FWO DMP Template - Flemish Standard Data Management Plan

Version KU Leuven

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	**Neesha Rajesh Shewakramani, ORCID: 0000-0002-4577-3491**
Contributor name(s) (+ ORCID) & roles	**Prof. Kalyan Das, ORCID: 0000-0002-8897-324X - Promoter** **Dr. Julien Olivet, ORCID: 0000-0003-4115-5546 – Co-promoter** **Prof. Michael Washburn, University of Kansas, USA - Professor of collaborating group** **Prof. Marc Vidal, Harvard Medical School, USA – Professor of collaborating group** **Prof. Gregory Bowman, Johns Hopkins University, USA, ORCID: 0000-0001-8025-4315 – Professor of collaborating group** **Dr. Ilana Nodelman – Research scientist from Bowman group** **Joseph Cesare, ORCID: 0000-0002-5569-1422 - PhD student from Washburn group**
Project number [1] & title	**3M230090** **Structural study and integrative modelling of the yeast Sin3/Rpd3 histone deacetylase complexes**
Funder(s) GrantID [2]	**11PN624N**
Affiliation(s)	☒KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: ROR identifier KU Leuven: 05f950310

[1] "Project number" refers to the institutional project number. This question is optional. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

Please provide a short project description	Histone deacetylase (HDAC) enzymes are the main transcriptional repressors in eukaryotes. Most of these enzymes do not function in isolation but as part of multi-subunit complexes. Among HDAC machineries, the evolutionarily conserved Sin3 HDAC complex is involved in various cellular processes and diseases. However, the molecular assemblies of the Sin3 HDAC complexes remain poorly understood with limited knowledge on how the 3D organization of their subunits can impact diverse functions. Here, to address this question, we use single-particle cryogenic-electron microscopy (cryo-EM) and integrative modelling approaches to determine high-resolution 3D structures of the conserved Sin3/Rpd3 HDAC complexes from yeast *Saccharomyces cerevisiae*. These structures and subunit interaction mapping information will serve as a foundation for understanding the molecular mechanisms behind HDAC complex-mediated transcriptional regulation in eukaryotes, and how these epigenetic molecular machines can be reprogrammed for therapeutic interventions.

2. 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 [3].

				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 (GB, TB)	Physical Volume
Sequences	Sequences of plasmids and primers for protein purification	☒ Generate new data ☒ Reuse existing data	☒ Digital	☒ Textual	.fasta .txt .dna	☒ < 1 GB	
Biophysical characterization results	Results from biophysical characterization of purified proteins using techniques such as NanoDrop, SDS-PAGE, Native PAGE, DLS	☒ Generate new data	☒ Digital	☒ Images ☒ Textual ☒ Numerical	.tiff .png .jpg / .jpeg .dexp .pdf .txt .csv .xlsx	☒ < 1 GB	
Biochemical assay results	Results from enzymatic activity assays of purified proteins	☒ Generate new data	☒ Digital	☒ Images ☒ Numerical	.tiff .png .pzfx .raw .csv .xlsx .pdf	☒ < 1 GB	
Phenotypic assay results	Results from phenotypic assays of strains used protein production	☒ Generate new data	☒ Digital	☒ Images	.tiff .png .jpg / .jpeg .pdf	☒ < 1 GB	

[3] Add rows for each dataset you want to describe.

Physical samples from protein purification	Plasmids, ordered primers, cell strains, cell pellets, purified protein complex samples	☒ Generate new data	☒ Physical	NA	NA	NA	2-3 boxes of Eppendorf tubes
Cryo-EM data	Raw data (movies) generated during collection of data from cryo-EM of purified protein (complex) samples	☒ Generate new data	☒ Digital	☒ Audiovisual (video)	.mrc .eer .gain .dm .jpg / .jpeg .xml .tiff	☒ < 5 TB (per dataset)	
Protein density maps	3D electron density maps generated from raw cryo-EM data using software such as Relion and cryoSPARC	☒ Generate new data	☒ Digital	☒ Audiovisual ☒ Numerical ☒ Model	.mrc .star .cs .txt .map	☒ < 100 GB	
Protein structures	Structures modelled from electron density maps using software like Phenix, Coot and CCP4	☒ Generate new data	☒ Digital	☒ Audiovisual ☒ Numerical ☒ Model	.pdb .cif .txt .geo .eff .smiles .smi	☒ < 10 GB	
Images and movies of protein maps and structures	Images* and movies* of electron density maps and solved structures generated using Chimera, ChimeraX and PyMol	☒ Generate new data	☒ Digital	☒ Audiovisual ☒ Images	.txt .cxs .py .pyc .tiff .png .jpg / .jpeg	☒ < 10 GB	

	*made with (un)published maps and structures				.mp4		
Documentation	Experimental protocols (adjusted from literature), scripts for generating movies, scripts for data analysis	☒ Generate new data ☒ Reuse existing data	☒ Digital ☒ Physical	☒ Textual	.txt .docx .pdf .R .py	☒ < 10 GB	Notebooks and printed protocols

GUIDANCE:
The data description forms the basis of your entire DMP, so make sure it is detailed and complete. It includes digital and physical data and encompasses the whole spectrum ranging from raw data to processed and analysed data including analysis scripts and code. Physical data are all materials that need proper management because they are valuable, difficult to replace and/or ethical issues are associated. Materials that are not considered data in an RDM context include your own manuscripts, theses and presentations; documentation is an integral part of your datasets and should described under documentation/metadata.
RDM Guidance on data

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.	Plasmid sequences from literature are available upon request from the respective authors, if not publicly available in the supplementary information of the paper. Experimental protocols are publicly available in the literature. Published protein structures used for comparison and analysis of generated structures along with their electron density maps are publicly available in the Protein Data Bank (PDB) and Electron Microscope Data Bank (EMDB) respectively.
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, human subject data; provide SMEC or EC approval number: ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☒ No Additional information:

Will you process personal data[4]? 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).	☐ Yes (provide PRET G-number or EC S-number below) ☒ No Additional information:
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 ☒ No If yes, please comment:
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 to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:
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 to what data they relate and which restrictions will be asserted.	☐ Yes ☒ No If yes, please explain:
3. Documentation and Metadata	

[4] See Glossary Flemish Standard Data Management Plan

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). *RDM guidance on documentation and metadata*.	**Sequences:** Reused plasmid and primer sequences will be named as per the original naming convention. Sequences of generated plasmids will be named in the following format – YYYYMMDD_originalPlasmidName_modification. Sequences of primers generated during the project will be named as NRSxxx_plasmidName_FW or NRSxxx_plasmidName_RV in a self-explaining manner. All information will be listed in an Excel file including the location of the file and the physical sample. All files will be stored in appropriate folders in my KU Leuven OneDrive. **Digital data from biophysical characterization:** Experimental data obtained from biophysical characterization of purified protein/protein complexes will be stored on the shared J: drive in the appropriate folders. In the shared J: drive, there are folders for results of each type of biophysical characterization with sub-folders for the results of each member of the laboratory. These files are named in the following convention: YYYYMMDD_purificationNo._strain_purificationStep with the date describing the date of the characterization experiment. The purificationNo. denotes the number of purification from the complete list of purifications performed. The strain denotes the strain from which protein(s) were purified and the purificationStep clarifies the stage of the purification protocol the characterization was performed at. All files are stored in an hierarchical order. **Digital data from biochemical and phenotypic assays:** Experimental data obtained from biochemical and phenotypic assays will be stored in the KU Leuven OneDrive, named in the following format YYYYMMDD_self-explainingName. The name describes the date of the experiment, along with the sample/strain used in the experiment and the type of assay performed. All files are stored in an hierarchical order in appropriate folders based on the type of assay. **Physical samples from protein purification:** Physical samples of purified protein/protein complexes are stored in a box labelled with my name in the -80°C named with the date of protein (complex) production in YYYYMMDD format along with the protein(s') name(s), buffer components, concentration, etc. A complete list of all samples will be maintained. The location of the samples will also be available for all laboratory members to search on using the laboratory inventory maintenance platform, Quartzy. Glycerol stocks of cell strains will be named with the strain name and date of glycerol stock creation and saved in the -80°C freezer as well. A list of all cell strains will be maintained with locations as well. Plasmids and primers will be saved in the same naming convention as their sequences and stored in the -20°C freezer. **Cryo-EM data:** Each cryo-EM data collection generates a dataset with metadata generated automatically. This metadata includes information on the date and time of data collection, the user, the equipment parameters, etc. Each dataset is saved on an in-line server and is backed up on an external 600 TB RAID

	server in the laboratory. Each data collection is named with the date in YYYYMMDD format with a self-explaining name explaining the name and most important parameters of the data collection, for example: YYYYMMDD_proteinName_sampleConcentration_gridType. All raw data and metadata will stored on the server and backed up for several years with the end responsibility of data preservation with Prof. Kalyan Das. A master list of all datasets is maintained as well. **Protein electron density maps and structures:** Protein density maps and solved structures are stored on the shared drive in the appropriate folders as described above. Density maps are also backed up on the external 600 TB large capacity storage along with the raw cryo-EM dataset. If generated by cryoSPARC, the density maps are additionally stored on the user's account on the VSC. Density maps and protein structures are named in a similar naming convention as the raw cryo-EM data (YYYYMMDD_proteinName_description). **Images and movies of protein maps/structures:** Images and movies will be stored on the KU Leuven OneDrive named in a similar naming convention as the protein structure with additional descriptions such as the orientation of the structure, etc. **Documentation:** All experiments and results will be recorded in physical lab notebooks. Experimental protocols will also be stored as Word or Excel files. A summary of the results and analyses performed will be stored in Word or PowerPoint files as well. During the project, data documentation, storage and backup will be my responsibility. In summary, all data will be stored in appropriate folders on the KU Leuven OneDrive or shared J: drive except for large volume data which will be stored on the in-line server and backed up on the external 600 TB drive. All data will be named with the date in YYYYMMDD and a clear description of the sample. All physical samples will be stored in the appropriate freezer and lists of all locations will be maintained.

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. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☒ Yes ☐ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: All cryo-EM datasets will be backed up on the 600 TB RAID configured data storage server and named in the following convention: YYYYMMDD_sampleName_sampleDetails_cryoEMParameters. If no, please specify (where appropriate per dataset or data type) which metadata will be created:

4. Data Storage & Back-up during the Research Project	
Where will the data be stored? *Consult the interactive KU Leuven storage guide to find the most suitable storage solution for your data.*	☒ Shared network drive (J-drive) ☐ Personal network drive (I-drive) ☒ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☐ Large Volume Storage ☐ Digital Vault ☒ Other: **Digital data** generated during the project (including, but not limited to, protocols, raw data, analysed data, images/movies, density maps and protein structures) will be stored on the shared network (J-drive) which restricts access to the members of the group. In addition to this, digital data is also stored on KU Leuven OneDrive, which can only be accessed by me. These drives are centrally managed by KU Leuven ICTS and have a built-in back-up capacity guaranteeing long-term availability of the data as well as recovery, if needed. **Cryo-EM data** will be saved on an in-line server (Athena) and will be backed-up on external 600TB server and a new 1.2PB server available in our lab. Notebooks and other **hard copy material** (including printed protocols) will be stored physically in the laboratory.

How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS?*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution. ☒ Personal back-ups I make (specify) ☐ Other (specify) All **digital data** (except cryo-EM data) is automatically backed-up when stored on drives managed by KU Leuven, including the J-drive or KU Leuven OneDrive. On the faculty network J-drives, "snapshot" technology is used to actively back-up files are different timestamps. A mirror copy of all the data is stored in a second data center as well, to prevent data loss. This occurs every hour, which allows restoration of data to within the hour in case the primary data center is compromised. Large-volume **cryo-EM data** is backed-up on external 600TB capacity and 1.2PB RAID servers, both of which are available in our lab.
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 ☐ No If yes, specify concisely: The KU Leuven OneDrive provided to each staff member has a capacity of 2 TB, which is sufficient for the storage of all digital data except for cryo-EM data. For cryo-EM data, the 600TB and 1.2PB external servers have the capacity to store and backup the data as needed. Additionally, for the storage of cryo-EM data generated during the processing steps, an extra 60TB is available on the Vlaams Supercomputer Centrum (VSC) HPC.
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *CLEARLY DESCRIBE THE MEASURES (IN TERMS OF PHYSICAL SECURITY, NETWORK SECURITY, AND SECURITY OF COMPUTER SYSTEMS AND FILES) THAT WILL BE TAKEN TO ENSURE THAT STORED AND TRANSFERRED DATA ARE SAFE.* *Guidance on security for research data*	The external drives can only be accessed by the computers in the lab, which require a KU Leuven account and password to log in. Therefore, unauthorized people cannot log in to the computers and access the external drives. The shared J-drive can also only be accessed by members of the lab using their lab PCs and their KU Leuven ID and passwords. Furthermore, at all these stages, two-factor authentication is needed during log in.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The external 600TB and 1.2PB data storage servers are already available in our lab and hence no costs are incurred with their use. The use of the shared J-drive amounts to €51.90/100GB/year, which is covered by the laboratory. Additionally, data stored on the HPC costs €20/TB/year, which with the current capacity of 60TB costs €1200 which is covered by the host institution and the laboratory.

5. 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...). *Guidance on data preservation*	☒ All data will be preserved for 10 years according to KU Leuven RDM policy ☐ All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans ☐ Certain data cannot be kept for 10 years (explain)
Where will these data be archived (stored and curated for the long-term)? *Dedicated data repositories are often the best place to preserve your data. Data not suitable for preservation in a repository can be stored using a KU Leuven storage solution, consult the interactive KU Leuven storage guide.*	☐ KU Leuven RDR ☐ Large Volume Storage (longterm for large volumes) ☐ Shared network drive (J-drive) ☒ Other (specifiy): Digital data (except cryo-EM data) will be stored on the KU Leuven K: drive which is specifically meant for archive storage and has an unlimited maximum size. This drive can only be accessed by the researchers involved in the project and the data cannot be modified. The cryo-EM data (raw and metadata) will be compressed and stored on the external 600TB or 1.2PB server in the lab.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	The cost for the KU Leuven K: drive is €11.384/100GB/year, half of which is covered by the KU Leuven Group Biomedical Sciences, meaning the lab will have to pay €5.692/100GB/year. No costs are associated with the external 600TB or 1.2PB servers where the cryo-EM data will be stored.

6. 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. *NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS*	☒ Yes, as open data ☐ Yes, as embargoed data (temporary restriction) ☒ Yes, as restricted data (upon approval, or institutional access only) ☐ No (closed access) ☐ Other, please specify:
If access is restricted, please specify who will be able to access the data and under what conditions.	Prof. Kalyan Das will be the lead contact for all requests to access data from this project.

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, privacy aspects ☐ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☒ No If yes, please specify:
Where will the data be made available? If already known, please provide a repository per dataset or data type.	☐ KU Leuven RDR ☒ Other data repository (specify) ☐ Other (specify) Cryo-EM structures will be made available in the Protein Data Bank (PDB) with cryo-EM density maps being made available in the Electron Microscopy Data Bank (EMDB). Raw data will be made available upon request.
When will the data be made available?	☒ Upon publication of research results ☐ Specific date (specify) ☐ Other (specify)

Which data usage licenses are you going to provide? If none, please explain why. *A DATA USAGE LICENSE INDICATES WHETHER THE DATA CAN BE REUSED OR NOT AND UNDER WHAT CONDITIONS. IF NO LICENCE IS GRANTED, THE DATA ARE IN A GREY ZONE AND CANNOT BE LEGALLY REUSED. DO NOTE THAT YOU MAY ONLY RELEASE DATA UNDER A LICENCE CHOSEN BY YOURSELF IF IT DOES NOT ALREADY FALL UNDER ANOTHER LICENCE THAT MIGHT PROHIBIT THAT.* *Check the RDR guidance on licences for data and software sources code or consult the License selector tool to help you choose.*	☒ CC-BY 4.0 (data) ☐ Data Transfer Agreement (restricted data) ☐ MIT licence (code) ☐ GNU GPL-3.0 (code) ☐ Other (specify)
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *INDICATE WHETHER YOU INTEND TO ADD A PERSISTENT AND UNIQUE IDENTIFIER IN ORDER TO IDENTIFY AND RETRIEVE THE DATA.*	☒ Yes, a PID will be added upon deposit in a data repository ☐ My dataset already has a PID ☐ No
What are the expected costs for data sharing? How will these costs be covered?	Data will be shared in open access repositories such as the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB).

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Data documentation and metadata management will be the responsibility of the PhD fellow (Neesha Rajesh Shewakramani) with the supervision of the promoter (Prof. Kalyan Das).
Who will manage data storage and backup during the research project?	Data storage and backup will be the responsibility of the PhD fellow (Neesha Rajesh Shewakramani) with the supervision of the promoter (Prof. Kalyan Das).

Who will manage data preservation and sharing?	The end responsibility for data preservation will be with Prof. Kalyan Das. Prof. Kalyan Das will be the lead contact for all requests to access data from this project.
Who will update and implement this DMP?	The PhD fellow (Neesha Rajesh Shewakramani) will be responsible to implement this DMP and update it as and when necessary. Prof. Kalyan Das will have the end responsibility to implement and update this DMP.