## DMP TITLE

RATIONAL DESIGN OF BIOLOGICS FOR THERAPEUTIC DEVELOPMENT (FOLDCO)

### ADMIN DETAILS

**Project Name:** Rational design of biologics for therapeutic development (FoldCo)

**Project Identifier:** S000722N

**Project type:** SBO/ES Strategic Basic Research-Economic Spinoff

**Principal Investigator / Researcher:** Frederic Rousseau

**Project Data Contact:** Béla Z Schmidt

**Description:** Biologics represent an increasingly important class of drugs yet biologic drug development still suffers from a high attrition rate, i.e. many molecules that enter the costly development pipeline fail before they reach the market because of inferior properties regarding manufacturability, long-term stability, immunogenicity. Furthermore, many antibody targets are hard to hit with current technologies. This project is aiming to overhaul the biologics development process by developing an integrated platform for the generation of innovative bio-therapeutics with a superior profile using in silico rational protein design methods.

The project brings together top-notch protein engineering expertise in Europe to build an integrated protein design and validation platform, based on our proprietary FoldX software suite and FoldX force field. The platform will rely on scoring functions that simultaneously calculate all the relevant parameters that need to be controlled during the antibody design process. In addition, our computational protein design pipeline will be extended with new modelling capacities and fully integrated into a professional high-performance platform.

We will use the platform to generate a comprehensive proprietary data package on high-profile cases of actual product development to allow for the valorisation of this highly innovative platform. The ultimate goal of the project is to launch FoldCo, the protein design company in Flanders, centred around our proprietary FoldX protein design platform. The company will have its own biotherapeutics pipeline combined with an active partnering strategy to maximally exploit the potential of the technology.

**Institutions:** KU Leuven, Universiteit Gent, Centre for Genomic Regulation (CRG, Barcelona, Spain)

### 1. GENERAL INFORMATION

**Name applicant**

Frederic Rousseau

**FWO Project Number & Title**

Application number: S000722N

**English Title** Rational design of biologics for therapeutic development (FoldCo)

**Dutch Title** Rationeel ontwerp van biologische geneesmiddelen voor therapeutische ontwikkeling (FoldCo)

**Affiliation**

* KU Leuven

### 2. DATA DESCRIPTION

**Will you generate/collect new data and/or make use of existing data?**

* Generate new data
* Reuse existing data

**Describe the origin, type and format of the data (per dataset) and its (estimated) volume, ideally per objective or WP of the project. You might consider using the table in the guidance.**

Please see data table in the following pages.

| **WP** | **Dataset** | **Purpose** | **New/Existing (source)** | **Data type** | **Data subtype** | **Data format** | **Size** | **Unit** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Scoring function algorithms | ranking candidate protein sequences | New data | Derived\_and\_compiled\_data | Research\_documentation | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 20 | MB |
| 1 | Observed Antibody Space (OAS) database | existing human antibodies - used to develop humaneness scroing function | Existing data | Canonical\_data | Protein\_sequences | Nucleotide and protein sequences: raw sequence data trace (.ab1), textbased format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk); | 100 | MB |
| 1 | Purified antibodies bought from Gescript (protein) | controls for biophysical characterization of candidate proteins | Existing data | Experimental\_data | Antibodies | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 24 | mg |
| 1 | Sequence of FDA-approved antibodies | modelling antibodies of known structure | Existing data | Canonical\_data | Protein\_sequences | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 1 | MB |
| 1 | PDB structures of FDA-approved antibodies | comparing modelled and actual structure | Existing data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 10 | MB |
| 1 | Post-translational modification databases | modelling post-translational databases | Existing data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 1 | GB |
| 1 | Structures of MHCII complexes with peptides | developing T-cell epitope scoring function | Existing data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 500 | MB |
| 1 | Human proteome sequences from the full human genome (for linear epitopes) | Developing humaneness scrorign function | Existing data | Canonical\_data | Protein\_sequences | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 20 | MB |
| 2 | Scoring function algorithms from WP1 | Integrating scoring fundtion into computational pipeline | Existing data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 20 | MB |
| 2 | Entire PDB with water molecules | Building computationally efficient way of modelling of structural water molecules | Existing data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 5 | GB |
| 2 | Optimized predcition algorithms | A computationally efficient way of modelling of structural water molecules | Existing data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 5 | MB |
| 2 | Algorithm to move the backbone | Modelling conformational flexibility | New data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 5 | MB |
| 2 | The structure of all protein structures that from complexes | Building protein docking algorithm | Existing data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 1 | GB |
| 2 | Fragment libraries generated from the PDB | Building protein docking algorithm | Existing data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 2 | GB |
| 2 | Algorithm for protein docking | Modelling protein docking | New data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 5 | MB |
| 2 | ProThem core dataset | Training neural network to estimate adjusted DeltaG | Existing data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); possibly in instrument-specific proprietary format | 100 | MB |
| 2 | SKEMPI interface dataset | Training neural network to estimate adjusted DeltaG | Existing data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); possibly in instrument-specific proprietary format | 100 | MB |
| 2 | AI algorithm | Produce a more realistinc estimation of DeltaG | Existing data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 100 | MB |
| 2 | Integrated pipeline for interface design | Designing antibodies with optimal CMC characteristics | New data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 100 | MB |
| 2 | Feedback of all output of WP4 and 5 | Improve integrated pipeline based on the success of earlier designs | Existing data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 100 | MB |
| 2 | Optimized integrated pipeline for interface design | Pipeline adjusted according to feedback gained from WP4 and WP5 | New data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 100 | MB |
| 3 | Target list with documentation | Select optimal target for designing both linear- and structural epitope | New data | Derived\_and\_compiled\_data | Research\_documentation | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 1 | MB |
| 4 | Integrated pipelene form WP2.5 | Designing antibody against a structural epitope from the target list | New data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 100 | MB |
| 4 | In silico library for clustering | Starting list on entibody candidates for development of antibody against structural epitope | New data | Canonical\_data | Protein\_sequences | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 5 | GB |
| 4 | 100 selected sequences output | Shortlist of candidate Fabs selected for developability and functional assays | New data | Canonical\_data | Protein\_sequences | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 1 | MB |
| 4 | Pichia pastoris | Organism of expressing Fabs | Existing data | Experimental\_data | Genetically\_modified\_organisms | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 3 | vials of 1 ml |
| 4 | 100 plasmids for antibody protein production | Producing Fabs in Pichia for developability and functional assays | New data | Experimental\_data | Vectors | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 100 | vials of 20 ul |
| 4 | Experimental data on stability, aggregatio propensity, solubility, etc. | Assessing the actual developability and functional properties of Fab candidates | New data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 2 | GB |
| 4 | List of 10 selected Fabs | Shortlist of best-performing Fabs for affinity maturation | New data | Derived\_and\_compiled\_data | Research\_documentation | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 10 | KB |
| 4 | Phage library generated with error-prone PCR of more than a million phages per Fab | Results of affinity maturation of best performing candidate Fabs | New data | Experimental\_data | Genetically\_modified\_organisms | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 10 | vials of 1 ml |
| 4 | Sequencing data of successful phages | Affinity maturation of best performing candidate Fabs | New data | Canonical\_data | Nucleic\_acid\_sequences | Nucleotide and protein sequences: raw sequence data trace (.ab1), textbased format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk); | 5 | MB |
| 4 | Purified Fabs protein | Fabs in the form of purified protein to be used for further analysis of affinity and CMC properties | New data | Experimental\_data | Antibodies | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 10 | vials of 1 mg |
| 4 | Alphascreen binding assay results | Confirming phage display results | New data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 10 | MB |
| 4 | Experimental data on stability, aggregatio propensity, solubility, etc. of affinity-matured Fabs | Assessing the actual developability and functional properties of Fab candidates | New data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 5 | GB |
| 4 | Structures of Fabs | Comparing the predicted and actual structure of Fabs | New data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 10 | GB |
| 4 | Reporter cell lines | Functional testing of Fabs | Existing data | Experimental\_data | Cell\_lines | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 10 | vials of 1 ml |
|  | Results of using reporter cell lines | Functional testing of Fabs | New data | Experimental\_data | Digital\_images | Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; | 5 | GB |
| 4 | Mouse models | Functional testing of Fabs | Existing data | Experimental\_data | Genetically\_modified\_organisms | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 1 | colony |
| 4 | Results of mouse experiments | Functional testing of Fabs | New data | Experimental\_data | Digital\_images | Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; | 10 | GB |
| 5 | Integrated pipelene form WP2.5 | Designing antibody against a structural epitope from the target list | New data | Derived\_and\_compiled\_data | Algorithms\_and\_scripts | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 150 | MB |
| 5 | In silico library for clustering | Starting list on entibody candidates for development of antibody against structural epitope | New data | Canonical\_data | Protein\_sequences | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 6 | GB |
| 5 | 100 selected sequences output | Shortlist of candidate Fabs selected for developability and functional assays | New data | Canonical\_data | Protein\_sequences | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Mark-up Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 1 | MB |
| 5 | Pichia pastoris | Organism of expressing Fabs | Existing data | Experimental\_data | Genetically\_modified\_organisms | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 3 | vials of 1 ml |
| 5 | 100 plasmids for antibody protein production | Producing Fabs in Pichia for developability and functional assays | New data | Experimental\_data | Vectors | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 100 | vials of 20 ul |
| 5 | Experimental data on stability, aggregatio propensity, solubility, etc. | Assessing the actual developability and functional properties of Fab candidates | New data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 3 | GB |
| 5 | List of 10 selected Fabs | Shortlist of best-performing Fabs for affinity maturation | New data | Derived\_and\_compiled\_data | Research\_documentation | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 10 | KB |
| 5 | Phage library generated with error-prone PCR of more than a million phages per Fab | Results of affinity maturation of best performing candidate Fabs | New data | Experimental\_data | Genetically\_modified\_organisms | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 10 | vials of 1 ml |
| 5 | Sequencing data of successful phages | Affinity maturation of best performing candidate Fabs | New data | Canonical\_data | Nucleic\_acid\_sequences | Nucleotide and protein sequences: raw sequence data trace (.ab1), textbased format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk); | 6 | MB |
| 5 | Purified Fabs protein | Fabs in the form of purified protein to be used for further analysis of affinity and CMC properties | New data | Experimental\_data | Antibodies | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 10 | vials of 1 mg |
| 5 | Alphascreen binding assay results | Confirming phage display results | New data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 10 | MB |
| 5 | Experimental data on stability, aggregatio propensity, solubility, etc. of affinity-matured Fabs | Assessing the actual developability and functional properties of Fab candidates | New data | Experimental\_data | Biophysics\_data | Quantitative tabular data: commaseparated value files (.csv), tabdelimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb); | 5 | GB |
| 5 | Structures of Fabs | Comparing the predicted and actual structure of Fabs | New data | Canonical\_data | Protein\_structures | Protein structures: Protein Data Bank format (.pdb / .pdbx); | 12 | GB |
| 5 | Reporter cell lines | Functional testing of Fabs | Existing data | Experimental\_data | Cell\_lines | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 10 | vials of 1 ml |
| 5 | Results of using reporter cell lines | Functional testing of Fabs | New data | Experimental\_data | Digital\_images | Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; | 6 | GB |
| 5 | Mouse models | Functional testing of Fabs | Existing data | Experimental\_data | Genetically\_modified\_organisms | Biological and chemical samples: live animals, frozen samples in cryovials, samples stored at 4°C. | 1 | colony |
| 5 | Results of mouse experiments | Functional testing of Fabs | New data | Experimental\_data | Digital\_images | Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; | 12 | GB |
| 6 | Meeting notes | Records of valorization efforts | New data | Derived\_and\_compiled\_data | Research\_documentation | Text files: Rich Text Format (.rtf), plain text data (Unicode, .txt), MS Word (.doc/.docx), eXtensible Markup Language (.xml), Adobe Portable Document Format (.pdf), LaTex (.tex) format; | 10 | MB |

### 3. LEGAL & ETHICAL ISSUES

**Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.**

* No

**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, add the reference to the formal approval by the relevant ethical review committee(s)**

* Yes

Animal studies will be performed towards the end of the project in WP4 and WP5. We will seek approval of The Ethical Committee for Animal Experimentation (ECD) in due time.

**Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?**

* Yes

We do not exclude that the proposed work could result in research data with potential for tech transfer and valorisation. VIB and KU Leuven has a policy to actively monitor research data for such potential. If there is substantial potential, the invention will be thoroughly assessed, and in a number of cases the invention will be IP protected (mostly patent protection or copyright protection). As such the IP protection does not withhold the research data from being made public. In the case a decision is taken to file a patent application it will be planned so that publications need not be delayed. Further research beyond the scope of this project may be necessary for developing a strong IP portfolio.

**Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?**

* No

### 4. DOCUMENTATION & METADATA

**What documentation will be provided to enable reuse of the data collected/generated in this project?**

Metadata will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the electronic laboratory notebook (E-notebook) and/or in hard copy lab notebooks that refer to specific datasets. All datasets will be accompanied by a README.txt file containing all the associated metadata (see more details below). The data will be generated following standardized protocols. Clear and detailed descriptions of these protocols will be stored in our lab protocol database, and published along with the results.

**Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.**

* The following metadata standards will be used for certain datasets
  + Nucleotide sequence files (vectors and sequencing) : GenBank Sequence Format (<https://fairsharing.org/FAIRsharing.rg2vmt>)
  + Proteomics data: PRoteomics IDEntifications database (PRIDE, <https://www.ebi.ac.uk/pride/>)
* For instrument-specific datasets, additional metadata will be associated with the data file as appropriate.
* For other datasets, the metadata will include the following elements:
  + Title: free text
  + Creator: Last name, first name, organization
  + Date and time reference
  + Subject: Choice of keywords and classifications
  + Description: Text explaining the content of the data set and other contextual information needed for the correct interpretation of the data, the software(s) (including version number) used to produce and to read the data, the purpose of the experiment, etc.
  + Format: Details of the file format,
  + Resource Type: data set, image, audio, etc.
  + Identifier: DOI (when applicable)
  + Access rights: closed access, embargoed access, restricted access, open access.

The final dataset will be accompanied by a README.txt document. This file will be located in the top-level directory of the dataset and will also list the contents of the other files and outline the file-naming convention used. This will allow the data to be understood by other members of the laboratory and add contextual value to the dataset for future reuse.

### 5. DATA STORAGE & BACK UP DURING THE FWO PROJECT

**Where will the data be stored?**

Digital files will be stored either on KU Leuven servers or in shared laboratory folders of an off-site online backup service. The researchers working on the project will have copies of the data files as well as of the derived and compiled data stored on their personal computers.

The Switch Lab has a professional subscription to an off-site online backup service with unlimited space, version control and roll-back capability, which will be used for storage during the project and after. There is a secondary on-campus physical backup of the online storage which synchronizes with the online content with a one-day delay.

Algorithms, scripts and softwares: All the relevant algorithms, scripts and software code driving the project will be stored in a private online git repository from the GitHub account of the department (https://github.com/vibcbd).

The screening core has a database system in place to handle the data stream from the high content imaging screen, including archiving facilities and will store the data during the project. Representative images and the quantitation of the images will be transferred to the Switch laboratory storage for long term storage.

Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacteria glycerol stock (-80°C). All published vectors and the associated sequences will be sent to the non-profit plasmid repository Addgene, which will take care of vector storage and shipping upon request.

Cell lines: Newly created cell lines will be stored locally in the laboratory in liquid nitrogen storage and will be deposited in the UZ Leuven-KU Leuven Biobank.

Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate.

**How is back up of the data provided?**

The Switch Lab has a professional subscription to an off-site online backup service with unlimited space, version control and roll-back capability, which will be used for storage during the project and after. There is a secondary on-campus physical backup of the online storage which synchronizes with the online content with a one-day delay.

**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

The Switch Lab has a professional subscription to an off-site online backup service with unlimited space, which will be used for storage during the project and after.

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

Data storage and backup costs are included in general lab costs. The Switch Lab has a yearly subscription to an off-site online backup service paid from the general budget of the laboratory. The yearly cost of the service is 5500 Euros. This cost includes unlimited data storage, not only the data belonging to the present project.

Electricity costs for the -80° and -20° freezers and refrigerators present in the labs as well as the cost of liquid nitrogen cryostorage are included in general lab costs.

**Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?**

All notebooks and physical data are stored in the labs. Entry to the lab requires ID-card and key. Access to the digital data is u-number and password controlled.

### 6. DATA PRESERVATION AFTER THE FWO PROJECT

**Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).**

The minimum preservation term of 5 years after the end of the project will be applied to all datasets.

**Where will the data be archived (= stored for the longer term)?**

The goal of this project is to launch FoldCo, a protein design company in Flanders. The data generated by the project will form the portfolio of FoldCo therefore it will be proprietary and not shared widely.

For the datasets that will be made openly accessible, we will use, whenever possible, the existing platforms that support FAIR data sharing (www.fairsharing.org), at the latest at the time of publication.

For all other datasets, long term storage will be ensured as follows: -Digital datasets will be stored on storage space of an online data-backup service. -Vectors: As a general rule at least two independently obtained clones will be preserved for each vector, both under the form of purified DNA (in -20°C freezer) and as a bacterial glycerol stock (-80°C). -Other biological and chemical samples: storage at 4°C and/or as frozen samples in cryovials as appropriate.

**What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?**

Electricity costs for the -80° and -20° freezers and refrigerators present in the labs as well as for in liquid nitrogen cryostorage are included in general lab costs. The cost of the laboratory's professional subscription to the online data backup service is 5500 Euros per year (27 500 Euros for 5 years). This cost includes unlimited data storage, not only the data belonging to the present project. Data storage and backup costs are included in general lab costs.

### 7. DATA SHARING AND REUSE

**Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?**

* No

**Which data will be made available after the end of the project?**

Participants to the present project are committed to publish research results to communicate them to peers and to a wide audience. All research outputs supporting publications will be made openly accessible. Depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or publicly via repositories (e.g. negative data). We aim at communicating our results in top journals that require full disclosure upon publication of all included data, either in the main text, in supplementary material or in a data repository if requested by the journal and following deposit advice given by the journal. Depending on the journal, accessibility restrictions may apply. Physical data (e.g. cell lines) will be distributed to other parties if requested.

**Where/how will the data be made available for reuse?**

* The data will be shared upon request by mail.
* Possible ways of sharing the generated data:
  1. nucleic acid sequences: GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)
  2. protein sequences: UniProt KB (<https://www.uniprot.org/>)
  3. vectors: AddGene (<http://www.addgene.org/depositing/start-deposit/>)
  4. cell lines: direct mailing on dry ice
  5. microscope images: Image Data Resource (<http://idr.openmicroscopy.org/about/>)
  6. proteomics data: PRIDE (<https://www.ebi.ac.uk/pride/>)
  7. manuscripts: bioRxiv (<https://www.biorxiv.org/>)
  8. other digital data: Zenodo data repository (<https://zenodo.org/>)

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

* Upon publication of the research results

Generally, research outputs will be made openly accessible at the latest at the time of publication. No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements – note that patent application filing will be planned so that publications need not be delayed - or ongoing projects requiring confidential data. In those cases, datasets will be made publicly available as soon as the embargo date is reached.

**Who will be able to access the data and under what conditions?**

Whenever possible, datasets and the appropriate metadata will be made publicly available through repositories that support FAIR data sharing. As detailed above, metadata will contain sufficient information to support data interpretation and reuse and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, a Creative Commons Attribution (CC-BY)  or an ODC Public Domain Dedication and Licence, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. For data shared directly by the PI, a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.

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

It is the intention to minimize data management costs by implementing standard procedures e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget.

The receiving party will pay for sharing physical data (e.g. cell lines).

### 8. RESPONSIBILITIES

**Who will be responsible for data documentation & metadata?**

Metadata will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the electronic laboratory notebook (E-notebook) that refer to specific datasets.

**Who will be responsible for data storage & back up during the project?**

The research and technical staff will ensure data storage and back up, with support from René Custers and Alexander Botzki for the electronic laboratory notebook (ELN) and from Raf De Coster for the KU Leuven drives.

**Who will be responsible for ensuring data preservation and reuse ?**

The PI is responsible for data preservation and sharing, with support from the research and technical staff involved in the project, from René Custers and Alexander Botzki for the electronic laboratory notebook (ELN) and from Raf De Coster for the KU Leuven drives.

**Who bears the end responsibility for updating & implementing this DMP?**

The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.