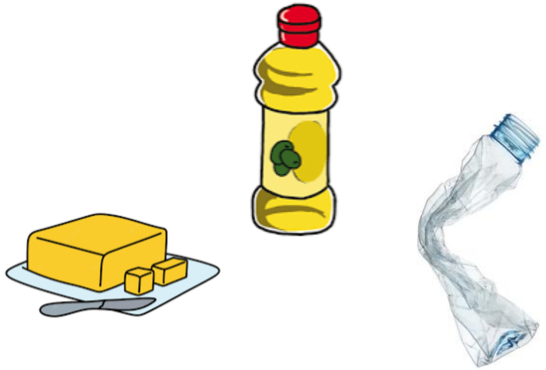
Triglycerides



Polyethylene terephthalate (PET) tetramer



Ligand name	Identification
U10	C10:0
U12	C12:0
GMY	C14:0
U16	C16:0
I16	C16:1
U17	C17:0
U18	C18:0
I18/TOL	C18:1
D18	C18:2
T18	C18:3
PT4	PET tetramer

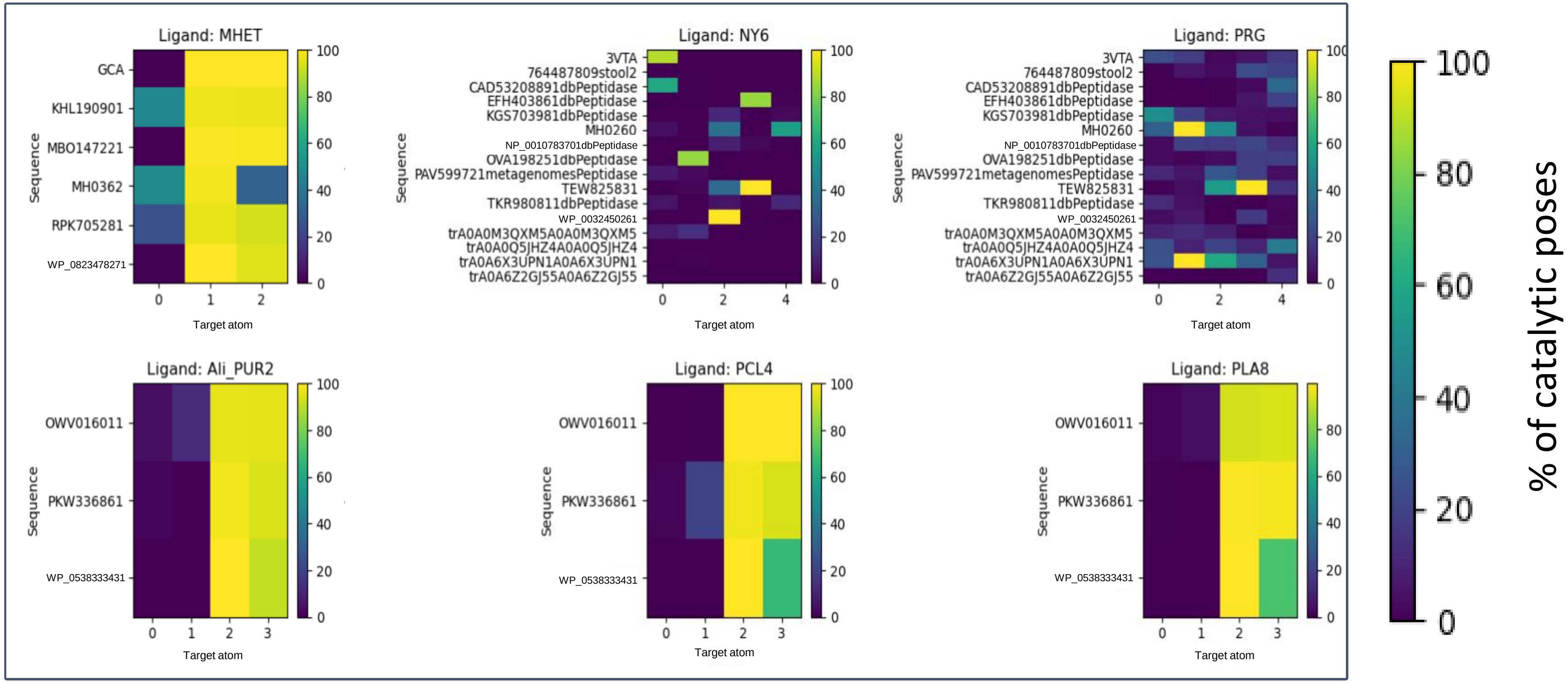


Triglycerides are commonly found in many natural oils and fats and are often used as substrates in enzymatic assays for lipases and esterases. As Lip9 can degrade PET, we also simulate PET tetramers.

Tasks 2.2-2.3: Explanation of the work carried

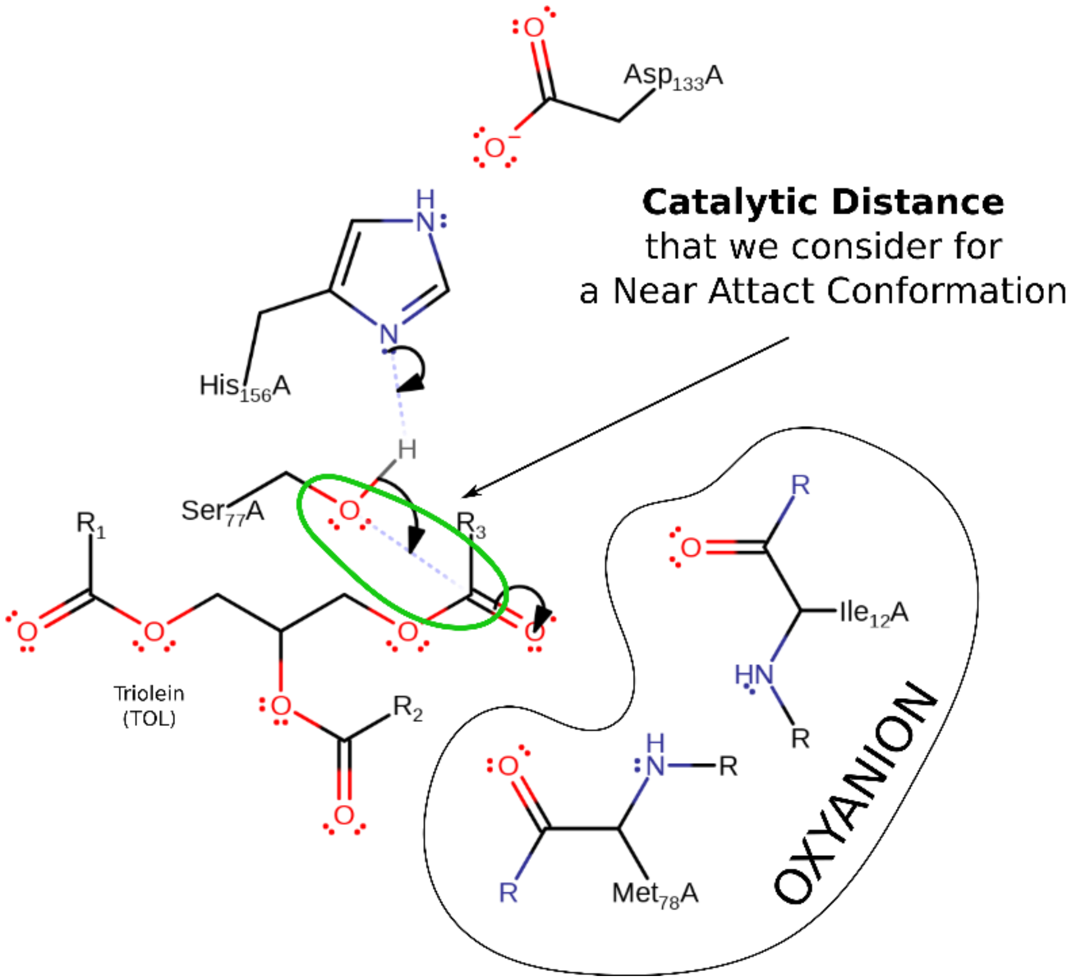
Progress undertaken and outputs achieved M18-M24

- Examples of simulations of sequences selected in Task 2.2



Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

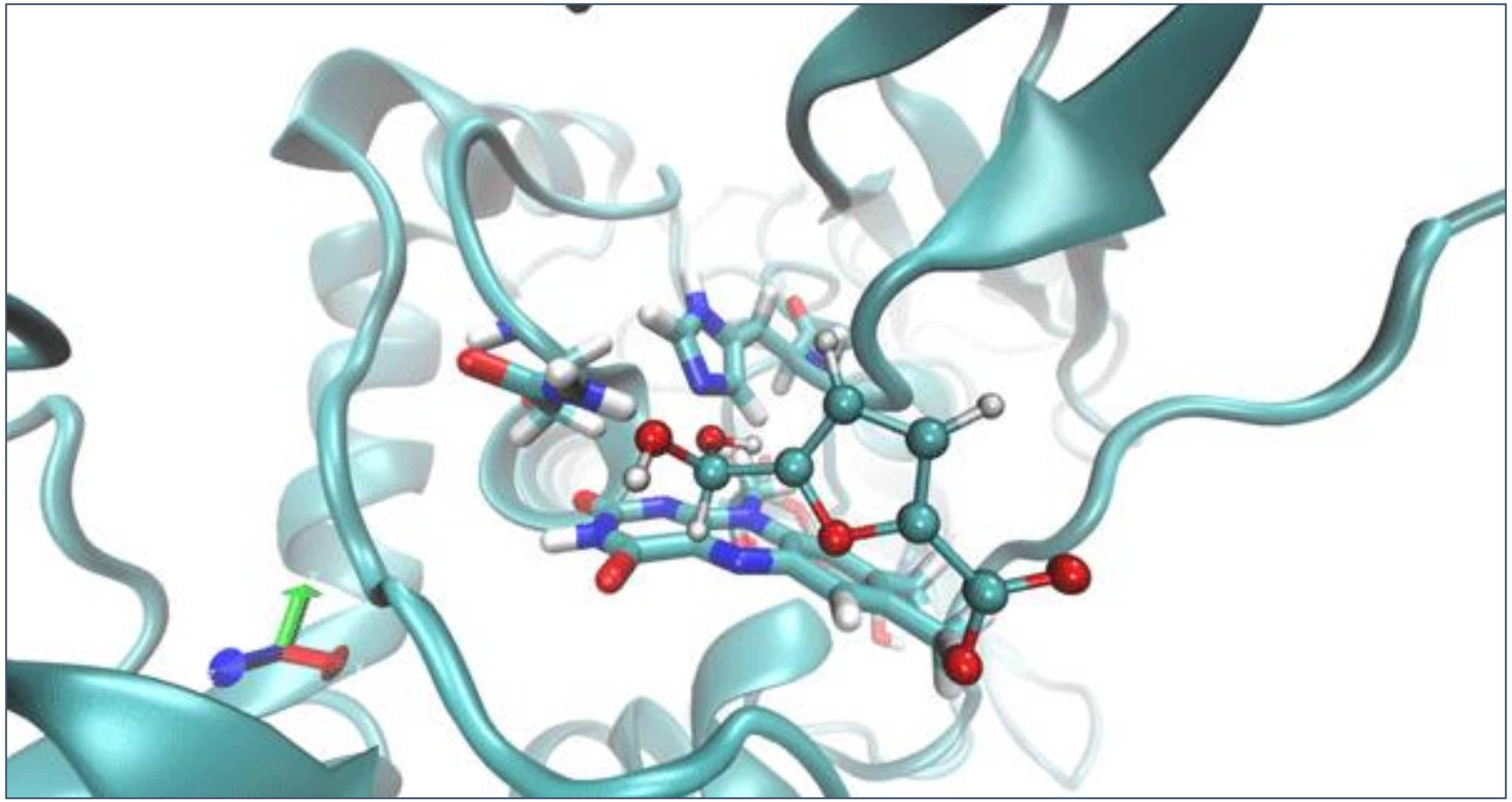


Catalytic Distance
that we consider for
a Near Attact Conformation

Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

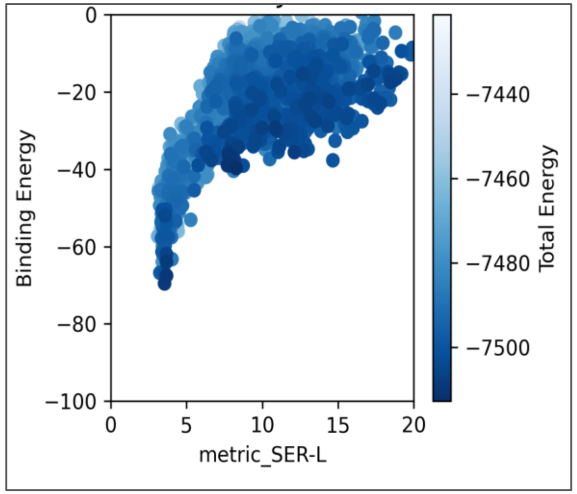
- PELE Induced Fit Simulations of Substrate/Active site interaction



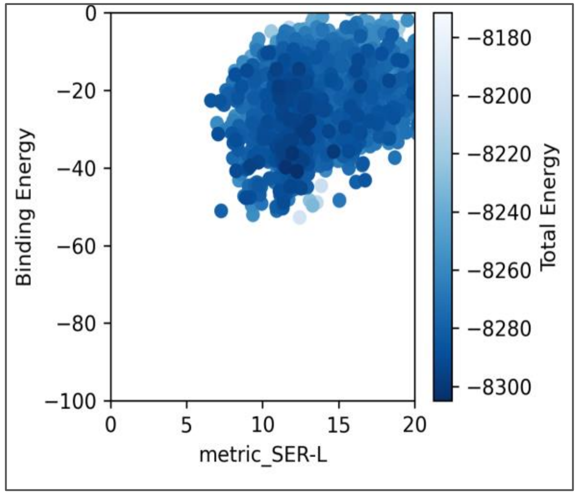
Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

A0A410WJ00-U18



A0A6I6D8S4-U18

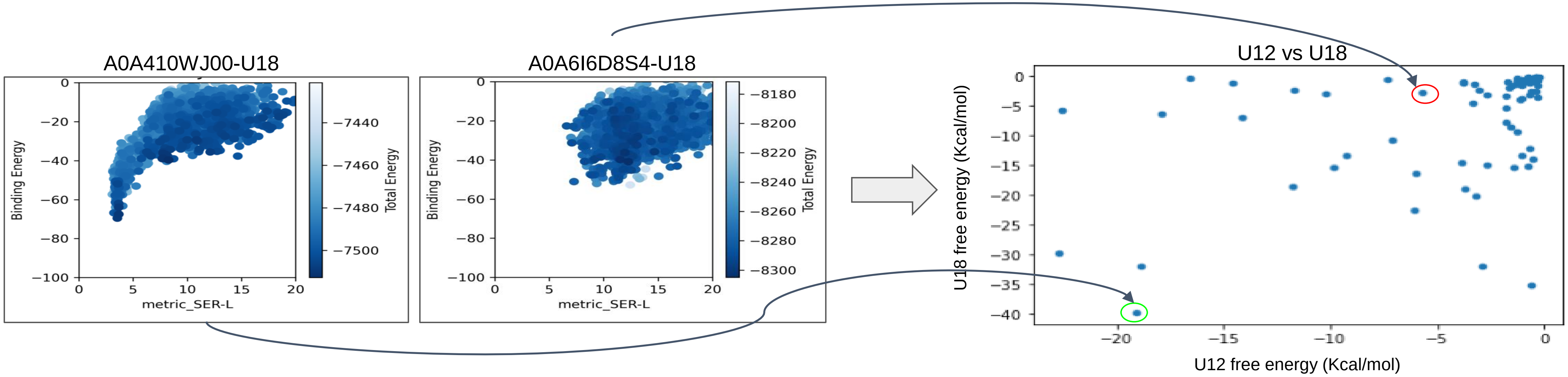


Binding free energy: a single value that can be used to compare among simulations

$$p_B^i = \frac{e^{\frac{-E_i}{K_B T}}}{\sum_i^N e^{\frac{-E_i}{K_B T}}} \qquad A_B = \sum_i^N E_b^i p_B^i \qquad E_B^C = \sum_i^{N_C} E_b^i {p_B}^i, i \in S_C$$

Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24



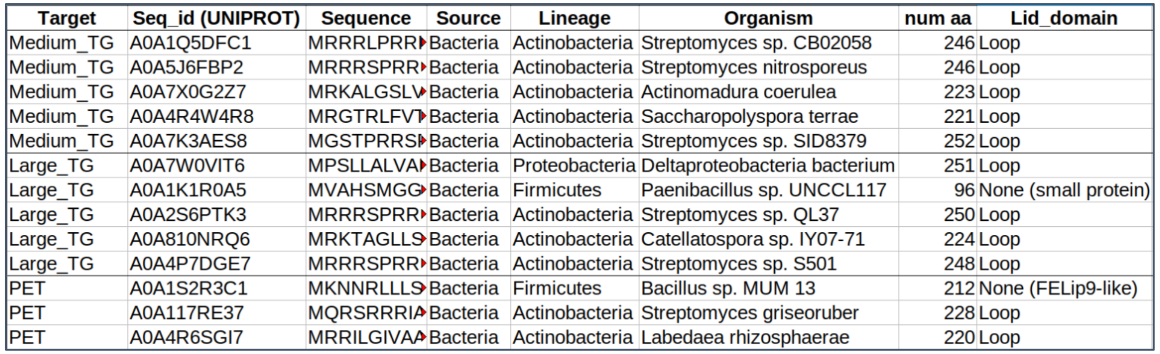
Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

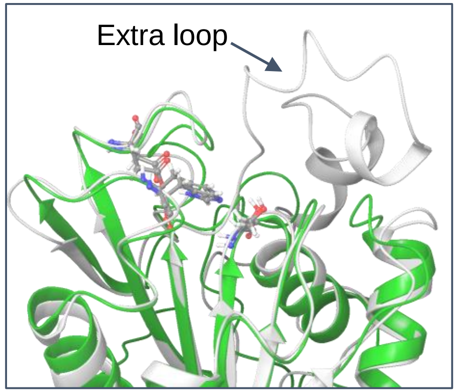
13 selected esterases

For medium and large triglycerides, selected candidates should bind all the ligands in the subgroup and bind the best one specific ligand.

For PET, the candidates which best bind that ligand were selected.



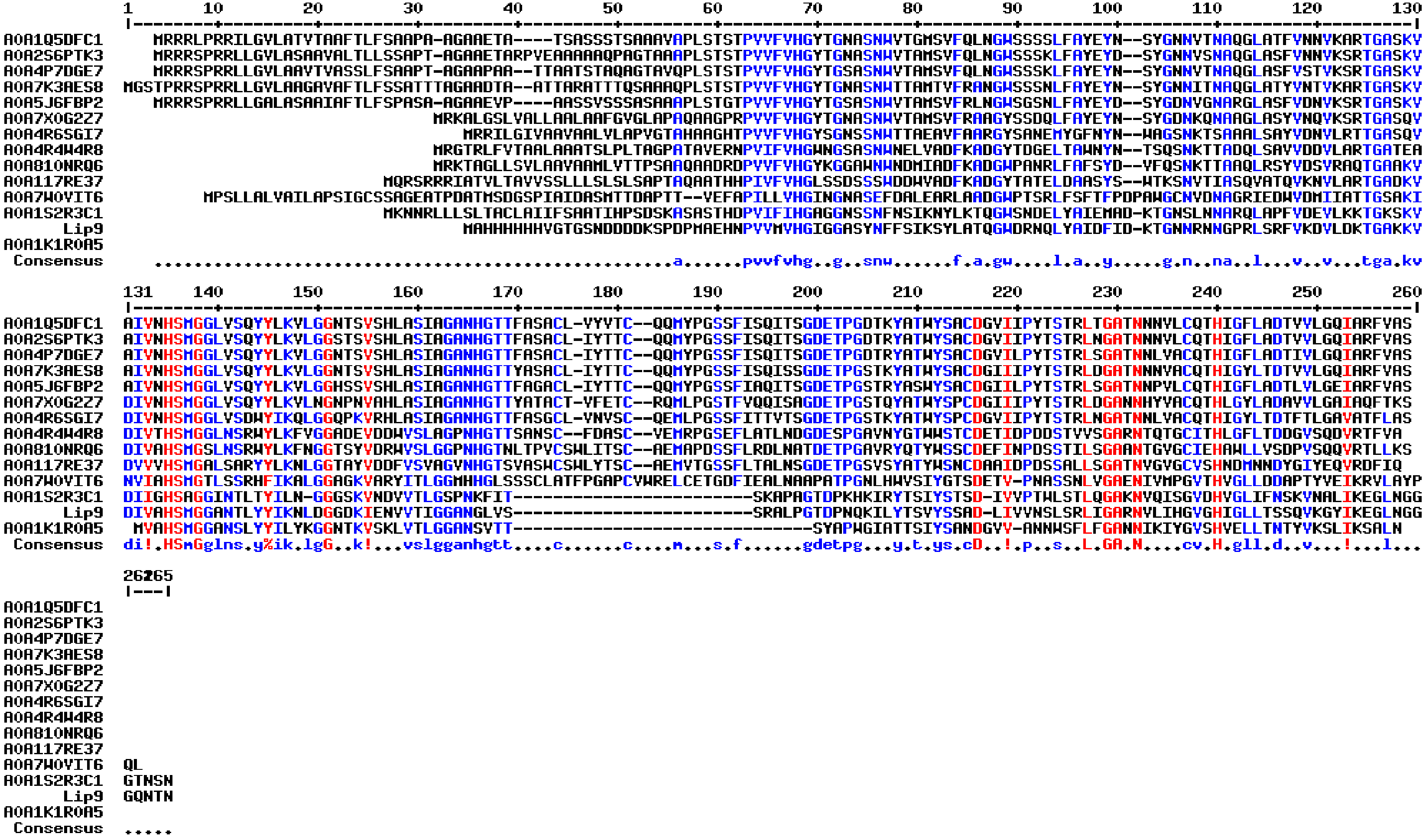
Target	Seq_id (UNIPROT)	Sequence	Source	Lineage	Organism	num aa	Lid_domain
Medium_TG	A0A1Q5DFC1	MRRRLPRR	Bacteria	Actinobacteria	Streptomyces sp. CB02058	246	Loop
Medium_TG	A0A5J6FBP2	MRRRSPRR	Bacteria	Actinobacteria	Streptomyces nitrosporeus	246	Loop
Medium_TG	A0A7X0G2Z7	MRKALGSLV	Bacteria	Actinobacteria	Actinomadura coerulea	223	Loop
Medium_TG	A0A4R4W4R8	MRGTRLFVT	Bacteria	Actinobacteria	Saccharopolyspora terrae	221	Loop
Medium_TG	A0A7K3AES8	MGSTPRRSI	Bacteria	Actinobacteria	Streptomyces sp. SID8379	252	Loop
Large_TG	A0A7W0VIT6	MPSLLALVA	Bacteria	Proteobacteria	Deltaproteobacteria bacterium	251	Loop
Large_TG	A0A1K1R0A5	MVAHSMGG	Bacteria	Firmicutes	Paenibacillus sp. UNCCL117	96	None (small protein)
Large_TG	A0A2S6PTK3	MRRRSPRR	Bacteria	Actinobacteria	Streptomyces sp. QL37	250	Loop
Large_TG	A0A810NRQ6	MRKTAGLLS	Bacteria	Actinobacteria	Catellatospora sp. IY07-71	224	Loop
Large_TG	A0A4P7DGE7	MRRRSPRR	Bacteria	Actinobacteria	Streptomyces sp. S501	248	Loop
PET	A0A1S2R3C1	MKNNRLLLS	Bacteria	Firmicutes	Bacillus sp. MUM 13	212	None (FELip9-like)
PET	A0A117RE37	MQRSRRRIA	Bacteria	Actinobacteria	Streptomyces griseoruber	228	Loop
PET	A0A4R6SGI7	MRRILGIVAA	Bacteria	Actinobacteria	Labedaea rhizosphaerae	220	Loop



Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

Too similar



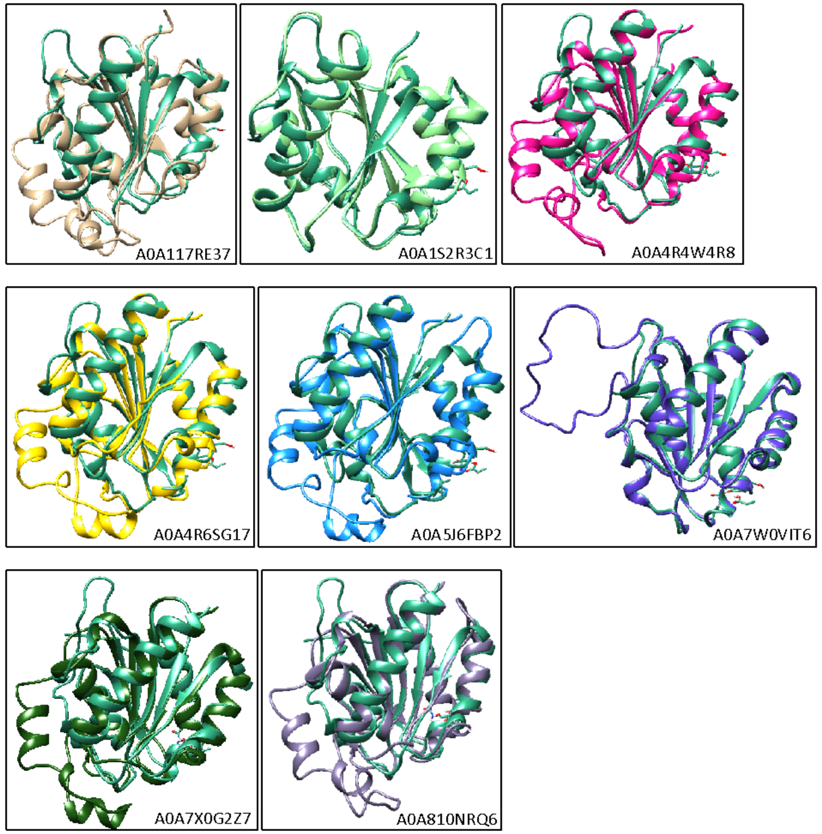
```
            1       10        20        30        40        50        60        70        80        90       100       110       120       130
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
A0A1Q5DFC1     MRRRLPRRILGVLATVTAAFTLFSAAPA-AGAAETA----TSASSSTSAAAVAPLSTSTPVVFVHGYTGNASNWVTGMSVFQLNGWSSSSLFAYEYN--SYGNNVTNAQGLATFVNNVKARTGASKV
A0A2S6PTK3     MRRRSPRRLLGVLASAAVALTLLSSAPT-AGAAETARPVEAAAAAQPAGTAAAPLSTSTPVVFVHGYTGNASNWVTAMSVFQLNGWSSSKLFAYEYD--SYGNNVSNAQGLASFVNNVKSRTGASKV
A0A4P7DGE7     MRRRSPRRLLGVLAAVTVASSLFSAAPT-AGAAPAA--TTAATSTAQAGTAVQPLSTSTPVVFVHGYTGSASNWVTAMSVFQLNGWSSSKLFAYEYN--SYGNNVTNAQGLASFVSTVKSRTGASKV
A0A7K3AES8  MGSTPRRSPRRLLGVLAAGAVAFTLFSSATTTAGAADTA--ATTARATTTQSAAAQPLSTSTPVVFVHGYTGNASNWTTAMTVFRANGWSSSNLFAYEYN--SYGNNITNAQGLATYVNTVKARTGASKV
A0A5J6FBP2     MRRRSPRRLLGALASAAIAFTLFSPASA-AGAAEVP----AASSVSSSASAAAPLSTGTPVVFVHGYTGSASNWVTAMSVFRLNGWSGSNLFAYEYD--SYGDNVGNARGLASFVDNVKSRTGASKV
A0A7X0G2Z7                                 MRKALGSLVALLAALAAFGVGLAPAQAAGPRPVVFVHGYTGNASNWVTAMSVFRAAGYSSDQLFAYEYN--SYGDNKQNAAGLASYVNQVKSRTGASQV
A0A4R6SGI7                                    MRRILGIVAAVAALVLAPVGTAHAAGHTPVVFVHGYSGNSSNWTTAEAVFAARGYSANEMYGFNYN--WAGSNKTSAAALSAYVDNVLRTTGASQV
A0A4R4W4R8                                 MRGTRLFVTAALAAATSLPLTAGPATAVERNPVIFVHGWNGSASNWNELVADFKADGYTDGELTAWNYN--TSQSNKTTADQLSAVVDDVLARTGATEA
A0A810NRQ6                                 MRKTAGLLSVLAAVAAMLVTTPSAAQAADRDPVVFVHGYKGGAWNWNDMIADFKADGWPANRLFAFSYD--VFQSNKTTAAQLRSYVDSVRAQTGAAKV
A0A117RE37                            MQRSRRRIATVLTAVVSSLLLSLSLSAPTAQAATHHPIVFVHGLSSDSSSWDDWVADFKADGYTATELDAASYS--WTKSNVTIASQVATQVKNVLARTGADKV
A0A7W0VIT6          MPSLLALVAILAPSIGCSSAGEATPDATMSDGSPIAIDASMTTDAPTT--VEFAPILLVHGINGNASEFDALEARLAADGWPTSRLFSFTFPDPAWGCNVDNAGRIEDWVDMIIATTGSAKI
A0A1S2R3C1                            MKNNRLLLSLTACLAIIFSAATIHPSDSKASASTHDPVIFIHGAGGNSSNFNSIKNYLKTQGWSNDELYAIEMAD-KTGNSLNNARQLAPFVDEVLKKTGKSKV
      Lip9                                    MAHHHHHHVGTGSNDDDDKSPDPMAEHNPVVMVHGIGGASYNFFSIKSYLATQGWDRNQLYAIDFID-KTGNNRNNGPRLSRFVKDVLDKTGAKKV
A0A1K1R0A5
 Consensus     ....................................................a......pvvfvhg..g..snw......f.a.gw....l.a..y.....g.n..na..l...v..v...tga.kv

            131     140       150       160       170       180       190       200       210       220       230       240       250       260
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
A0A1Q5DFC1  AIVNHSMGGLVSQYYLKVLGGNTSVSHLASIAGANHGTTFASACL-VYVTC--QQMYPGSSFISQITSGDETPGDTKYATWYSACDGVIIPYTSTRLTGATNNNVLCQTHIGFLADTVVLGQIARFVAS
A0A2S6PTK3  AIVNHSMGGLVSQYYLKVLGGSTSVSHLASIAGANHGTTFASACL-IYTTC--QQMYPGSSFISQITSGDETPGDTRYATWYSACDGVIIPYTSTRLNGATNNNVLCQTHIGFLADTVVLGQIARFVAS
A0A4P7DGE7  AIVNHSMGGLVSQYYLKVLGGNTSVSHLASIAGANHGTTFASACL-IYTTC--QQMYPGSSFISQITSGDETPGDTRYATWYSACDGVILPYTSTRLSGATNNLVACQTHIGFLADTIVLGQIARFVAS
A0A7K3AES8  AIVNHSMGGLVSQYYLKVLGGNTSVSHLASIAGANHGTTYASACL-IYTTC--QQMYPGSSFISQISSGDETPGSTKYATWYSACDGIIIPYTSTRLDGATNNNVACQTHIGYLTDTVVLGQIARFVAS
A0A5J6FBP2  AIVNHSMGGLVSQYYLKVLGGHSSVSHLASIAGANHGTTFAGACL-IYTTC--QQMYPGSSFIAQITSGDETPGSTRYASWYSACDGIILPYTSTRLSGATNNPVLCQTHIGFLADTLVLGEIARFVAS
A0A7X0G2Z7  DIVNHSMGGLVSQYYLKVLNGNPNVAHLASIAGANHGTTYATACT-VFETC--RQMLPGSTFVQQISAGDETPGSTQYATWYSPCDGIIIPYTSTRLDGANNHYVACQTHLGYLADAVVLGAIAQFTKS
A0A4R6SGI7  DIVNHSMGGLVSDWYIKQLGGQPKVRHLASIAGANHGTTFASGCL-VNVSC--QEMLPGSSFITTVTSGDETPGSTKYATWYSPCDGVIIPYTSTRLNGATNNLVACQTHIGYLTDTFTLGAVATFLAS
A0A4R4W4R8  DIVTHSMGGLNSRWYLKFVGGADEVDDWVSLAGPNHGTTSANSC--FDASC--VEMRPGSEFLATLNDGDESPGAVNYGTWWSTCDETIDPDDSTVVSGARNTQTGCITHLGFLTDDGVSQDVRTFVA
A0A810NRQ6  DIVAHSMGSLNSRWYLKFNGGTSYVDRWVSLGGPNHGTNLTPVCSWLITSC--AEMAPDSSFLRDLNATDETPGAVRYQTYWSSCDEFINPDSSTILSGAANTGVGCIEHAWLLVSDPVSQQVRTLLKS
A0A117RE37  DVVVHSMGALSARYYLKNLGGTAYVDDFVSVAGVNHGTSVASWCSWLYTSC--AEMVTGSSFLTALNSGDETPGSVSYATYWSNCDAAIDPDSSALLSGATNVGVGCVSHNDMNNDYGIYEQVRDFIQ
A0A7W0VIT6  NVIAHSMGTLSSRHFIKALGGAGKVARYITLGGMHHGLSSSCLATFPGAPCVWRELCETGDFIEALNAAPATPGNLHWVSIYGTSDETV-PNASSNLVGAENIVMPGVTHVGLLDDAPTYVEIKRVLAYP
A0A1S2R3C1  DIIGHSAGGINTLTYILN-GGGSKVNDVVTLGSPNKFIT------------------------SKAPAGTDPKHKIRYTSIYSTSD-IVVPTWLSTLQGAKNVQISGVDHVGLIFNSKVNALIKEGLNGG
      Lip9  DIVAHSMGGANTLYYIKNLDGGDKIENVVTIGGANGLVS------------------------SRALPGTDPNQKILYTSVYSSAD-LIVVNSLSRLIGARNVLIHGVGHIGLLTSSQVKGYIKEGLNGG
A0A1K1R0A5   MVAHSMGGANSLYYILYKGGNTKVSKLVTLGGANSVTT------------------------------SYAPWGIATTSIYSANDGVV-ANNWSFLFGANNIKIYGVSHVELLTNTYVKSLIKSALN
 Consensus  di!.HSmGglns.y%ik.lgG..k!...vslgganhgtt....c......c....m...s.f......gdetpg...y.t.ys.cD..!.p..s..L.GA.N....cv.H.gll.d..v...!...l...

            26265
            |---|
A0A1Q5DFC1
A0A2S6PTK3
A0A4P7DGE7
A0A7K3AES8
A0A5J6FBP2
A0A7X0G2Z7
A0A4R6SGI7
A0A4R4W4R8
A0A810NRQ6
A0A117RE37
A0A7W0VIT6  QL
A0A1S2R3C1  GTNSN
      Lip9  GQNTN
A0A1K1R0A5
 Consensus  .....
```

Tasks 2.2-2.3: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24



Target	Seq_id (UNIPROT)
Medium_TG	A0A1Q5DFC1
Medium_TG	A0A5J6FBP2
Medium_TG	A0A7X0G2Z7
Medium_TG	A0A4R4W4R8
Medium_TG	A0A7K3AES8
Large_TG	A0A7W0VIT6
Large_TG	A0A1K1R0A5
Large_TG	A0A2S6PTK3
Large_TG	A0A810NRQ6
Large_TG	A0A4P7DGE7
PET	A0A1S2R3C1
PET	A0A117RE37
PET	A0A4R6SGI7

🟥	discarded
🟨	low priority
🟩	good
🟩	best

ranking

(left) homologs model structures superposed to Lip9 model structure (dark green). (right) table of sequences with target substrate colored by ranking.

Task 2.4: Explanation of the work carried

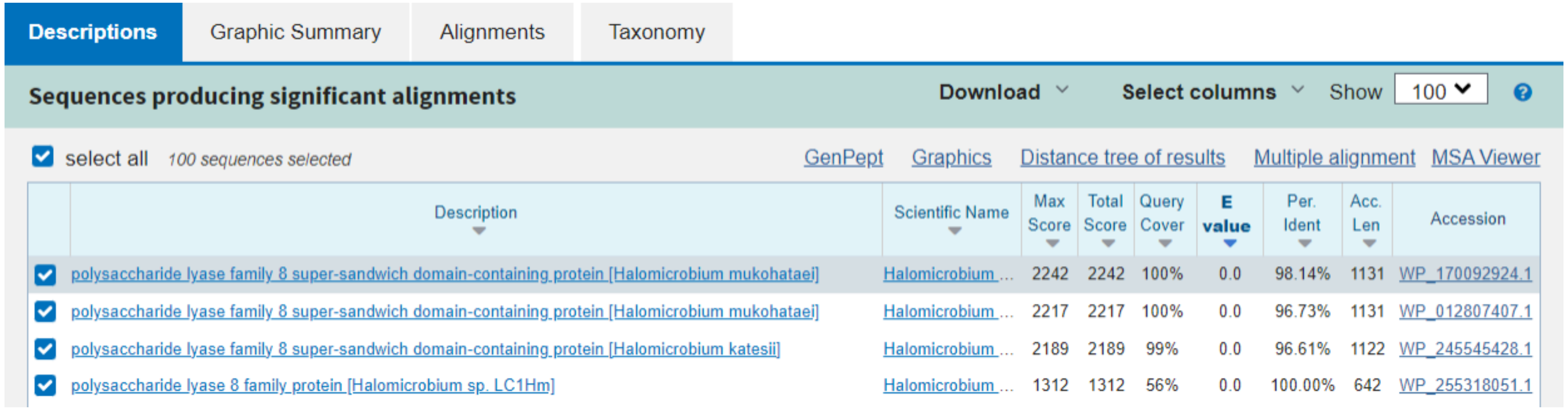
Progress undertaken and outputs achieved

- Pre-selecting candidate sequences through extensive homology search
  - Second round of the iterative bioprospecting
  - 1 Hyaluronidase, LC1Hm_4133 from partner CNR

```
MSDGWSRRSVLKSSLGLSLAGVSLSGTTETVTGASEYETLRQRWAQLLTGGDFDATQFEYQDPLAELDETAQDHWETMDTSADRDRLWSDLPIPASSSASA
SESNITDSYGRLQEMAMAYATNGSSLEDDSALVADIVDGLDFLYDRVYNEDQSQFGNWWHWEIGSPMRLVSVCALVGDELSSTQETNYTNAVGAHTGTP
YEYTEYDVTSGGANRVDMCIITALRGAISGTDSTIALARDCIEESDIFQYNTSGGGNGLYRDGSYVYHKEIPYIGSYGAILLEGLGELFTVLDGTTWEITDVDHDVI
YDAVGDAVAPFMYRGLMMDAVSGRSISRADQTDHVRGHGITATVLRLANTAPEPYASEFRSLAKGWIENDTWDSFLSDADVPDIANATAVLDDSTISAADE
PVRHDVFHNMDRVVHNRSEWAYTISMCSERIARYEAINEENLRGWYTGAGMTQLYNDDLGHYTDGYWPTVDPYRLPGTTVDTRERSTLDGTHHPRPSTQ
WVGGASVDEFGIAGMEFDAEGASLTGKKSWLHLDDTVVALGADITSSDGRPIETTVENRNLHTDGSETLTVDDTEKSTTPDWSETLTDVSWAHLDGVGGYL
FPNQPTLEAKREERTGSWQEINAGGPSESLTREYQTLWLDHGVDPSAETYAYALLPGHTASETRQRSQEPGFEIVANDATVQAVTVPRLGLTAANFWSSGSI
TVPGSERTLSVSGPAAVVVRHRNDELVIGVADPSRTQETVTVEYEHYTDGIVSTDSAVGVTQFRPGVTMEVAVGGTRGATHSATFDAPVTELSPRADTFVRD
GSYSGDNYGSWSSLVVKGGPTGYSRESYLAFDLASVAGEVQEAVLDVYGAVTDDNGGASVDCTVAAVDDDSWTEDGLTWDTKPDLGSSLGSLTVTRERR
WWREDVTEFVQTAASGDGIASVALRQPNDERYASFDSREADENPPSLRVTTSRPDTTALTPTADTFVRDGSYSGDNYGSWSSLVVKNAATDYSRQGYLTFD
LSALSGSIDEAVLYLYGAVTDDSGGDAVDCAINAVGDDSWTESGLTWDTKPDLGSALGSVTVTRTPQWWTVDVTEFVQSEAGGDGVVSLAVQQPQSGLYT
DFNSRDADEKVPTLRVQTS
```

Task 2.4: Explanation of the work carried

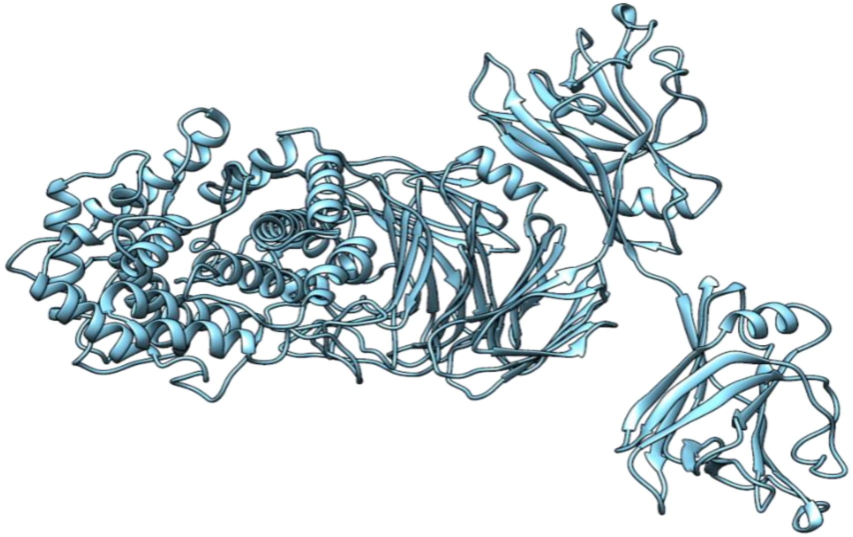
Progress undertaken and outputs achieved M18-M24



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	polysaccharide lyase family 8 super-sandwich domain-containing protein [Halomicrobium mukohataei]	Halomicrobium ...	2242	2242	100%	0.0	98.14%	1131	WP_170092924.1
☑	polysaccharide lyase family 8 super-sandwich domain-containing protein [Halomicrobium mukohataei]	Halomicrobium ...	2217	2217	100%	0.0	96.73%	1131	WP_012807407.1
☑	polysaccharide lyase family 8 super-sandwich domain-containing protein [Halomicrobium katesii]	Halomicrobium ...	2189	2189	99%	0.0	96.61%	1122	WP_245545428.1
☑	polysaccharide lyase 8 family protein [Halomicrobium sp. LC1Hm]	Halomicrobium ...	1312	1312	56%	0.0	100.00%	642	WP_255318051.1

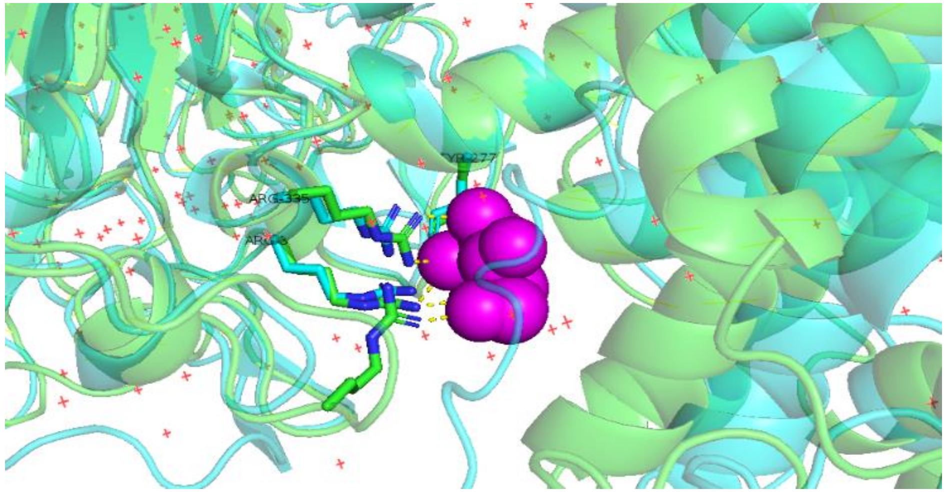


The model obtained this time is created with alphafold2 directly from blast page.

When performing a swissmodel calculation, the template obtained is 2e24.1 Xanthan lyase, another similar template 2e22.1 has a mannose residue inside the crystal.

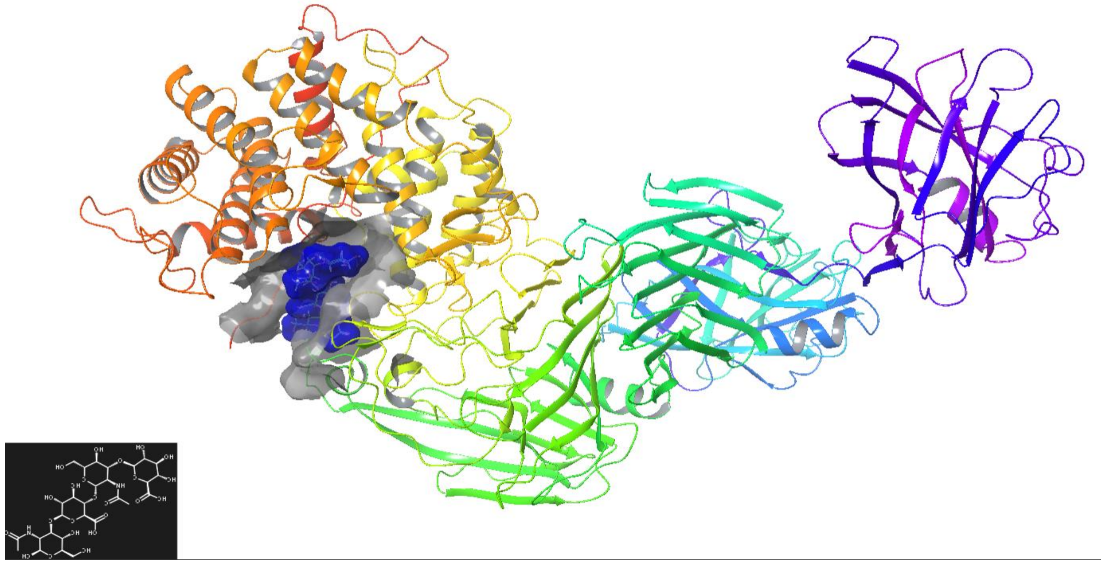
Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24



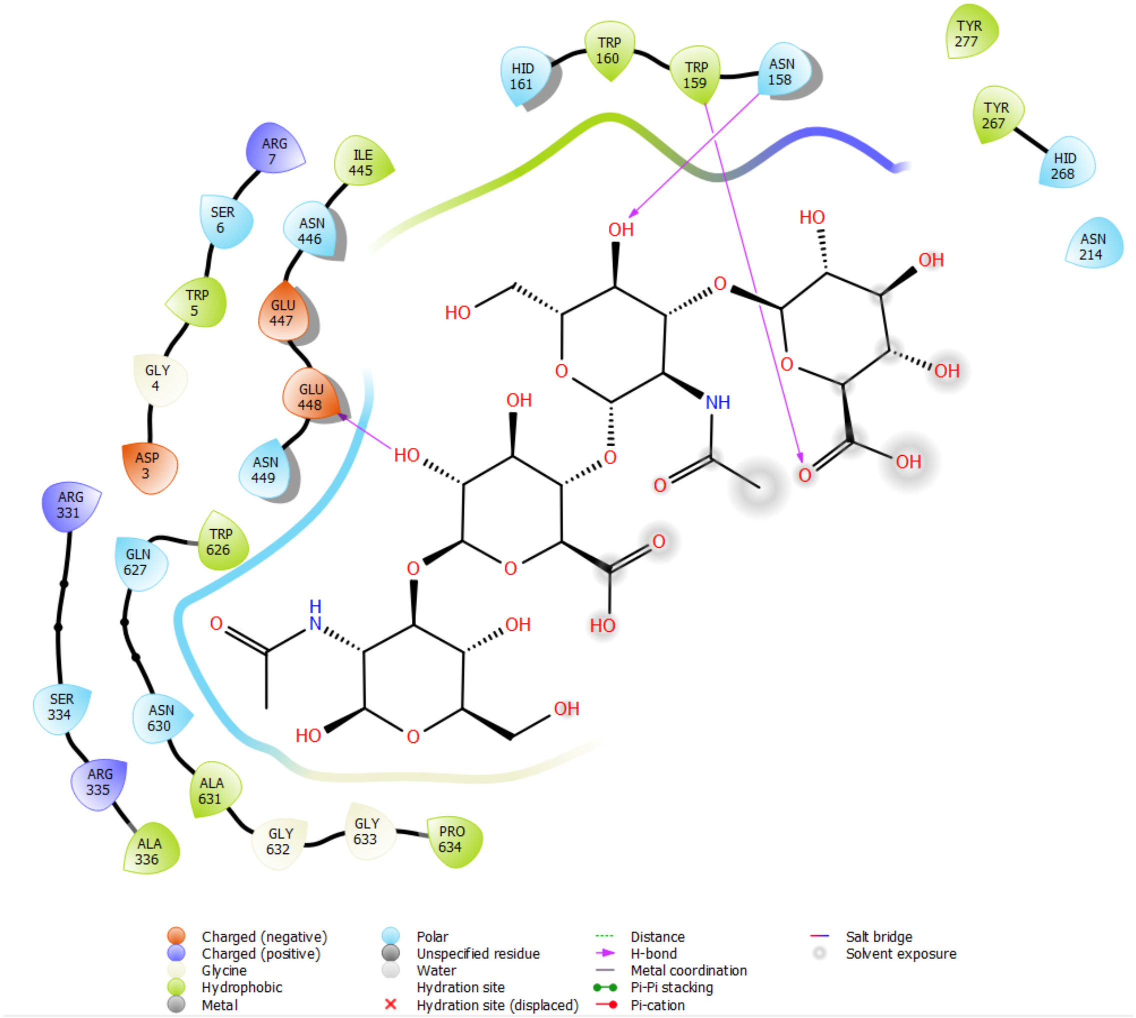
We aligned the crystal with the mannose and the model of LC1Hm_4133 to see which amino acids are in contact with the substrate. 2e22.1 is represented in green, LC1Hm_4133 is represented in blue and mannose is represented in magenta spheres. We also show the polar contacts between mannose and the residues from both enzymes that are shown is sticks.

We performed a docking (swissdock) with hyaluronic acid obtained from chemspider directed against residue Arg331 atom Ccz with a 10 angstroms window. As the protein is too much big, we use only the first domain (450 aa).



Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24



We conclude that LC1Hm_4133 is a good sequence but we suggest to cut the sequence in Val799.

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- Machine Learning methods for enzyme bioprospecting
  - Ever growing databases that are waiting to be explored and too much for experimental testing
    - We are not only interested in function which can be inferred from homology comparisons but also properties like thermostability, substrate specificity, etc.
    - Several examples like Soluprot or DeepLoc as tools that can increase the success of bioprospecting

Task 2.4: Explanation of the work carried

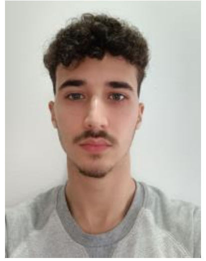
Progress undertaken and outputs achieved M18-M24

- **Machine Learning methods for enzyme bioprospecting**

EP-Pred	Promiscuity predictor (Published)
Bio-ML	Enzymatic searcher
SCOT	Stability predictor
Noname yet	Expression predictor (In early stages)

Albert Cañellas

Ruite Xiang

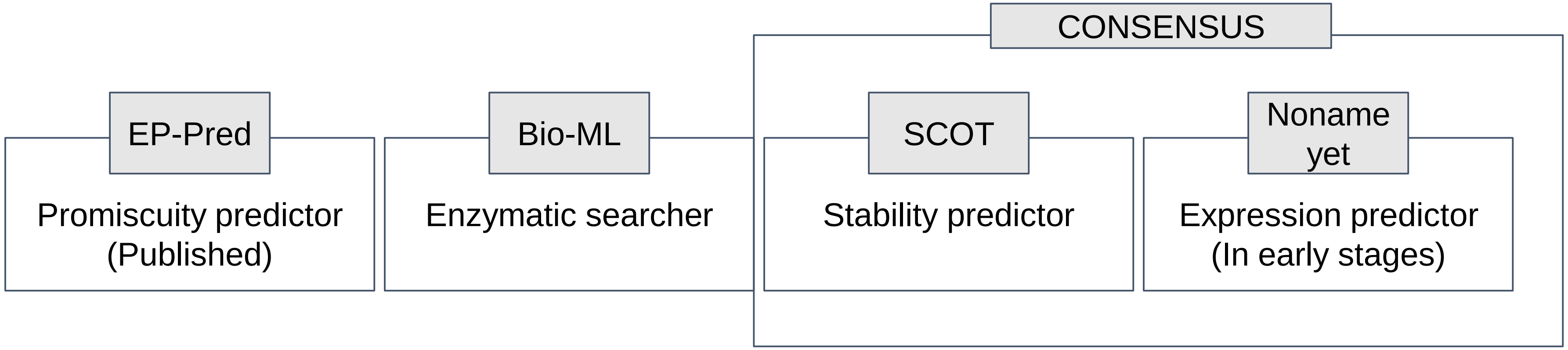


José María Romero

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- **Machine Learning methods for enzyme bioprospecting**



CONSENSUS

EP-Pred

Promiscuity predictor (Published)

Bio-ML

Enzymatic searcher

SCOT

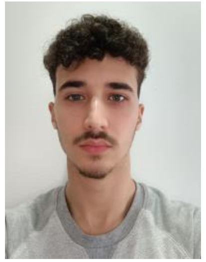
Stability predictor

Noname yet

Expression predictor (In early stages)

Albert Cañellas

Ruite Xiang

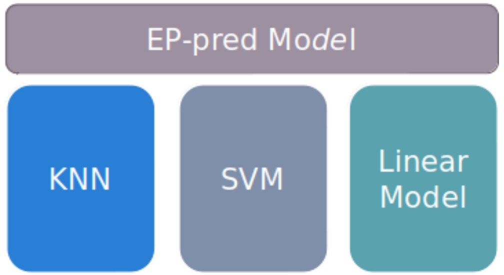


José María Romero

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

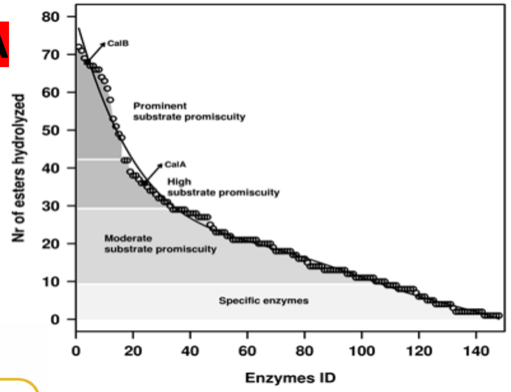
EP-Pred: A Machine Learning Tool for Bioprospecting Promiscuous Ester Hydrolases

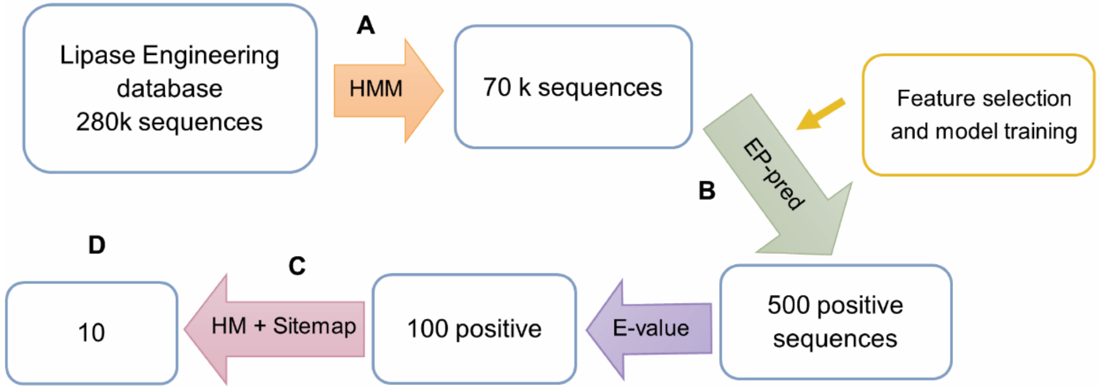


⇨ EP-pred is an ensemble classifier formed by 3 models

The models predicts promiscuity from the sequences and were tested on the Lipase Engineering database

DATA!



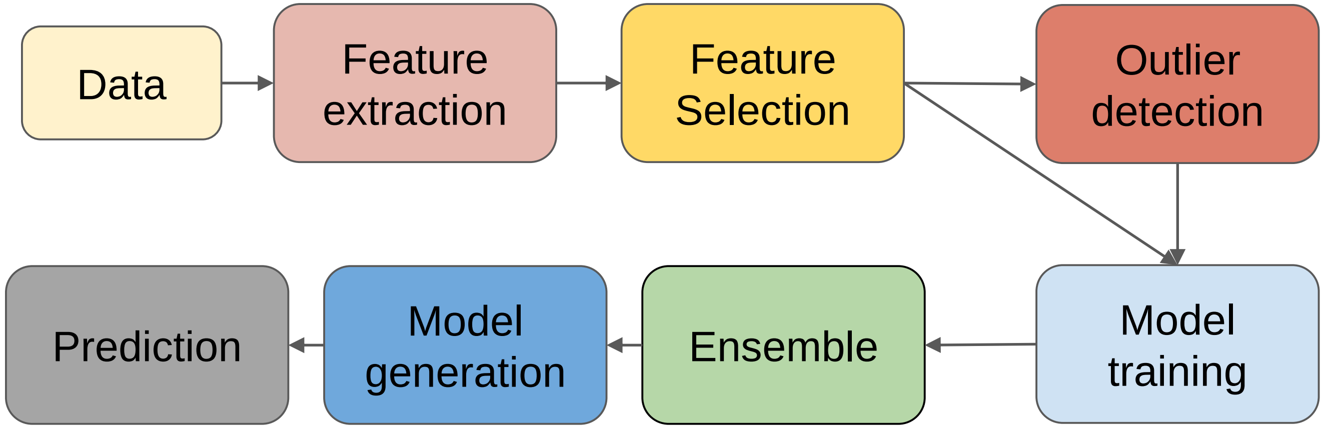


Biomolecules. 2022 Oct 21;12(10):1529

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

Bio-ML: A Machine Learning Tool for Bioprospecting Enzymes with specific activities

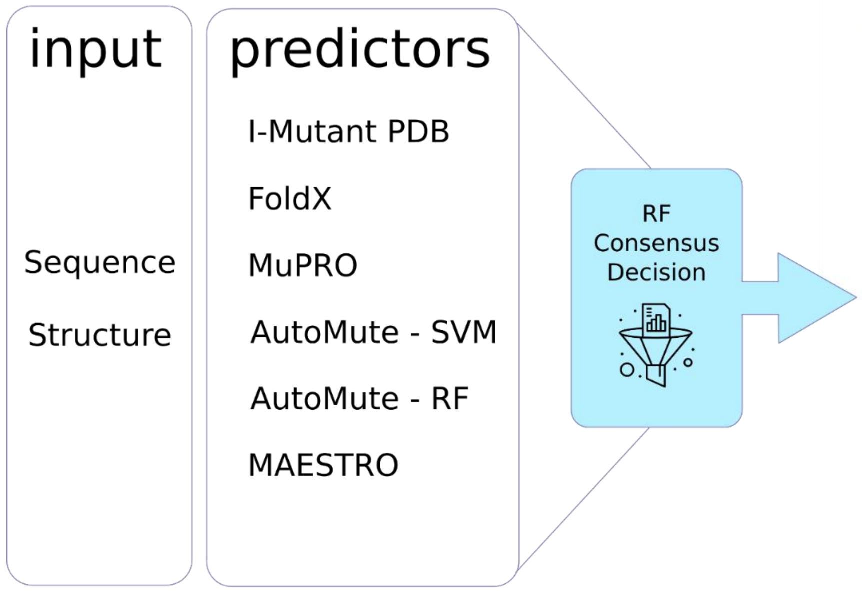


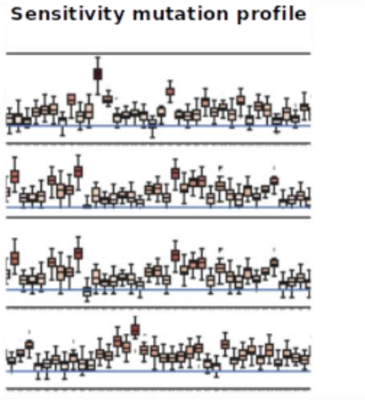
Bio-ML takes the same idea as EP-Pred but instead of predicting promiscuity, its target is activity

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

SCOT: Stability COnsensus Metapredictor





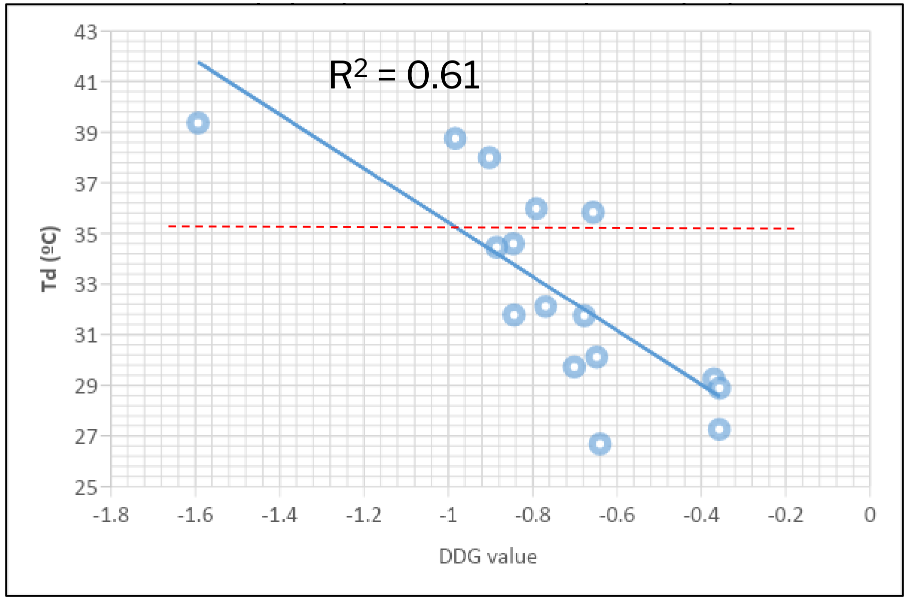
RANKING	MUTATION	RFC %	RFR ΔΔG (Kcal/mol)
1	44L	98	-0.5004
2	284M	97	-0.6562
3	302A	97	-0.3168
4	45I	96	-0.2228
5	283I	95	-0.8898
6	284L	95	-0.8608
7	265M	95	-0.7479
8	26M	95	-0.439
9	265I	94	-1.1778
10	152L	94	-0.8301
11	62L	94	-0.7276
12	284V	94	-0.406
13	22L	94	-0.3843
...	...	...	...
350	55I	71	-0.0056

- SCOT is a Random Forest based Machine Learning metapredictor that combines the estimations of already published protein stability predictors and a molecular filter to produce a more reliable result.
- Predictors: MAESTRO, AUTOMUTE-SVM and AUTOMUTE-TR, FOLDX, MUPRO and I-MUTANT.

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

SCOT: Stability COnsensus Metapredictor

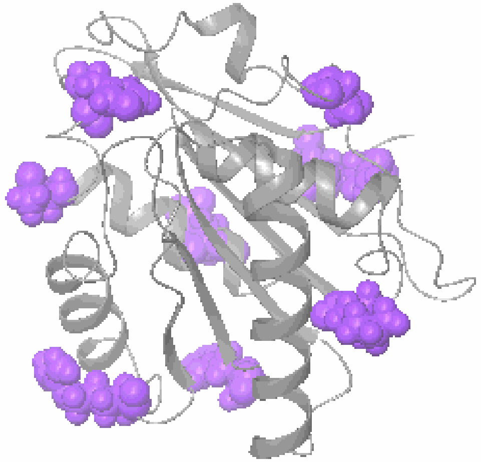


- SCOT is a Random Forest based Machine Learning metapredictor that combines the estimations of already published protein stability predictors and a molecular filter to produce a more reliable result.
- Predictors: MAESTRO, AUTOMUTE-SVM and AUTOMUTE-TR, FOLDX, MUPRO and I-MUTANT.

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

SCOT: Stability COnsensus Metapredictor



MUT	Conf. %	DDG regressor estimation
138W	0.89	-1.57259999256581
178L	0.89	-1.55980000313371
138Y	0.85	-1.53439999027178
138M	0.74	-1.69759999528527
155L	0.82	-1.50200000040233
99M	0.88	-1.28880002802238
53P	0.88	-1.23099999995902
159L	0.92	-0.98160000288859
138V	0.6	-1.5024999842979

- SCOT is a Random Forest based Machine Learning metapredictor that combines the estimations of already published protein stability predictors and a molecular filter to produce a more reliable result.

- Predictors: MAESTRO, AUTOMUTE-SVM and AUTOMUTE-TR, FOLDX, MUPRO and I-MUTANT.

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

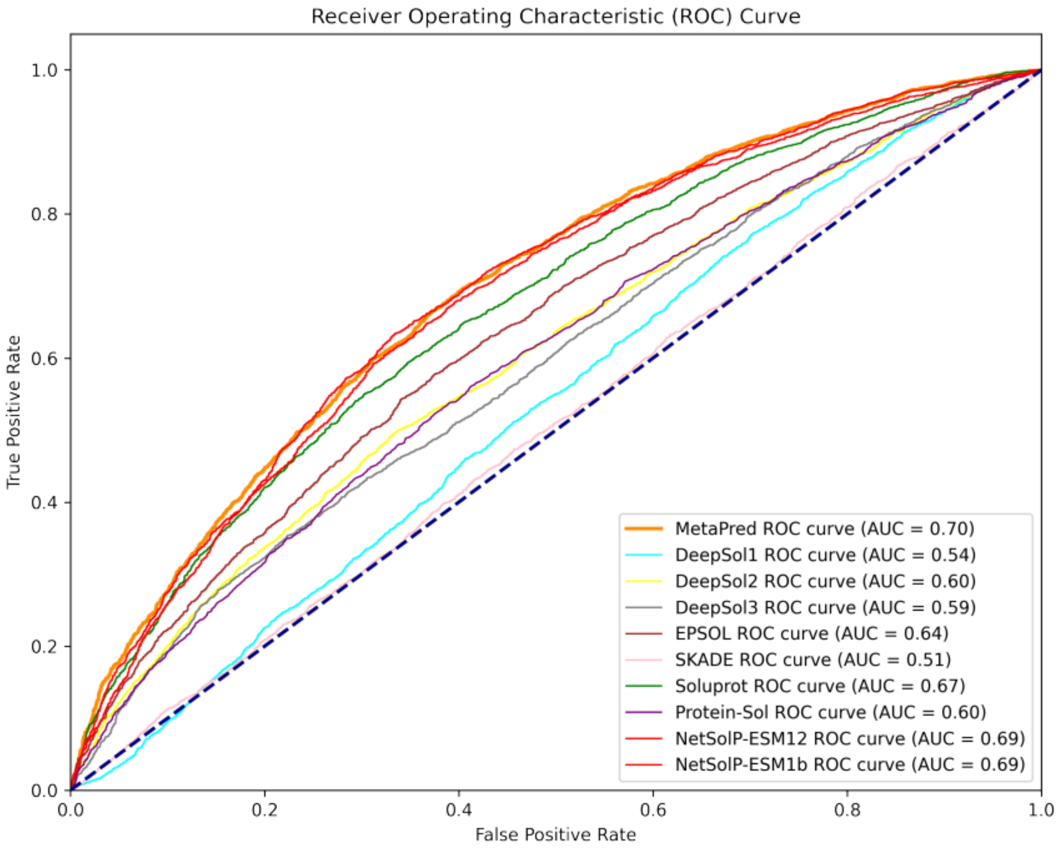
Expression Metapredictor (early stages)

We have review the state-of-the-art solubility/expression predictors:

DeepSol, EPSOL, SKADE, Soluprot, Protein-SOL and NetSolP

- XGBoost Decision Tree Consensus Model (with sequence embedding as features)
- NetSolP-esm model is based on deep learning protein language models called transformers
- Protein language model seems to be the way to go to create a more accurate sequence embedding that extract protein properties

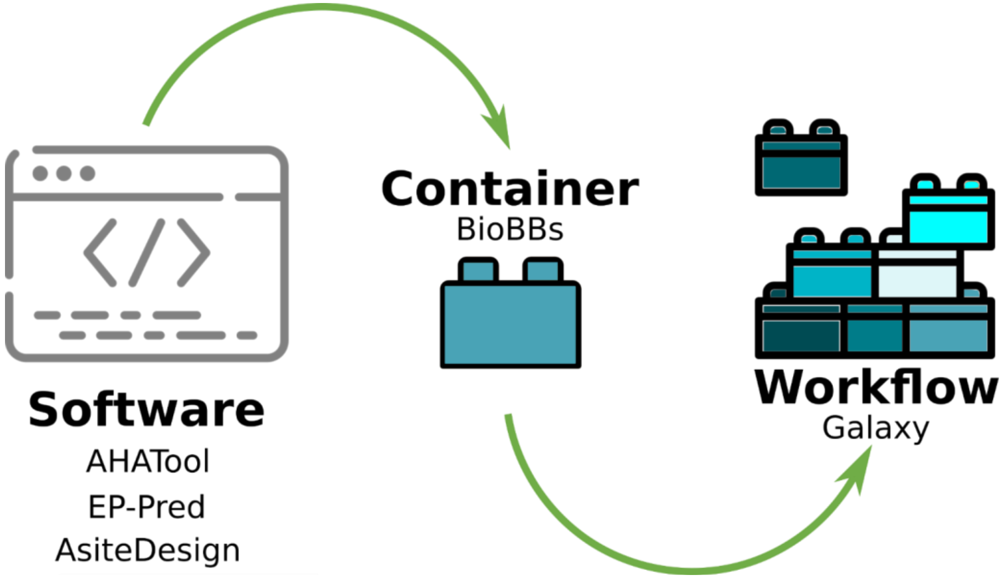
We need More Expression/Solubility Data



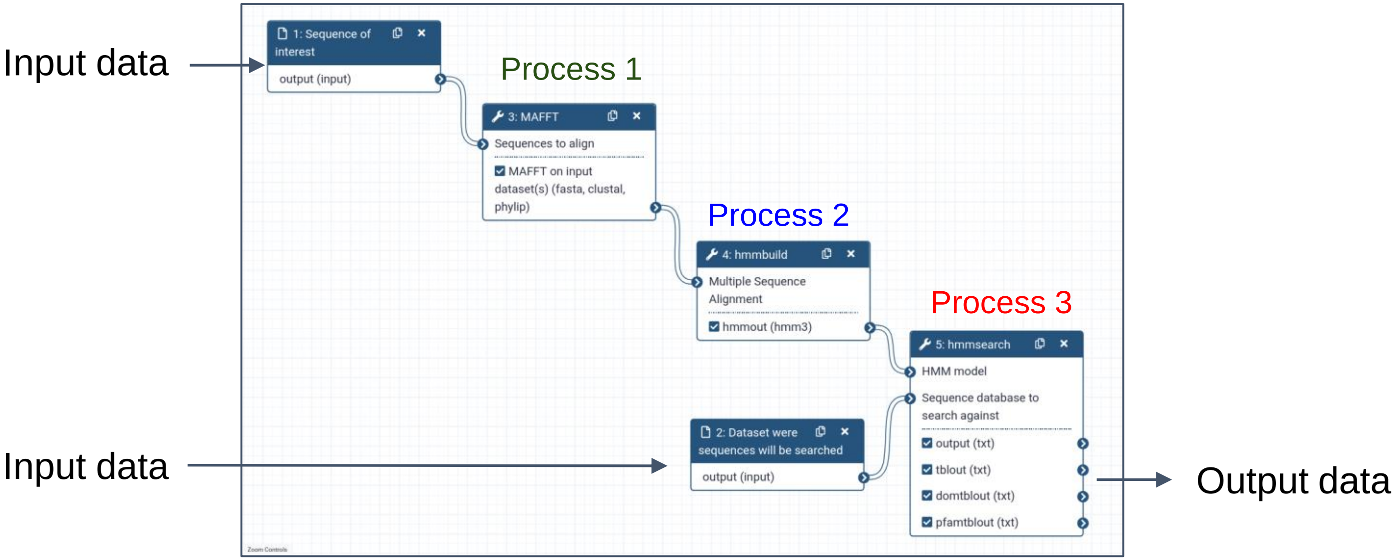
Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

- Several biocontainers have already been developed for bio-prospecting and engineering, including the AHATool, EP-Pred and AsiteDesign ones.



Galaxy: an open tool to create workflows



Albert Cañellas

Task 2.4: Explanation of the work carried

Progress undertaken and outputs achieved M18-M24

Event 7 host in BSC

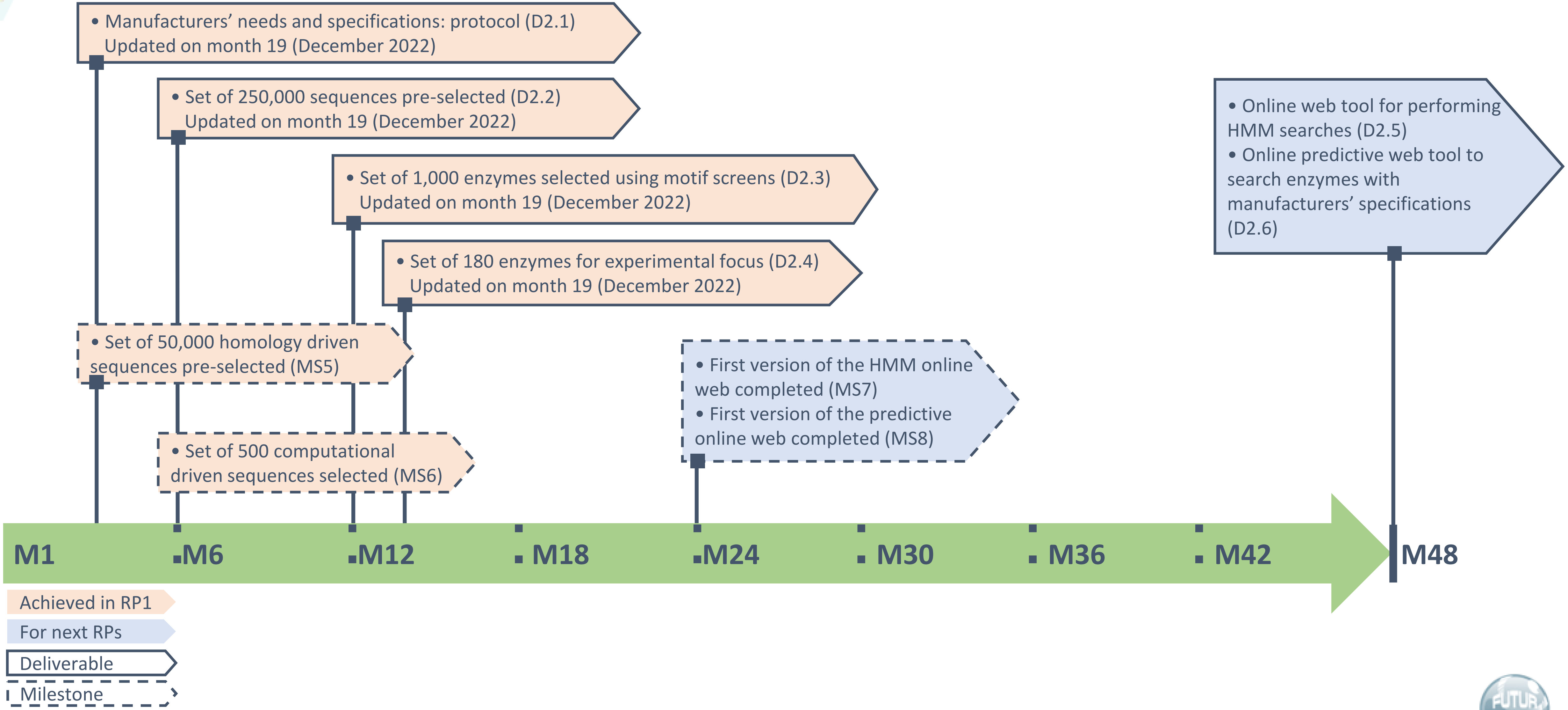
On the 22nd and 23rd of may, we had the privilege to host the First BSC workshop on Computational Enzyme Bioprospecting and Engineering

Conclusions

Key bullet points (scientific)

- Joint efforts are being made to design an intelligent bioprospecting workflow based on iterations.
- A second bioprospecting iteration is running to find a more active and more stable lipase and active hyaluronidases than the best ones found in the first iteration.
- New predictors and bioinformatic tools are being developed, and parallely, with efforts to make them available through both biocontainers and web servers.

WP2 – Deliverables and milestones



- Manufacturers' needs and specifications: protocol (D2.1)
  Updated on month 19 (December 2022)

- Set of 250,000 sequences pre-selected (D2.2)
  Updated on month 19 (December 2022)

- Set of 1,000 enzymes selected using motif screens (D2.3)
  Updated on month 19 (December 2022)

- Set of 180 enzymes for experimental focus (D2.4)
  Updated on month 19 (December 2022)

- Set of 50,000 homology driven sequences pre-selected (MS5)

- Set of 500 computational driven sequences selected (MS6)

- First version of the HMM online web completed (MS7)
- First version of the predictive online web completed (MS8)

- Online web tool for performing HMM searches (D2.5)
- Online predictive web tool to search enzymes with manufacturers' specifications (D2.6)

Achieved in RP1

For next RPs

Deliverable

Milestone

WP2 - Expected and achieved outputs

Key bullet points

- Key bullet points (non-scientific)
  - A total of 16.62 P/M (out of 51 total, a 32.59%) at M18.
  - The work plan is proceeding as planned (see **Table 2.1**)

Table 2.1. Brief summary of clear and measurable details and achievements in WP2 (as in the GA)

Name of activity	Achieved	Achievement (%)	Status
1 Protocol with manufacturers' needs & specifications (D2.1)	Yes	100%	Completed
250,000 Sequences pre-selected (D2.2)	Yes	1300%	Open
1,000 Enzymes selected using motif screens (D2.3)	Yes	140%	Open
180 Enzymes for experimental focus (D2.4)	Yes	377%	Open
Online web tool for performing HMM searches (D2.5)	Partially	40%	Open
Online predictive web tool to search project enzymes (D2.6)	Partially	20%	Open
Deliverables (4, at M24)	Yes	100%	Completed
Milestones (4, at M18)	Yes	100%	Completed

WP2 – Future actions

- Future actions (six months ahead)
  - Continue new rounds of enzyme bio-prospecting, if needed
  - Integrate meta-data to find correlations between computationally predicted parameters and enzyme parameters, and further integrate the different bio-containers being developed into a graphical web application (Galaxy already ongoing), to guide robust pre-selection based on those calculations.

WP2 – Deviations

- Deviations
  - No deviations found in the activities planned in the GA, and no mitigation actions are required.

WP2 – Highlights from the review report RP1

- Highlights
  - No criticisms or concerns.

FuturEnzyme Technologies of the FUTURe for low-cost ENZYMEs for environment-friendly products

FuturEnzyme: 2nd annual meeting
Start date: 1 June 2021 - End date: 31 May 2025
Proposal number: 101000327 - Consortium: 16 partners
Requested EU Contribution: 5,995,035.13 €

Project funded by the European Union’s Horizon 2020
Research and Innovation Programme under grant agreement No [101000327]