# Plan Overview

*A Data Management Plan created using DMPonline.be*

**Title:** AI-guided computational design of anti-virulence proteins as sustainable alternative antibacterials

**Creator:** n.n. n.n.

**Principal Investigator:** n.n., First Name Surname, n.n., n.n.

**Data Manager:** n.n., Nick Geukens

**Affiliation:** KU Leuven (KUL)

**Template:** KU Leuven BOF-IOF

**Principal Investigator:** n.n. n.n., First Name Surname, n.n. n.n., n.n. n.n.

**Data Manager:** n.n. n.n., Nick Geukens https://orcid.org/0000-0001-5706-1072

**Project abstract:**

Antibiotics are some of the most essential drugs used in the clinic. The livestock industry also currently relies extensively on antibiotics for disease prevention and growth promotion. Significant ongoing concerns about the further emergence of antibiotic-resistant bacteria are thus a major medical, veterinary, and economic problem. In response to this pressing issue, we propose a groundbreaking project aimed at developing
sustainable biologic alternatives to antibiotics. Our approach builds on the recently developed SymBody platform (Voet lab), which are ultrastable proteins that do not provoke an immune response and that can be produced sustainably. They can be administered orally through the gastrointestinal track without degradation, as well as intravenously.
Rather than relying on classical display methods to optimize SymBodies, in this project we will combine protein engineering with machine learning experience to enhance the platform. This will result in an AI-based protein engineering method that can more rapidly design the desired protein for a given target. During the AI-design and evaluation phase, we will also take into account the optimization of the upscaling of the production
through bacterial fermentation. The Symbody-antibacterials (SymbAs) will be designed to be evolutionarily robust by following a double safety approach. On the one hand, the SymbAs will be iteratively designed by AI to interact with the target in such a way that it is more difficult for resistant mutants to originate. On the other hand, by focusing on a virulence factor as target we aim to reduce the selective pressure on resistant mutants if they still would originate, and as such avoid their spread and enrichment within the pathogen population. Indeed, contrary to traditional antibiotics that target essential enzymes, anti-virulence drugs targeting specific types of virulence factors (such as public virulence factors or coincidental virulence factors) are predicted to be less prone to resistance selection. In this project, we will design SymbAs targeting the bacterial virulence factor Sortase A of both human pathogens and those relevant to animal husbandry. SortA is a validated target that is present in most Gram-positive bacteria. Over several iterative rounds, we will create and validate optimized proteins that can neutralize the virulence of pathogenic bacterial strains (e.g. MRSA, VRE, E. cecorum, S. suis,...) as well as validate the virulence targets as evolutionarily robust.
Hence, this project will provide a platform to further develop SymbAs targeting other virulence factors for different pathogens, as well as the expertise in AI-based protein design for different synthetic biology projects.

**ID:** 214177

**Start date:** 01-10-2024

**End date:** 30-09-2028

**Last modified:** 14-04-2025

# AI-guided computational design of anti-virulence proteins as sustainable alternative antibacterials

**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.**

| Dataset name / ID | Description | New or reuse | Digital or Physical data | Data Type | File format | Data volume | Physical volume |
|---|---|---|---|---|---|---|---|
| | | *Indicate: **N**(ew data) or E(xisting data)* | Indicate: **D**(igital) or **P**(hysical) | Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify) | | Indicate: <1GB <100GB <1TB <5TB >5TB NA | |
| protein_models | 3D structures and sequences of designed protein complexes | N | D | SO | PDB | 2TB | |
| protein_structure | crystallographic determined structures | N | D | SO | PDB+MTZ | 2TB | |
| protein_biophys | expression, stability and affinity data of evaluted proteins | N | D | N | csv | 1GB | |
| protein_crystal | crystallisatie condities van proteines | N | D | IT | png, csv | 10GB | |
| scripts_cpd | all scripts related to CPD data generation, analysis and | N | D | T | py | <1GB | |
| scripts_bio | all scripts related to bioreactor optimisation | N | D | T | matlab | <1GB | |
| scripts_AI | all scripts related to AI modelling | N | D | T | py,C | <1GB | |
| AI models | AI based algorithms predicting sequences | N | D | T | py,C | <10GB | |
| protein samples | all proteins and bacterial targets and designer proteins) | N | D | experimental | | | ~1000 samples,plasmids and purified proteins in cryvials stored per 100 in boxes at -20/-80 degrees. |
| bioreactor protocols and results | all protocols and results from bioreactor optimisation expeirments | N | D | experimental | word,excell, csv | <7GB | |
| biopharmaceutical profiles | all analysis performed by pharmabs on developmentablity and immunogenecities of select designer proteins | N | D | experimental | word, excel, csv | <10GB | |
| Microbial cell counts via flow cytomtery | Read-out for all the bacterial competition and fitness assays (both in vitro and in vivo) | N | D | experimental | .xls | <1GB | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Microbial cell counts via CFU | Read-out for all the bacterial competition and fitness assays (both in vitro and in vivo) | N | D | experimental | .fcs | <100GB | |
| Optical density measurements | Measurements of bacterial population densities based on optical density | N | D | experimental | .xls | <100MB | |
| Whole genome sequencing | WGS data of the ancestral and selected evolved strains | N | D | experimental | .fasta/.dbam | <1TB | |
| RNA sequencing | RNA-sequencing of the ancestral and selected evolved strains | N | D | experimental | /fastq/.dbam | <1TB | |
| evolved bacteria | Evolved bacterial populations and isolated clones from the evolved population | N | D | experimental | | | ~1000 samples, stored in 96-well plate format |
| | | | | | | | |
| | | | | | | | |

**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:**

Question not answered.

**Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.**

- No

**Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).**

- 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 datasets created are not valuable on their own. The only potential valorisation lays in the final application of our AI-designed proteins which could become biopharmaceuticals. This will be covered by the sequences (part of **protein_models**) combined with the **protein_biophys** dataset.

**Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or 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

While the starting point of our reaearch has been patented (by the Voet-lab) we have full freedom to operate during this project.

### 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).**

1. The data will be accompanied by a detailed metadata txt file denoting important characteristics (e.g. strain, material, timepoint,..) necessary for interpretation of the results. The key characteristics will also be denoted in the filename of the data files, as well as in the benchling digital lab book files
2. A detailed experimental protocol will be added to the directory of the corresponding experimental results. This step-wise description will facilitate potential future reproduction of the experiments.

3. For every deliverable in the project, a general outline txt file will be created. This file provides an overview of all the available data, the design of the experiment and the structure of the data saving.

**Will a metadata standard be used to make it easier to find and reuse the data?**
**If so, please specify which metadata standard will be used.**

**If not, please specify which metadata will be created to make the data easier to find and reuse.**

- Yes

Sequencing data: MIxS metadata standard
Flow cytometry data: MiFlowMe metadatastanrd
For others we will follow the Dublin Core metadata standard and work with Benchling digital lab books.

### Data Storage & Back-up during the Research Project

**Where will the data be stored?**

- Shared network drive (J-drive)
- Personal network drive (I-drive)
- Large Volume Storage

Data will be both stored In I and J drives as well as NAS storage.
All digital data will be registered using an electronic lab notebook (ELN) based on the Sharepoint platform and backed up on the internal server of KU Leuven, which is maintained by the IT service of KU Leuven. Due to the size of the raw sequencing data, this data will be exempt from this rule and only be stored on an NAS. Proteins, bacterial strains and populations will be stored in a secured -80°C freezer at the labs of respectively prof . Voet(proteins) or Stenackers (bacterial strains).

**How will the data be backed up?**

- Standard back-up provided by KU Leuven ICTS for my storage solution
- Personal back-ups I make (specify below)

The Sharepoint is backed up three times a day. The internal servers are managed by the KU Leuven IT department and backed-up according to their procedures. The NAS is double backed up in the cloud via a Sinology provided service.

**Is there currently sufficient storage & backup capacity during the project?**

**If no or insufficient storage or backup capacities are available, explain how this will be taken care of.**

- Yes

We have budgetted extra drives for datastorage within this project.

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

1. Access to the digital data will be limited to members of the research group who contribute to the research.
2. Critical documents can be (temporarily) locked by the author(s).
3. Sharepoint provides a changelog for detecting and reverting possible unauthorized changes.
4. The internal storage provides a back-up for the sharepoint and vice versa.
5. Physical data is stored in a secured -80°C freezer at the facility with limited-access
6. Our NAS system is user protected. Only researchres and root/admin have access to the data.

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

Only the computational researchers will exceed capacity from what is provided by the kuleuven. This is covered by the NAS storage for which extra drives are acquired from the consumables of this project (500 euro for drives)

**Data Preservation after the end of the Research Project**

**Which data will be retained for 10 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 will be preserved for 10 years according to KU Leuven RDM policy

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

- Large Volume Storage (longterm for large volumes)
- Other (specify below)

The laboratory of the main promotor houses a NAS of 128Tb, and continously expanding to ensure all data can be saved. Some of the data will be released online as part of publications via online repositories e.g. the PDB for structural protein data. Here also data from the group of Yves Moreau will be stored as the researchers are jointly promoted.
Other researches will use the one drive / Sharepoint repositories provided by ICTS. All scripts will also be place don Github repositories. At the end the datawill
At the end of the project all data will be transferred to the NAS drives of the main promotor

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

The NAS data storage drives were budgetted within this project.

Data Sharing and Reuse

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

- Yes, as open data
- Other (specify below)
- No (closed access)

All data which is academic in nature, including negative results will be released. However as the ultimate goal is the develop a novel biopharmaceutical certain proteins and sequences may not be released to protect IP , or only after IP protection is in place. Certain specificities may never be released as part of an IP protection strategy.

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

All data will be accessible to the reseachers and promotors of the project during the course of the project. Common essential results will be stored on Sharepoint and a teams channel accessible to all involved. But more specialised data (eg. scripts and raw data) will be localy managed. When access is needed to other peoples dataset it will be made available through involved phd student, supervising postdoc or Promotor.

**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 per dataset or data type where appropriate.**

- Yes, intellectual property rights

In certain cases a designer protein contains a unique feature (e.g. the oligomerisation interface of a SymBody) which prevents renal clearance. As this is a unique features which gives us strategic advantage over potential competitors (after consultation with LRD) it was decided this mechanism could not be part of any future publication or patent to prevent me-too type workarounds.

**Where will the data be made available?**

**If already known, please provide a repository per dataset or data type.**

- KU Leuven RDR (Research Data Repository)
- Other data repository (specify below)

structural and sequential information will be released on the PDB database.

**When will the data be made available?**

- Upon publication of research results

All non confidential data will be made available via online repositories. e.g. PDB database also non succesfull desings will be released in order for other researchers to learn.

**Which data usage licenses are you going to provide?**

**If none, please explain why.**

- CC-BY 4.0 (data)
- GNU GPL-3.0 (code)
- Data Transfer Agreement (restricted data)

Some scripts will build upon GNU scripts which need to remain in the GNU format. Other formats will depends on agreement. This is of importance as we aim to establish an spin off or valorise the results and may want to transfer scripts and data.
Other non-valorisation sensitive scientific data will be available via CC-BY 4.0 or equivalent license types.

**Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.**

- Yes, a PID will be added upon deposit in a data repository

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

The sequence and structural databases are free. Also github is free.

**Responsibilities**

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

The researchers who generate the data are responsible for the correct documentation of the data and metadata.

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

The researchers who generate the data are responsible for the storage and back-up, supervised by the promotors of this project.

**Who will manage data preservation and sharing?**

Arnout Voet has the end responsibility and manages long term preservation and sharing for protein design biophysics and

structural biology, biopharmaceutical profiling as well as the computational scripts involved (the latter is shared with Yves Moreau)

Hans Steenackers has the end responsibility and manages long term preservation and sharing regarding (evolved) bacterial cultures and populations and the effects of the designer proteins on the virulence and resistance from this bacteria.

Christal Bernaerts has the end responsiblity and manages long term preservation and sharing regarding (evolved) bioreactor optimisation

**Who will update and implement this DMP?**

Arnout Voet