Uncovering the implications of N-glycosylation and protein aggregation in human disease

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	
Mass spectrometry data	Proteomic data identifying aggregated proteins in different N-linked CDGs.	N	D	Numerical	mzML, CSV	<5TB	NA
Transcriptomics data	RNA-Seq data to analyze ER stress and cellular responses in CDG-affected cells.	N	D	Numerical	FASTQ, BAM, CSV	<1TB	NA
Patient derived fibroblasts	Fibroblast cells obtained from patients with N-linked CDGs for proteomics and aggregation studies.	E	P	Organism	NA	NA	<1m³
Healthy fibroblasts	Control fibroblasts	N	P	Organism	NA	NA	<1m³
Western Blot data	Images of protein expression and aggregation analysis using Western blotting.	N	D	Images, Numerical	PNG, TIFF, CSV	<10 GB	NA
Cell viability assays data	Results from viability assays measuring the impact of aggregation on cellular health.	N	D	Numerical	CSV	<1 GB	NA
Amyloid light chain sequences	Amyloid Light Chain sequences from a public repository (AL-Base)	E	D	Textual	CSV	<1 GB	NA
Purified Light Chains	Purified κ light chains from cells.	N	D	Protein samples	NA	NA	NA
Transmission electron microscopy (TEM)	Transmission electron microscopy images of amyloid fibril morphology	N	D	Images	TIFF, PNG	<100 GB	NA
Biophysical characterization data	Biophysical data from stability assays and high throughput protein aggregation assays	N	D	Numerical	CSV	<100 GB	NA
HEK 293 T cells	Human embryonic kidney cells (HEK293T) used for protein expression and aggregation studies.	N	P	Organism	NA	NA	<1m³
In cellulo aggregation data	Confocal imaging data on κ LC aggregation patterns in live cells.	N	D	Images, numerical	TIFF, CSV	<100GB	NA
Experimental protocols	Experimental protocols including MS proteomics, etc.	N	D	Textual	TXT	<1 GB	NA

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:

- **Amyloid Light Chain Sequences:**
  - Source: Amyloid Light Chain Database (AL-Base)
  - DOI: [10.1080/13506120802676781](10.1080/13506120802676781)
- **Human derived fibroblasts:**
-
  - Provided by Prof. Gert Matthijs from the Department of Human Genetics

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.

- Yes, animal data (Provide ECD reference number below)

The project involves the use of patient-derived fibroblasts provided by Prof. Gert Matthijs from the Department of Human Genetics. These fibroblasts are pre-existing materials and will be reused for this research. Ethical approval for the collection and use of these cells is covered under the EC S-number: **S67875**. All samples will be anonymized in compliance with GDPR and institutional ethical guidelines.

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.

- No

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.

- Yes

The use of patient-derived fibroblasts is subject to restrictions. We do not have a data transfer agreement to use these cells in our lab. As a result, all experiments involving the patient-derived fibroblasts will be conducted in the laboratory of Prof. Gert Matthijs at the Department of Human Genetics. Data generated from these experiments will adhere to the ethical and dissemination policies of that lab.

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

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).

We will ensure all datasets are easy to understand and use by:
- **Clear Documentation**:

- Use README.txt files to explain dataset contents, methods, variables, and units.
- Create codebooks for numerical data to define variables and explain processing steps.
- Record experimental steps and protocols in **LabCollector**, the lab's Electronic Lab Notebook (ELN).
- Include links in the ELN to connected datasets stored on file servers for easy access.

- **Organized Metadata**:

- Mass spectrometry will follow mzML standards.
- Transcriptomics data will use MIAME standards.
- Imaging data (e.g., TEM) will include metadata for imaging settings and conditions.

- **File Organization and Storage**:

- Data will be stored in clearly named folders with subfolders for documentation.
- **Dropbox (Professional Subscription)** will serve as the main storage for datasets up to medium size.
- Larger datasets will be stored on **Mango**, KU Leuven's file server.

- **Sharing and Long-term Use**:

- Assign persistent identifiers (like DOIs) to datasets.
- Include full documentation when archiving data.

Examples:

- **Mass Spectrometry Data**: mzML/CSV files with README files describing software and processing workflows.
- **Western Blots**: PNG/TIFF files with experiment notes, including sample IDs and antibody sources.

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

1. Metadata Standards:

- **DataCite** will be used for descriptive and citation metadata to ensure datasets are findable and citable. This includes authors, title, abstract, keywords, and identifiers like DOIs.
- Community-specific metadata standards will be applied where relevant:
    - **HUPO-PSI mzML** for mass spectrometry data.
    - **MIAME** for transcriptomics data.
    - **OME-TIFF** for imaging datasets (e.g., TEM and confocal microscopy).

2. Structural Metadata:

- Metadata will outline dataset contents, including units of analysis, sample types (e.g., fibroblasts, HEK 293T cells) and instruments and settings (e.g., mass spectrometer configurations, imaging parameters).
- These will be included as part of the dataset's documentation and stored alongside the data.

3. Metadata Creation and Management:

- Instruments (e.g., mass spectrometers, confocal microscopes) will generate metadata during data collection.
- Additional metadata will be created manually and managed using tools like LabCollector and the KU Leuven Research Data Repository.

Data Storage & Back-up during the Research Project

Where will the data be stored?

- Other (specify below)
- ManGO

The data will be stored in the following locations:
1. **ManGO (KU Leuven File Server)**:
·   Used for large datasets, such as mass spectrometry, transcriptomics, and imaging data, that exceed the storage capacity of Dropbox.
2. **Dropbox (Professional Subscription)**:
·   Used for small to medium-sized datasets and project documentation.
·   Selected for its ease of use, professional-grade storage capabilities, and collaboration features. Links to these datasets will be maintained in LabCollector for accessibility.
Both storage solutions are secure, regularly backed up, and align with the project's requirements. No external hard drives or free cloud storage will be used.

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution
- Other (specify below)

1. **KU Leuven ICTS Standard Backup**:

- Data stored on **ManGO** will be backed up automatically through the standard backup services provided by KU Leuven ICTS.

2. **Dropbox Professional Backup**:

- Dropbox Professional includes:
    - Continuous data synchronization and versioning.
    - A 180-day recovery window for deleted files or previous versions.
    - AES 256-bit encryption for stored data and SSL/TLS encryption during data transfer.

3. **Additional Measures**:

- Links to datasets stored on Dropbox will be recorded in LabCollector to ensure data traceability.
- For critical datasets, manual periodic checks and redundant storage strategies will be implemented as an extra precaution.

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

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

We ensure the security of our data through the following measures:

1. **Network Security**:

    - Data stored on ManGO (KU Leuven) benefits from the university's secure infrastructure, which includes access controls and regular monitoring by KU Leuven ICTS.
    - Dropbox Professional uses industry-standard AES 256-bit encryption for data at rest and SSL/TLS for data in transit.

1. **Access Control**:

    - Access to KU Leuven storage drives (ManGO) and Dropbox is limited to authorized users only.
    - Permissions are assigned based on project roles and are regularly reviewed to ensure they remain appropriate.

1. **Secure Data Transfer**:

    - Data transfers are conducted using encrypted connections (e.g., SFTP or Dropbox's secure protocols).
    - Sensitive or large datasets are shared using ManGO's secure file-sharing features or approved KU Leuven platforms.

1. **Physical Security**:

    - The lab is located in Building ON1 bis on the GHB campus, where access to lab spaces is controlled by badge access.

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

1. **ManGO (KU Leuven Large File Repository)**:
    - **Cost**: ManGO incurs a fee of €95.14 per TB per year for storage. Based on the estimated data volume, the lab anticipates storing up to 5 TB, resulting in an annual cost of approximately €475,70.
    - **Coverage**: This cost will be covered by the research project's budget, allocated under data management expenses.
2. **Dropbox Professional**:
    - **Cost**: The annual subscription fee for Dropbox Professional is approximately €200 per user.
    - **Coverage**: This is also covered by the project's operational budget.
3. **Backup Costs**:
    - Both storage solutions include integrated backup services (KU Leuven ICTS for ManGO and Dropbox's built-in recovery and versioning features), so no additional backup costs are expected.

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 data will be archived using:

- **ManGO (KU Leuven Large Volume Storage)**: For large datasets requiring secure, long-term storage.
- **Dropbox Professional**: For smaller datasets and project documentation, with links maintained in the ELN for traceability.

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

ManGO (KU Leuven Large Volume Storage):

- **Cost**: €95,14 per TB per year. For an estimated 5 TB of storage, the annual cost is approximately **€475,70**. Over a 10-year retention period, the total cost is **€4757**.

Dropbox Professional:

- **Cost**: €200 per year. Over 10 years, the total cost is **€2000**.

Total Costs:

- - Combined, the expected cost for 10 years of data preservation is approximately **€6757**.

Coverage: These costs will be covered under the lab's general operational expenses.

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 restricted data (upon approval, or institutional access only)
- Yes, as open data

Mass Spectrometry and Transcriptomics Data:

- **Availability**: These datasets will be made available as **open data** in public repositories (e.g., PRIDE for MS data and GEO for transcriptomics data) after publication.
- **Conditions**: Accompanied by proper documentation and persistent identifiers (DOIs).

Patient-derived Fibroblasts and Related Data:

- **Availability**: Restricted to researchers under **Material Transfer Agreements (MTA)** to ensure compliance with ethical and legal requirements.

Other Datasets (e.g., Western Blots, Imaging, Protocols):

- **Availability**: These datasets will be available as **restricted data** upon approval to ensure proper use.

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

Access to the data will be restricted to ensure compliance with ethical and institutional requirements. Researchers or institutions with a valid research purpose may request access, but approval will be granted only after careful evaluation by the principal investigator (PI) and, where necessary, relevant ethical committees. All data sharing will be governed by a Material Transfer Agreement (MTA) or a similar data sharing agreement. These agreements outline the conditions for use, including proper citation of the data source, restrictions on further sharing, and adherence to ethical guidelines.

Requests for access must be submitted in writing, specifying the intended use of the data. Approved users will receive the data through secure transfer methods, such as encrypted file sharing provided by KU Leuven systems. Compliance with the terms of the agreement will be monitored, and access may be revoked if the data is used improperly. This controlled approach ensures that the data remains protected while supporting its responsible use for research.

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
- Yes, ethical aspects
- Yes, privacy aspects

The following factors restrict or prevent sharing for certain datasets:

1. Control and Patient-derived Human Fibroblasts:
Restrictions: Ethical and legal aspects. Sharing requires Material Transfer Agreements (MTAs) with the provider and compliance with GDPR.

2. HEK 293T Cells:
Restrictions: Legal and intellectual property (IP) aspects. These cells are obtained from a commercial source and sharing is subject to provider agreements and IP restrictions.

3. Other Datasets (e.g., Mass Spectrometry, Transcriptomics):
No restrictions: These datasets can be shared after publication in public repositories, following community standards and

repository guidelines

Where will the data be made available?

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

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

1. **Mass Spectrometry Data**:
Repository: **PRIDE** (Proteomics Identifications Database) for proteomics data.
2. **Transcriptomics Data**:
Repository: **GEO** (Gene Expression Omnibus) for RNA-Seq datasets.
3. **Other Datasets (e.g., Imaging, Western Blots)**:
Repository: **KU Leuven RDR** for sharing with access restrictions or general data not covered by disciplinary repositories.
4. **Physical Data (e.g., cell lines)**:

- Availability: Shared under **Material Transfer Agreements (MTA)** only, not via repositories.

When will the data be made available?

- Upon publication of research results

Which data usage licenses are you going to provide?

If none, please explain why.

- Data Transfer Agreement (restricted data)

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

- **Mass Spectrometry Data**: An accession number will be assigned upon deposit in **PRIDE**.
- **Transcriptomics Data**: An accession number will be assigned upon deposit in **GEO**.
- **Other Datasets**: For datasets deposited in **KU Leuven RDR**, a DOI will be generated to ensure long-term identification and retrieval.

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

1. **Mass Spectrometry and Transcriptomics Data**: Data sharing through **PRIDE** and **GEO** repositories is typically free of charge, as these are publicly funded repositories.
2. **KU Leuven RDR**: Sharing via the KU Leuven Research Data Repository does not incur additional costs for researchers affiliated with KU Leuven.

Hence, no significant costs are anticipated for data sharing.

Responsibilities

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

Ramon Duran, the fellowship holder, will manage the documentation and metadata for all datasets generated during the project. He will be supported by **Joost Schymkowitz** and **Frederic Rousseau** (Switch Lab PIs), with additional guidance from **Gert Matthijs** (co-promotor), particularly for datasets involving patient-derived fibroblasts.

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

- Data storage and backup will be managed by **Ramon Duran**, who will ensure proper organization and adherence to institutional policies.
- Storage solutions (ManGO and Dropbox Professional) are supported by KU Leuven ICTS and the lab's infrastructure.
- Oversight and compliance will be ensured by **Joost Schymkowitz** and **Frederic Rousseau**.

Who will manage data preservation and sharing?

Joost Schymkowitz and **Frederic Rousseau** will oversee long-term data preservation and ensure compliance with KU Leuven and repository policies.
Ramon Duran will prepare datasets for sharing during the project, including depositing in public repositories (e.g., PRIDE, GEO, KU Leuven RDR).

Who will update and implement this DMP?

- **Ramon Duran** will handle updates and ensure that the DMP is implemented throughout the project.
- **Joost Schymkowitz**, **Frederic Rousseau**, and **Gert Matthijs** will provide oversight and approve any major revisions to the DMP.