# DMP title

**Project Name** FWO DMP Jose Ignacio Alvira - DMP title

**Grant Title** 1SF1522N

**Principal Investigator / Researcher** Jose Ignacio Alvira Larizgoitia

**Description** Our objectives are: integrate several data modalities measured in cells to create a common representation that better separates cell types and states, create a meaningful representation of cellular microenvironments and see how these affect cell types and improve current methods for the classification of whole-slide images by using the information captured in spatial transcriptomics.

**Institution** KU Leuven

## 1. General Information

**Name applicant**

Jose Ignacio Alvira Larizgoitia

**FWO Project Number & Title**

Development of multimodal deep neural networks for spatial multi-omics data fusion
1SF1522N

**Affiliation**

- KU Leuven

## 2. Data description

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

- Reuse existing data

**Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).**

| Type of data | Origin | File format | Estimated volume |
|---|---|---|---|
| Tabular data (single cell measurements) | Public databases and ongoing collaborations | .csv, ,xls, .Rds | 20-80 GBs |
| Nucleic acid sequence data | scRNAseq, scATACseq, scCITEseq, scNMTseq, Visium, etc. experiments of current collaborators | .fastq, .Rds and similar | 10-50 GBs |
| Code (text data) | Written from scratch for this project | .R, .py, .m, .sh, and similar | 1 GB |
| Computational models | Created and trained for this project | .h5 or similar | 10-200GBs |
| Microscopy data | Experiments of current collaborators | .tiff, .jpg, .png or similar | 100 GBs - 15 TB |
| Spatial transcriptomics | Visium, MERFISH. Experiments done by current collaborators | .tiff, .Rds or similar | 10 - 100 TBs |

## 3. Legal and ethical issues

**Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). 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)**

- No

**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?**

- No

Not at this moment. In case there rise IP possibilities, we will contact LRD.

**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 and metadata

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

1. The new metadata created in single-cell analyses will be added to the relevant object so that it stores all information and previous metadata (for example identiffied cell-types in a scRNAseq experiment).

2. The code used for this research will be documented in between the code and in GitHub so that its reuse or implementation is easiest. Besides, jupyter notebooks will be uploaded to make the working pipeline easier to follow.

3. A guide on how to read in and use trained models will also be available on GitHub for each specific model.

4. Metadata will be added about the processing of microscopy images.

**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.**

- Yes
- No

The datatypes used in this project are very different from one another so no metadata standard will be used. However, standard practices in metadata documentation will be used in each object separately to make it possible to understand and reuse every dataset used individually (for example, adding metadata to Seurat objects for single-cell analyses or documenting each function/class created in the code).

## 5. Data storage and backup during the FWO project

**Where will the data be stored?**

All of the data will be stored in local computers, hard drives and the VSC (Vlaamse Supercomputer Centrum). Besides, online servers will be used for code and notebooks (GitHub) and omics data (KU Leuven servers and KU Leuven OneDrive). In the case we need more space for very large-scale image data, we will explore cloud-based arctic cold storage solutions as well.

**How is backup of the data provided?**

The university servers, the VSC and GitHub are automatically backed-up.

**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

Sufficient back-up and storage capacity is already provided by the VSC and KU Leuven servers.

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

The expected costs are 20€ per TB per year. These costs will be covered by internal KU Leuven funds, VLIR infraestructure grant for compute and storage and project-related funding.

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

Access to the data is only granted to authorized researchers.

## 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, ...).**

All data. However spatial multiomics experiments comprise of large-scale data sets (10 TB/sample). In case the costs of storage
become prohibitive we will reevaluate the cost of long-term storage of raw imaging data compared to preprocessed/compressed data types.

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

The data will be archived in the VSC, KU Leuven servers, GitHub and research group hard drives.

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

The expected costs are 20€ per TB per year. The costs will be covered by internal KU Leuven funds, VLIR infraestructure grant for compute and storage and project-related funding.

The clinical data will be kept for 10 years and whis will be covered by VLIR infraestructure grant and Leuven Institute for Single Cell Omics (LISCO)

## 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?**

After the research, all data will be made openly accessible, whenever possible.

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

- In an Open Access repository
- In a restricted access repository
- Upon request by mail

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

- Upon publication of the research results

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

Not yet applicable

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

No costs are expected for data sharing.

## 8. Responsibilities

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

I, Jose Ignacio Alvira, as the grant holder, will be responsible for data documentation and addition

of new metadata.

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

I, Jose Ignacio Alvira, as the grant holder, will be responsible for data storage & back up during the project.

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

The PI Alejandro Sifrim

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

The PI bears the end responsibility of updating & implementing this DMP.