OPTIMISING THE ENZYME DISCOVERY PATHWAY THROUGH OMICS AND AI

A Data Management Plan created using DMPonline.be

Creator: Jaldert François

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1SE4623N

ID: 198084

Start date: 01-11-2022

End date: 31-10-2024

Project abstract:

Many industrial processes require harmful compounds. In the last decades enzymes became a sustainable and environmentally friendly alternative, replacing chemical catalysts. However, using enzymes is challenging due to the extreme conditions required in industrial applications. The demand for improved enzymes keeps rising yet enzyme screening remains costly and time-consuming. Computational enzyme screening can drastically reduce cost and speed up the discovery of industrial relevant enzymes. Yet, computational screening methods remain underdeveloped. We aim to develop a toolbox for computational enzyme selection. Additionally, we will develop new machine learning models for enzyme similarity search and parameter prediction. We will also build a specialized database tailored towards enzyme selection by integrating public databases, taking FAIR data principles into account. This database will be valuable for both discovering new candidate enzymes and developing new machine learning models. Our toolbox will advance computational enzyme screening and likely lead towards the discovery of new enzymes with industrial application through our collaboration with other research groups and industry. In addition, we will make our toolbox and subsequent database available to our partners and industry, enhancing their ability to identify new enzymes.

Last modified: 21-04-2023

OPTIMISING THE ENZYME DISCOVERY PATHWAY THROUGH OMICS AND AI
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

This project will reuse existing datasets from public databases for the development of a toolbox/database. This data includes sequences and experimental results such as enzyme reactions, validated pathways, and protein annotations. Additionally, during the project Python and R code will be developed for machine learning and data analysis. SQL code will be developed for the database, in combination with a Python code for the toolbox.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. The person responsible for data preservation will be Prof. Vera van Noort.
2. The storage capacity for data is 5TB (excluding code, which will be stored in KU Leuven Gitlab repositories). Data will be stored on the KU Leuven network drive (L-drive) both during and after the project. KU Leuven offers long term storage solution through their L-drive for large quantities of data.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

Some data (see "development group") is readily available through public databases (without the need of data cleaning steps). While this data is necessary for modelling and analysis, users can easily access this data through the public databases. As such, a DOI, link and version (or date) is sufficient to access the original data used from those databases.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Not applicable.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

Not Applicable.

OPTIMISING THE ENZYME DISCOVERY PATHWAY THROUGH OMICS AND AI
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

OPTIMISING THE ENZYME DISCOVERY PATHWAY THROUGH OMICS AND AI
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

OPTIMISING THE ENZYME DISCOVERY PATHWAY THROUGH OMICS AND AI
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Enzymares SQL database model	SQL file outlining the database structure. This file is required to build the database locally, and consists of SQL commands to create the SQL table structure.	• Generate new data	• Digital	• Other	• .sql • .png	• <100MB	
Enzyme pH annotation	Enzyme pH dataset for optimal pH and range. This file is required for pH modelling and prediction, and includes pH, sequence and source/evidence data.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv • .tab	• <1GB	
Enzyme temperature annotation	Enzyme temperature dataset for enzyme temperature optimal/range. This file is required for temperature modelling and prediction, and includes temperature, sequence and source/evidence data.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv • .tab	• <1GB	
Enzyme sequence properties and features	This dataset will contain sequence features/embeddings generated with ProtTrans, bio_embeddings, protlearn and peptides (R). These protein embeddings (vectors) are necessary to make sequences machine learning ready.	• Generate new data	• Digital	• Simulation data	• .csv • .tab	• <100GB	
Enzyme structural features	Protein structure based features for machine learning.	• Generate new data	• Digital	• Simulation data	• .csv • .tab	• <100GB	
Enzyme reaction/EC number annotation	Dataset with enzyme/sequence reactions, EC number and meta-data (phylum, source, etc.) for enzyme classification.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv • .tab	• <100GB	
Enzyme substrate annotation	A dataset containing enzyme—substrate annotation for machine learning models. The dataset includes sequence/organism annotation, substrate annotation/info, structure annotation.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv • .tab	• <100GB	
Sequence annotation	A dataset with sequence annotation including phylogenetic information and enzyme level data.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv • .tab	• <100GB	
Protein structures	Protein 3D structure data collected from open access databases.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv • .tab	• <1TB	

UniProt Annotation	UniProt JSON annotation file.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .json	- <100GB	
UniProt Sequence Data	UniProt fasta files.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .fa - .fasta	- <1TB	
Brenda Annotation	Brenda JSON annotation file.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .json - .txt - .dat	- <1GB	
Expasy Enzyme annotation	Data from Expasy Enzyme on ec numbers.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .txt - .dat	- <1GB	
IntEnz annotation	Data from intEnz on ec classification.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .dat - .xml	- <1GB	
Emble Ebi Enzyme portal data	Enzyme annotations on reactions, pathways, protein function and cofactors from embl ebi.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .tab - .csv - .json	- <100GB	
SabioRK annotation file	Enzyme annotations on kinetics and temperatures from Sabio-RK.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .tab - .csv - .json	- <100GB	
Enzyme reaction and metabolic pathway dataset	Enzyme reaction annotations.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv - .json	- <100GB	
Enzyme reaction dataset	Reaction annotations for enzymes with SMART/SMILES representations.	- Reuse existing data - Generate new data	- Digital	- Compiled/aggregated data - Simulation data	- .tab - .sdf - .rdf	- <100GB	
Enzyme domain dataset	Dataset with enzyme domains, families and functional sites.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .tab - .json - .csv	- <100GB	
Organism growth and phenotype dataset	Dataset with growth conditions and phenotypic data for mono cellular organisms.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .tab - .csv - .json	- <1GB	
Organism Taxonomy	Taxonomy dataset for relevant organisms.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .tab - .csv - .json	- <100GB	
Organism environment annotation	Dataset with information about organism environements.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .tab - .csv - .info	- <100GB	
Marine organism data	Dataset with marine organisms and related data.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <100GB	
Ontology definition dataset	File containing ontology definitions and references.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <1GB	
Organism db dataset	A file containing the organism data to load to the SQL database.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <100GB	
feature db dataset	A file containing sequence/enzyme features such as domains, family, binding sites, etc. to load the SQL database.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <1GB	
structure db dataset	A file containing the structure fractions and annotations to load the SQL database. This includes derived data such as secondary structure fractions and bonds.	- Reuse existing data - Generate new data	- Digital	- Compiled/aggregated data - Simulation data	- .csv	- <1GB	
GO_terms db dataset	A file containing the GO terms annotations to load the SQL database.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <1GB	
reaction db dataset	A file containing the reaction annotations to load the SQL database.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <1GB	
external_link db dataset	A file containing crosslinks between datasets to load the SQL database.	- Reuse existing data	- Digital	- Compiled/aggregated data	- .csv	- <1GB	

temperature db dataset	A file containing organism, environment, and sequence temperature data to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
ecnumber db dataset	A file containing ecnumber annotations with links to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
ecnumber_organism db link dataset	A file containing organism ecnumber links to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
kinetics db dataset	A file containing kinetics annotations for sequences and reactions to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
reaction_products db dataset	A file containing reaction participants (products) to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
ECnumber_organism_sequence db dataset	A file containing links between organisms, sequences and organisms to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
sequence db dataset	A file containing sequences and related statistics to load the SQL database.	• Reuse existing data • Generate new data	• Digital	• Compiled/aggregated data • Simulation data	• .csv	• <100GB	
ph db dataset	A file containing ph related annotations for reactions, sequences, organisms and environments, to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
reaction_participants db dataset	A file containing reaction participant annotations (substrates) to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
source db dataset	A file containing source annotations for each table entry in the database, to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
inhibitor db dataset	A file containing inhibitors and annotations for sequences, ecnumbers/organisms with kinetics information to load the SQL database.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <1GB	
text mining dataset	A file containing text mining results for different pipelines, organisms and sequences.	• Reuse existing data	• Digital	• Compiled/aggregated data	• .csv	• <100GB	
Enzymares database/toolbox Code	Code	• Generate new data	• Digital	• Software	• NA - code	• <100MB	
Enzyme Parameter Prediction	Code	• Generate new data	• Digital	• Software	• NA - code	• <100MB	
Enzyme classification and substrate specificity	Code	• Generate new data	• Digital	• Software	• NA - code	• <100MB	
Enzymatic screening dataset	A dataset with results of enzymatic and proteomic screening for selected marine organisms.	• Generate new data	• Digital	• Experimental	• .csv	• <1GB	
Genome/Transcriptome screening	A dataset with genomes/transcriptomes of selected microalgae.	• Generate new data	• Digital	• Experimental	• .csv • .bam	• <100GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

The "enzyme pH annotation" and "Enzyme temperature annotation" datasets use existing data from Brenda and Sabio-RK. Brenda DOI: 10.1093/nar/gkaa1025 URL: https://www.brenda-enzymes.org. Sabio-RK DOI: 10.1093/nar/gkx1065 URL: http://sabio.h-its.org/.

The "enzyme substrate annotation" dataset will use a combination of Rhea (DOI: 10.1093/nar/gkab1016 URL: https://www.rhea-db.org/), GO database (DOI: 10.1038/75556 URL: http://geneontology.org/), UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/), PubChem (DOI: 10.1093/nar/gkac956 URL: https://pubchem.ncbi.nlm.nih.gov/), ChEBI (DOI: 10.1093/nar/gkv1031 URL: https://www.ebi.ac.uk/chebi/), KEGG (DOI: 10.1093/nar/28.1.27 URL: https://www.genome.jp/kegg/) and/or M-CSA (DOI: 10.1093/bioinformatics/bti693 URL: https://www.ebi.ac.uk/thornton-srv/m-csa/). The "Sequence annotation" dataset will use UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/) and NCBI Taxonomy (DOI: 10.1093/database/baaa062 URL: https://www.ncbi.nlm.nih.gov/taxonomy) as resources. The "Protein structures" data will contain data from AlphaFold Protein Structure Database (DOI: 10.1093/nar/gkab1061 URL: https://alphafold.ebi.ac.uk/), PDB (DOI: 10.1093/nar/28.1.235 URL: https://www.rcsb.org/) and UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/).

The following datasets are sources for the database construction described in WP1.

- Both datasets "UniProt Annotation" and "UniProt Sequence Data" will use data from UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/). "Brenda Annotation" is collected from Brenda (DOI: 10.1093/nar/gkaa1025 URL: https://www.brenda-enzymes.org).
- The "Expasy Enzyme annotation" dataset will reuse data from ExpasyEnzyme (DOI: 10.1093/nar/28.1.304 URL: https://enzyme.expasy.org/).
- The "IntEnz annotation" dataset will use IntEnz (DOI: 10.1093/nar/gkh119 URL: https://www.ebi.ac.uk/intenz/) as resource.
- The "Emble Ebi Enzyme portal data" will use data from Enzyme Portal (DOI: 10.1093/nar/gks1112 URL: https://www.ebi.ac.uk/enzymeportal/).
- The "SabioRK annotation file" will be collected from Sabio-RK (DOI: 10.1093/nar/gkx1065 URL: http://sabio.h-its.org/).
- The "Enzyme reaction and metabolic pathway" datatype will reuse data from MetaCyc (DOI: 10.1093/nar/gkx935 URL: https://metacyc.org/), Rhea (DOI: 10.1093/nar/gkab1016 URL: https://www.rhea-db.org/), RetroRules (DOI: 10.1093/nar/gky940 URL: https://retrorules.org/).
- The "Enzyme reaction dataset" will use Rhea (DOI: 10.1093/nar/gkab1016 URL: https://www.rhea-db.org/), RetroRules (DOI: 10.1093/nar/gky940 URL: https://retrorules.org/) and ChEBI (DOI: 10.1093/nar/gkv1031 URL: https://www.ebi.ac.uk/chebi/) as resources, in combination with the rdkit for SMILES/SMART representations.
- The "Enzyme domain dataset" will be collected from PROSITE (DOI: 10.1093/nar/gkp885 URL: https://prosite.expasy.org), InterPro (DOI: 10.1093/nar/gkac993 URL: https://www.ebi.ac.uk/interpro/) and Cathdb (DOI: 10.1186/1479-7364-4-3-207 URL: http://www.cathdb.info).
- The "Organism growth and phenotype dataset" will include data from BacDive (DOI: 10.1093/nar/gkab961 URL: https://bacdive.dsmz.de), ThermoBase (DOI: 10.1371/journal.pone.0268253 URL: http://togodb.org/db/thermobase), TEMPURA (DOI: 10.1264/jsme2.ME20074 URL: http://togodb.org/db/tempura), AciDB (DOI: 10.1093/bioinformatics/btaa638 URL: https://acidb.cl) and Engqvist organism growth data (DOI: 10.5281/zenodo.1175609 URL: https://zenodo.org/record/1175609#.ZDxArS8Rq14).
- The "Organism Taxonomy" data will be collected from NCBI taxonomy (DOI: 10.1093/database/baaa062 URL: https://www.ncbi.nlm.nih.gov/taxonomy).
- The "Marine organism data" will reuse data from from OBIS (URL: https://obis.org).
- The "Ontology definition dataset" will be collected from GeneOntology (DOI: 10.1093/nar/gkaa1113 URL: http://geneontology.org) and/or OLS (DOI: 10.1093/nar/gkac240 URL: https://www.ebi.ac.uk/ols/docs/about).

The following datasets will be used to build the SQL database model, and use similar sources as the aforementioned datasets.

- The "Organism db dataset" will use NCBI taxonomy (DOI: 10.1093/database/baaa062 URL: https://www.ncbi.nlm.nih.gov/taxonomy) as a resource. For "feature db dataset" a combination of sources from "Enzyme domain dataset" and UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/) will be used.
- The "GO_terms db dataset" will use the same sources as the "Ontology definition dataset".
- The "reaction db dataset" will use Rhea (DOI: 10.1093/nar/gkab1016 URL: https://www.rhea-db.org/) and RetroRules (DOI: 10.1093/nar/gky940 URL: https://retrorules.org/).
- The "external_link db dataset" will include data from all of the aforementioned databases (using database provided crosslinks to other databases).
- The "temperature db dataset", "kinetics db dataset" and "ph db dataset" will use Brenda (DOI: 10.1093/nar/gkaa1025 URL: https://www.brenda-enzymes.org) as resource.
- The "ecnumber_db dataset", "ecnumber_organism db link dataset", "inhibitor db dataset" and "ECnumber_organism_sequence db dataset" will use a combination of ExpasyEnzyme (DOI: 10.1093/nar/28.1.304 URL: https://enzyme.expasy.org/), Brenda (DOI: 10.1093/nar/gkaa1025 URL: https://www.brenda-enzymes.org) and UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/).
- The "reaction_products db dataset" and "reaction_participants db dataset" will use Rhea (DOI: 10.1093/nar/gkab1016 URL: https://www.rhea-db.org/) as primary source in combination with ChEBI (DOI: 10.1093/nar/gkv1031 URL: https://www.ebi.ac.uk/chebi/).
- The "sequence db dataset" will use UniProt (DOI: 10.1093/nar/gkac1052 URL: https://www.uniprot.org/) as it's resource.
- The "text mining dataset" will use PMC (URL: https://www.ncbi.nlm.nih.gov/pmc/).
- The "source db dataset" will use all aforementioned databases, storing their link and entry identifier if available.

Both "Enzymatic screening dataset" and "Genome/Transcriptome screening" will be provided by consortium members of Enzymares, and generated in the lab. The datasets/datatypes for this project can be divided in 3 groups:

- **Machine learning group**: "Enzyme pH annotation", "Enzyme temperature annotation", "Enzyme sequence properties and features", "Enzyme structural features", "Enzyme reaction/EC number annotation", "Enzyme substrate annotation", "Sequence annotation" and "UniProt Sequence Data". These datasets will be used for machine learning model development.
- **Development group:** "Protein structures", "UniProt Annotation", "Brenda Annotation", "Expasy Enzyme annotation", "IntEnz annotation", "Emble Ebi Enzyme portal data", "Emble Ebi Enzyme portal data", "SabioRK annotation file", "Enzyme reaction and metabolic pathway dataset", "Enzyme reaction dataset", "Enzyme domain dataset", "Organism growth and phenotype dataset", "Organism Taxonomy", "Organism environment annotation", "Marine organism data", "Ontology definition dataset". These datasets are used in the development of machine learning models or the database tables/data, and involve copyright protection.
- **Database group:** "Organism db dataset", "feature db dataset", "structure db dataset", "GO_terms db datasetGO_terms db dataset", "reaction db dataset", "external_link db dataset", "temperature db dataset", "ecnumber db dataset", "ecnumber_organism db link dataset", "kinetics db dataset", "reaction_products db dataset", "ECnumber_organism_sequence db dataset", "sequence db dataset", "ph db dataset", "reaction_participants db dataset", "source db dataset", "inhibitor db dataset", "feature db dataset", "structure db dataset", "GO_terms db dataset", "reaction db dataset", "external_link db dataset", "temperature db dataset", "ecnumber db dataset", "ecnumber_organism db link dataset", "kinetics db dataset", "reaction_products db dataset", "ECnumber_organism_sequence db dataset", "sequence db dataset", "ph db dataset", "reaction_participants db dataset", "source db dataset", "inhibitor db dataset" and "text mining dataset". These datasets will be used for building the database and toolbox.

We will refer to these group names to keep this document comprehensive and concise.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

There is potential for commercial valorization concerning "Enzymares database/toolbox Code". This code repository provides the functionalities for an enzyme discovery/selection toolbox, and also includes the "Enzyme classification and substrate specificity", "Enzyme Parameter Prediction" and "Enzymares SQL database model" datasets. Additionally, all datasets with "db" in their name are required for the database, and as such are part of this commercial valorization potential. There are no restrictions for commercial exploitation concerning "Enzymares database/toolbox Code", "Enzyme classification and substrate specificity" and "Enzyme Parameter Prediction". However, restrictions apply to "Enzymares SQL database model" and all it's included datasets (referred to as "db" in their name). These datasets use public databases that are subjected to copyright and licencing agreements. The public databases used as resource include but are not restricted to: (1) UniProt (CC BY 4.0), (2) NCBI Taxonomy (GNU General Public Licence V3), (3) InterPro (CCO 1.0), (4) Cathdb (CC BY 4.0), (5) GeneOntology (CC BY 4.0), (6) OLS (CCO 1.0), (7) Rhea (CC BY 4.0), (8) Brenda (CC BY 4.0), (9) ExpasyEnzyme (CC BY 4.0), (10) M-CSA (CC BY 4.0) and (11) ChEBI (CC BY 4.0). As such, the use of public databases will limit commercial exploitation.

Additionally, users can opt to include additional (public) databases for the toolbox, for which code will be provided. However, they will be responsible for obtaining the necessary database licenses. Moreover, copyright and licenses can vary for different data entries because they are extracted from different public databases, similar to RetroRules licensing (https://retrorules.org/dl). The sharing of both "Enzymatic screening dataset" and "Genome/Transcriptome screening" will also be restricted and licensing/data sharing agreements will depend on the Enzymares consortium partners (what exact restrictions apply).

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Both "Enzymatic screening dataset" and "Genome/Transcriptome screening" datasets are created in a lab outside our research group, and as such are subjected to research transfer agreements. Researchers from the Enzymares consortium will retain all rights to the data. We will have access to this data for machine learning purposes, and will be allowed to incorporate this data in the toolbox given proper reference/citation to their work. The research transfer agreements will be finalized by the Enzymares consortium coordination partners.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

For documentation and metadata purpose we will make use of **project and data level documentation.**

At project level:

- A README file will be provided to keep track of the project structure and versions (what dataset, where, created when).
- Each WP will be accompanied by a README to keep track of dataset versions, origin, copyright/licencing, data cleaning and extraction steps, and repositories (code).
- Repositories will have an individual README to outline the code structure and functions, to make scripts reusable.
- Python code will be documented with the Sphinx or Google docstrings format (in the code files).

At data level:

- For each dataset (and database table) a data dictionary will be provided. This data dictionary will at least include (but not limited to) (1) variable names (2) human readable variable names (3) measurement units (4) allowed values (5) variable definitions (6) variable synonyms (7) variable description (8) comments. A REAME file will be used to document the data cleaning steps.
- The database model(s) will be accompanied by (1) an automatic generated UML diagram (2) DDL file (3) a README file in human understandable language outlining design choices and limitations. The variable names and functions will be captured in a data dictionary.

The README files will follow the KU Leuven Research Data Management guidelines and template.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

At project level:

- The RDR metadata format will be used.

At data level:

- Both DDL and UML standard syntax will be used.

The README files will outline data cleaning steps and database design choices.

3. Data storage & back-up during the research project

Where will the data be stored?

Code and accompanying README files will be stored in KU Leuven Gitlab repositories. All datasets, documentation and metadata will be stored on the research group L-drive within the secure KU Leuven environment. Access to this drive is limited to the PI and involved researchers.

Code and datasets will be changed and/or used locally or stored in the KU Leuven Onedrive (provided for each staff member). Any changes or updates will be uploaded daily to the L-drive and KU Leuven Gitlab.

How will the data be backed up?

KU Leuven provides automatic backup for their L-drives. Automatic version management occurs when storing data in the KU Leuven datacenters. "Snapshot" technology is used for version management, which keeps the previous versions of the changed file online on the same storage system. Additionally, KU Leuven mirrors these files to a second ICTS data center to enable disaster recovery in case of problems. Files are versioned once a day, and version are kept for 14 days. KU Leuven Gitlab ensures automatic backups, and all file versions are permanently stored.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

Our research group currently has an L-drive with a capacity of 5 TB for research. This should be sufficient for storage of the generated datasets. In case of capacity problems, the L-drive capacity can be expanded or data can be stored on the KU Leuven OneDrive for staff (4 TB).

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The KU Leuven network drives are incorporated within secured KU Leuven environments, are password-protected (including multi-factor identification) and are only accessible by involved researches. The PI can add or remove access for researchers to the network drive. A separate folder will be created to store our datasets. Only the PI and registered collaborating researchers will have access to this folder.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The L-drive storage has a yearly cost of 104,42/TB/year and is covered by the research group. Storage for more than 5 years after the project can be covered by the research group/department. Data will be maintained for at least 10 years after the project in line with KU Leuven RMD policy. Additional funding from the Enzymares consortium is available for long-term data storage if needed.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

Digital data: All data in the "machine learning group" and "database group" (see earlier) will be archived for minimally 10 years after the project's completion, in line with the KU Leuven RDM policy. The data from the "development group" contains data that is readily available from public databases, as such not stored permanently.

Where will these data be archived (stored and curated for the long-term)?

The generated research data, metadata and documentation necessary to reuse or generate the data will be maintained on the L-drive (LVS Network Drive) for long-term data archiving, managed by KU Leuven ICTS with automatic back-up procedures.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

The current costs for the L-Drive are € 104.42 / TB / year. The absolute upper limit of the storage capacity needed will be 3.7 TB. However, our expected storage capacity needed will be 1TB - 2.5TB, with an estimated yearly cost of € 104.42 to € 260.06 yearly. These costs will be covered by the research group (KU Leuven funding), but additional funding from the Enzymares consortium is available.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)

Datasets created for the development of machine learning models will be made available through https://zenodo.org/ or through KU Leuven RDR repository. Code for the development and application of machine learning models will be released through KU Leuven Gitlab or GitHub. Additionally, dataset version and metadata will be released to enable other researchers to extract similar data from public databases. Additionally, large datasets used to build the database can be accessed upon request through the L-drive, or will be shared with KU Leuven ManGO platform (during the active phase of the project). In short, both "database group" and "machine learning group" data will be shared and will be freely accessible. Data from the "development group" is readily available through public databases, and won't be shared due to copyright/license restrictions. This data can be shared upon request if this does not violate the databases copyright/license. Additionally, date/time, version, and possible cleaning steps (with code) will be made opensource enabling users to download and construct the "development group" datasets locally.

Datasets and code for the development of the database/toolbox will be released through open access repository. An expanded database/toolbox can be generated by the user through provided code if they manage the licence agreements themselves, but won't be shared due to copyright. Data from the "Enzymares Consortium" will be shared in this database under restricted access based on their data sharing agreement.

If access is restricted, please specify who will be able to access the data and under what conditions.

Both "Enzymatic screening dataset" and "Genome/Transcriptome screening" contain novel data generated by research partners from the Enzymares consortium, as such they retain all rights to this data. These datasets can be shared after approval by involved research partners. Datasets from the "development group" are downloaded from public databases and sharing is license/copyright restricted. Researchers will be granted access to these datasets upon request and if this does not violate license or data transfer agreements.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

Datasets from all 3 groups (machine learning, development and database group) contain data from public databases and are restricted by copyright/license agreements. Both "machine learning group" and "database" data can be shared due to opensource licenses. However, datasets contained in the "development group" are restricted from sharing due to copyright/licenses. Additionally, both "Enzymatic screening dataset" and "Genome/Transcriptome screening" will be obtained with research transfer agreements, and as such cannot be shared without restrictions.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Data will be made available through KU Leuven RDR, Zenodo, KU Leuven Gitlab and/or ManGO. All code will be made available through KU Leuven Gitlab repositories. Datasets used in publications will be shared either through KU Leuven RDR or Zenodo ("machine learning group"). Data used for the toolbox/database construction will be shared either through ManGO or KU Leuven RDR ("database group"). Datasets from the "development group" can be shared upon request through ManGO for external users.

When will the data be made available?

Data will be made available for the machine learning models upon publication and/or project end. Other datasets will be made available upon WP completion or at the end of the project.

Which data usage licenses are you going to provide? If none, please explain why.

Code will be shared under GNU General Public License. Datasets will be shared under CC BY 4.0 or with an individual database dependent license (public database restricted). The toolbox/database can have spin-off potential, and will likely be shared under CC BY-NC-SA (relevant datasets and code).

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

What are the expected costs for data sharing? How will these costs be covered?

There are no expected costs for data sharing, as RDR, Zenodo and GitLab/GitHub are free of charge. Costs for sharing large datasets through ManGO (35/TB/year) will likely not exceed 1TB/2TB and will be covered by the research group.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The PhD researcher (Jaldert François) will be responsible for data documentation and metadata during this research project.

Who will manage data storage and backup during the research project?

The PhD researcher (Jaldert François) will be responsible for data storage and backup during this research project.

Who will manage data preservation and sharing?

The PhD researcher (Jaldert François) will be responsible for data preservation and sharing. The PI (prof. dr. Vera van Noort) will be responsible for data preservation and sharing past this projects duration.

Who will update and implement this DMP?

The PhD researcher (Jaldert François) will be responsible for updating and implementing this DMP. The PI (prof. dr. Vera van Noort) will monitor the DMP implemetation.