Plan Overview

A Data Management Plan created using DMPonline.be

Title: AI-Driven Analysis and Design of Protein-Ligand Interactions for Advancing De Novo Drug Discovery

Creator: Robin Poelmans

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Project abstract:

Over the past years, deep learning methods have claimed an increasingly important place in the field of computer-aided drug design. However, since these models are usually exclusively trained on common drug targets such as kinases and GCPRs, they often have difficulties in generalizing to unexplored target proteins. Therefore I want to create an unbiased dataset for the binding of small ligands to underexplored protein pocket clusters in the human pocketome. To achieve this, I will chart the human pocketome using a novel hypervoxel descriptor method developed by the LBMD lab, and attempt to link this with the ligand chemical space by creating a wormhole algorithm. Subsequently, by using computationally designed protein scaffolds, I will design a set of protein pockets representative for the underexplored clusters in the human pocketome. Using the wormhole algorithm, I will match compounds from the chemical library at the LBMD lab to these pockets, and further optimize these pockets to create specific small molecule binders. I will then experimentally evaluate the binding of these chosen ligands and their derivatives to all the designed pockets using biolayer interferometry, in this way creating a low-sparsity and high-diversity dataset ideal for the training of deep learning methods. This dataset will then finally be released as a blind predictive community challenge within the SAMPL framework.

ID: 212134

Start date: 01-11-2024

End date: 31-10-2028

Last modified: 11-03-2025

AI-Driven Analysis and Design of Protein-Ligand Interactions for Advancing De Novo Drug Discovery
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**
PLINDER dataset	Raw data of protein-ligand interaction systems	Reuse existing data	Digital	Compiled/aggregated data	.cif, .mol2	• <1TB	/
PYkPocket training set	Pre-processed training data for joint manifold learning algorithm: e.g. cleaned structure files, voxel grids,...	Generate new data	Digital	Experimental	.cif, .csv, .pkl	• <5TB	/
PYkPocket software package	Repository with scripts for knowledge-based potentials, grid scoring and joint manifold learning algorithms	Generate new data	Digital	Software	.py	• <1GB	/
Designed protein sequences	DNA sequences for the modified SAKe constructs. Plasmids stored at -20°C, glycerol stocks stored at -80°C	Generate new data	Physical				Between 50 and 100 DNA constructs
Designed protein structures	Data from X-ray crystallography experiments	Generate new data	Digital	Experimental	.raw, .mtz, .cif	• <10TB	/
SPR data	Data from surface plasmon resonance experiments, measuring protein-ligand interactions	Generate new data	Digital	Experimental	.csv	• <1GB	/
SAMPL analysis scripts	Software to analyse predictions from SAMPL challenges	Reuse existing data	Digital	Software	.py	• <1GB	/
SAMPL raw prediction data	Compound rankings and binding poses sent in by competitors	Reuse existing data	Digital	Compiled/aggregated data	.mol2, .csv	• <1TB	/
SAMPL analysed results	Analysis of received prediction results	Generate new data	Digital	Experimental	.csv, .png	• <10GB	/

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:

- PLINDER: https://console.cloud.google.com/storage/browser/plinder
  https://github.com/plinder-org/plinder

  We make use of (a subset of) PLINDER to obtain the raw training data for the PYkPocket models

- SAMPL: https://github.com/samplchallenges

  The SAMPL challenges have a well-established workflow for data analysis, so scripts from previous challenges can be re-used for our new challenge

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

The designed protein-ligand complexes could potentially be valorised as biosensors or as inducible switches in regulatory gene networks for synthetic biology. However, these applications do not lie within the scope of this PhD project.
If the software developed in this project is found to perform very well, we can opt for paid licenses for commercial use.

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.

- No

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).

GitHub repositories will be created for both the PYkPocket and the SAMPL projects. These repositories will contain the raw data, processed data and all software created for data analysis. Version control of the Python scripts will be maintained through these Git repositories. Python scripts within these repositories will be documented and commented to facilitate reuse of the code.
Raw data and processed data for the crystallography and SPR experiments will be stored locally. We will keep track of the generated files using a well-documented file tree, with each folder containing a README.txt file, specifying the data formats stored within a folder and how they should be processed.
Location and contents of glycerol stocks and plasmids are documented on a shared Benchling database in the LBMD lab.

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

The RCSB PDB and ChEMBL databases both make use of specific metadata requirements for contributions. The PDB database for example makes use of the Crystallographic Information Framework standards, which will be followed in our research.

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

Where will the data be stored?

Software, training data and analysis results for the PYkPocket and SAMPL projects will be stored in their respective GitHub repositories.

- PYkPocket: https://github.com/robin-poelmans/PYkPocket/tree/main
- SAMPL: https://github.com/samplchallenges

These GitHub repositories will be kept private until the date of publication.
Crystallography and SPR data will be stored locally on the systems of the LBMD lab. Back-ups of these systems to a Network-Attached Storage (NAS) will be performed at regular time intervals and after important / intensive calculations. This centralized storage can only be accessed by other desktops within the LBMD local network.
Plasmids for the designer proteins are stored at -20°C in the LBMD lab, while glycerol stocks are stored within the -80°C freezer of LBMD. Exact locations for the plasmids and glycerol stocks can be found through the lab's Benchling database.

How will the data be backed up?

Back-ups of these systems to a Network-Attached Storage (NAS) will be performed at regular time intervals (by means of a cron job) and after important / intensive calculations.

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

/

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

Since we work in Linux environments on the local LBMD systems, writing rights can easily be limited for other users within the LBMD system. Access to my own user is password-protected.
During the research, writing rights to the data will be limited to myself and my supervisor.
Writing rights for the GitHub repositories are limited to my own account, unless other collaborators are explicitly invited by me.

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

Since the regular back-ups to the NAS are already in place, there are no additional costs for back-ups. All storage costs are covered by the LBMD.

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...).

All data specified above can and will be maintained for at least five years. Digital data will be maintained in GitHub repositories and/or local NAS storage, as specified above.
Physical data (in the form of plasmids and glycerol stocks) will be stored within the LBMD lab, with their location specified in the LBMD Benchling database.

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

Data will be archived in dedicated GitHub repositories for the PYkPocket and SAMPL projects. Solved crystal structures will be uploaded to the PDB database.

- PYkPocket: https://github.com/robin-poelmans/PYkPocket/tree/main
- SAMPL: https://github.com/samplchallenges

Raw data from X-ray diffraction experiments will be stored on the local NAS for at least five years, but this will not be archived elsewhere due to the large size of these datasets.

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

No additional costs are expected, since storage in the GitHub repositories is free of charge and the local NAS system is already operational.

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

Both raw and processed data for the PYkPocket and SAMPL projects will be made openly available through their respective GitHub repositories, while solved crystal structures will be available through the PDB database. Protein-ligand interaction measurements from the SPR experiments will also be archived on ChEMBL.

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

Not applicable.

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.

- No

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

GitHub repositories:

- PYkPocket: https://github.com/robin-poelmans/PYkPocket/tree/main
- SAMPL: https://github.com/samplchallenges

When will the data be made available?

Data will be made available upon publication of research results.

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

In principle, all software produced during this PhD will fall under the MIT license. This license gives express permission for users to reuse code for any purpose.
If the software is found to outperform other drug design and/or protein design tools, we can opt for dual licensing, where academic use falls under the MIT license and business use falls under a commercial license.
The data from the crystallography and SPR experiments will fall under the CC0 license, as requested by their respective repositories.

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

GitHub repositories have a numeric ID which, unlike the repository URL, is a unique and persistent identifier. The repositories for PYkPocket and SAMPL will also be linked to their respective publications.

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

The archived data should be sufficiently small to archive free of cost on GitHub repositories.

6. Responsibilities

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

Data documentation will be managed by myself.

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

Data storage will be managed by myself. Backups of locally stored data are performed on a regular basis through cron jobs

Who will manage data preservation and sharing?

Archiving data on the Git repositories will be managed by myself. Preservation of locally stored data on the LBMD systems will be managed by the supervisor, Prof. Arnout Voet.

Who will update and implement this DMP?

This will be managed by myself, Robin Poelmans