Deep learning solutions for deciphering gene regulation in the human brain

A Data Management Plan created using DMPonline.be

Creator: Niklas Kempynck

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1SH6J24N

ID: 206873

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

The human genome contains more than one million enhancers and promoters that regulate gene expression. The sequence of these regulatory regions determines which transcription factors can bind. Together they form a gene regulatory network that determines a cell's transcriptional identity. Decoding how the DNA sequence of enhancers underlies cell type specific activity is a key challenge in the further understanding of gene regulation. Moreover, it is largely unknown how multiple enhancers cooperate as a chromatin module to control gene expression. To address this challenge we will use large single-cell ATAC-seq and single-cell RNA-seq atlases and develop new computational strategies, largely based on convolutional neural networks and attention networks, to learn the logic of gene regulation in the genome. Key goals will be to incorporate large input DNA sequences around each gene to predict gene expression and chromatin accessibility and to model how dynamic regulatory programs are implemented in the genome such as feedback loops and repression. Next, we will use these approaches to decipher gene regulation in the human brain, in comparison to related mammalian species. These human brain models will allow for the scrutinization, validation and design of human brain enhancers with potential applications for gene therapy in neurodegenerative and neurodevelopmental diseases.

Last modified: 29-04-2024

Deep learning solutions for deciphering gene regulation in the human brain
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Deep learning solutions for deciphering gene regulation in the human brain
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Deep learning solutions for deciphering gene regulation in the human brain
Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

My project will mostly focus on utilizing existing multiome datasets of mouse and human brain tissue. We will potentially also generate a novel MPRA (Massively Parallel Reporter Assays) dataset to evaluate enhancer activity. Additionally, we will create high-resolution microscopy images from enhancer reporter assays. I will also write code for analysis in complete packages. Here is a detailed list of the data types and formats:
Multiome Data (mouse and human brain) (.bam, .fastq)
MPRA Dataset (.fastq, .xml, .tsv)
Microscopy Images (.tiff, .jpeg)
Code (.py, .ipynb, .sh)
Estimated volume: 20 TB

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. Designation of responsible person: Niklas Kempynck (the researcher)
2. To ensure compliance with FWO guidelines for data preservation:
   Storage and Security: We have secured sufficient storage capacity through the VSC, the VIB Data Core, and KU Leuven servers which provide robust data security measures including encrypted storage and regular backups. Data will be maintained here during the research and for at least five years afterward. Access to the data will be controlled and monitored by the responsible person.
   Data Sharing and Repositories:
   Multi-disciplinary Data: We plan to use established repositories like Zenodo for general datasets.
   Code: Private git repositories from the PI, public GitHub repositories and Zenodo will be used to store written code.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

N.A.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

N.A.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

N.A.

Deep learning solutions for deciphering gene regulation in the human brain
FWO DMP (Flemish Standard DMP)

1. 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.

				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 (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
existing scRNA-seq & scATAC-seq datasets	publicly available multiome datasets from mouse and human brain	Reuse existing data	Digital	Experimental	- textual data: FASTQ file (.fastq, zipped as .gz) - metadata: textual data(.rtf, .xml, .txt)	<10TB	
EnhancerAI code package	code package for analyzing enhancer code	New	Digital	Software	python (.py) files and notebooks (.ipynb)	< 1 GB	
Microscopy images	microscopy images of enhancer reporter assays	New	Digital	Experimental	nd2, tiff, png	< 100 GB	
AAV-MPRA plasmid libraries	enhancer reporter assays	New	Physical	Experimental			~200 µl per sample

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:

Publicly available mouse and human multiome/scATAC-seq datasets:
Zu et al. 2023 (Full mouse brain scATAC): https://www.nature.com/articles/s41586-023-06824-9
Li et al. 2021 (Mouse brain scATAC): https://www.nature.com/articles/s41586-021-03604-1
Bravo Gonzalez-Blas et al. 2023 (Mouse cortex multiome): https://www.nature.com/articles/s41592-023-01938-4
Bakken et al. 2021 (Human cortex multiome): https://www.nature.com/articles/s41586-021-03465-8
Ma et al. 2022 (Human cortex multiome): https://www.science.org/doi/10.1126/science.abo7257
Zemke et al. 2023 (Human, Macaque, Marmoset and Mouse cortex multiome): https://www.nature.com/articles/s41586-023-06819-6

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, animal data

Data re-use:
We will be re-using sequencing data generated by others. In doing so, we will adhere to the Data Transfer Agreement (DTA) and the Data Access Committee (DAC) guidelines of the data provider.
Laboratory Animal Experiments
We will be conducting experiments on mice and will seek approval from the Ethical Committee for Animal Experimentation before their initiation.

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

N/A

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

We expect that the proposed work could result in research data with potential for tech transfer and valorization. Ownership of the data generated belongs to KU Leuven and VIB in accordance with the framework agreement of both institutes. VIB 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.

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 in the comment section to what data they relate and what restrictions are in place.

- No

No third-party agreement restricts dissemination or exploitation of the data or strains generated from this project. Existing agreements between VIB and KU Leuven do not restrict publication of data.

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

N/A

2. 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).

Data will be accompanied by documentation containing all contextual and descriptive features of the research data, which allow to understand and (re)use the data. This includes data collection methods, protocols, and code explanation. Documentation is stored at the study- and the data-level, providing data provenance from the original source data to specific datasets linked to publications. Data will be generated following standardized protocols. Clear and detailed descriptions of these protocols will be stored in our lab protocol database and electronic laboratory notebook (E-notebook) and published along with the results, eg. on protocols.io (https://www.protocols.io/workspaces/aertslab/publications). Algorithms, scripts and software usage will be documented,e.g. using Jupyter Notebooks. Internally, we use git.aertlab.org to save and version the scripts. When scripts, algorithms and software tools are finalized, they will be additionally described in manuscripts and on GitHub (see www.github.com/aertslab for our previous scripts and tools). Metadata will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the E-notebook and/or in hard copy lab notebooks that refer to specific datasets. All datasets will be accompanied by metadata that is stored in our electronic lab notebook and in our central samplesheet. We have scripts that process the metadata, for example to obtain all fastq files of a certain project. Digital files will be named following a standard procedure, so that all the name of all files in a given dataset will be in the same format.
To allow long term access and use of research data will be stored or converted to open file formats as much as possible.
• Containers: TAR, ZIP
• databases: XML, CSV, JSON
• Statistics: DTA, POR, SAS, SAV
• Images: TIFF, JPEG 2000, PNG, GIF
• Tabular data: CSV, TXT
• Text: XML, PDF/A, HTML, JSON, TXT, RTF
• Sequencing data: FASTA, FASTQ
We use controlled vocabularies or ontologies when applicable to provide unambiguous meaning, for example:
• Gene Ontology: molecular function, cellular component, and biological role of RNA seq
• ENSEMBL or NBCI identifiers: gene identity
• HUGO Gene Nomenclature Committee: names and symbol of human genes
• Mouse Genome Informatics: names and symbol of mouse genes
• UniProt protein accessions: protein identity

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

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, which will include the following elements:
• Title: free text
• Creator: Last name, first name, organization
• Date and time reference
• Subject: Choice of keywords and classifications
• Structure: internal structure of the dataset, or the meaning of abbreviations (not necessary when it is clear from the in-file documentation).
• 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. Additionally, we will closely monitor MIBBI (Minimum Information for Biological and Biomedical

Investigations) for metadata standards more specific to our data type. For specific datasets, additional metadata will be associated with the data file as appropriate.

3. Data storage & back-up during the research project

Where will the data be stored?

Primary storage for active digital files will be on KU Leuven servers. KU Leuven offers fast ("J-drive) and slower ("L-drive") storage that allows reading/writing/modification of non-confidential, confidential, and strictly confidential data.
KU Leuven further offers the ManGO platform for storage and management of large volumes of active research data . This platform allows secure storage, manual and automated metadata coupling, data workflows, and file sharing.
Data that is no longer active, can be archived on the KU Leuven K-drive, which allows reading of non-confidential, confidential, and strictly confidential data.
The VSC and VIB Data Core will also be used to store data which is actively being used for research.
Algorithms, scripts and software: All the relevant algorithms, scripts and software code will be stored on the lab GitHub account (https://github.com/aertslab).

How will the data be backed up?

KU Leuven drives are backed-up according to the following scheme:
- data stored in manGO: Snapshots are made at regular intervals (hourly, daily and monthly) in case data needs to be recovered. The data itself is synchronized on two separate hardware storage systems, each 6 PB large, located at Leuven and at Heverlee (ICTS). The data is protected against calamities at either site by synchronizing it in real-time at hardware level.
- data stored on the “L-drive” is backed up daily using snapshot technology, where all incremental changes in respect of the previous version are kept online; the last 14 backups are kept.
- data stored on the “J-drive” is backed up hourly, daily (every day at midnight) and weekly (at midnight between Saturday and Sunday); in each case the last 6 backups are kept.
- data stored on the digital vault is backed up using snapshot technology, where all incremental changes in respect of the previous version are kept online. As standard, 10% of the requested storage is reserved for backups using the following backup regime: an hourly backup (at 8 a.m., 12 p.m., 4 p.m. and 8 p.m.), the last 6 of which are kept; a daily backup (every day) at midnight, the last 6 of which are kept; and a weekly backup (every week) at midnight between Saturday and Sunday, the last 2 of which are kept.

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

KU Leuven servers offer sufficient storage for active data (J/L-drive, ManGO) and archived data (K-drive). Required data-storage volumes can be easily scaled up.

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

The buildings on our campus are restricted by badge system so only employees are allowed in and visitors are allowed under supervision after registration.
Access to the “L-drive”, “J-drive”, and ManGO servers is possible only through using a KU Leuven user-id and password, and user rights only grant access to their own data, or data that was shared to them. Data in these drives are mirrored in the second ICTS datacenter for business continuity and disaster recovery so that a copy of the data can be recovered within an hour. Access to the digital vault is possible only through using a KU Leuven user-id and password, and user rights only grant access to the data in their own vault. Sensitive data transfer will be performed according to the best practices for “Copying data to the secure environment” defined by KU Leuven. The operating system of the vault is maintained on a monthly basis, including the application of upgrades and security patches. The server in the vault is managed by ICTS, and only ICTS personnel (bound by the ICT code of conduct for staff) have administrator/root rights. A security service monitors the technical installations continuously, even outside working hours. Only the PI and medical team members will be granted access to the server to deposit private data. The PI and medical team members will be the only responsible for linking patient information and/or samples, and will strictly respect confidentiality.

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

-The costs of digital data storage are as follows: 569,2€/5TB/Year for the “L-drive”, 519€/TB/Year for the “J-drive”, and 35€/TB/Year for the ManGO platform. Data storage and backup costs are included in general lab costs.

4. 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...).

According to KU Leuven RDM policy, relevant research data will be preserved on the university's servers for a minimum of 10 years. Such data include data that are at the basis of a publication, that can only be generated or collected once, that are generated as a result of a substantial financial or personal effort, or are likely to be reused within the research unit or in wider contexts.

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

As a general rule all research outputs (data, documentation, and metadata) related to publications will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org). 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 separate data repository.
Other research data will be archived on KU Leuven servers as described above.

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

The costs of digital data storage are as follows: 569,2€/5TB/Year for the "K-drive" and the “L-drive”, 519€/TB/Year for the “J-drive”, and 35€/TB/Year for the ManGO platform. Data storage

and backup costs are included in general lab costs.

5. Data sharing and reuse

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

- Yes, in an Open Access repository

Computational workflows, models, and metadata will be stored on platforms such as Github, Kipoi, and Zenodo with proper versioning.
To ensure data findability, links and references these datasets, workflows and modes will be included in the data availability statements of the associated publication.

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

N.A.

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 in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights
- No

The researchers involved and the IP team of the VIB TechTransfer Office shall make the necessary arrangements in order to maintain an embargo on the public access of research data, at least until the essential steps in securing intellectual property (e.g. the filing of a patent application) have been taken. 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.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All research outputs (data, documentation, code, and associated metadata) 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.
Computational workflows, models, and metadata will be stored on platforms such as Github, Kipoi, and Zenodo with proper versioning.

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.

Datasets will not be licensed
Code will be licensed according to VIB standards: open for academic, license for industry

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

10.5281/zenodo.10868679

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

It is the intention to minimize the cost by using free-to-use data repositories whenever possible. Data sharing costs will be covered by the laboratory budget.

6. Responsibilities

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

The researcher (Niklas Kempynck) and collaborators who generate date for the project.

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

The researcher (Niklas Kempynck) with the hulp from lab technicians.

Who will manage data preservation and sharing?

The researcher (Niklas Kempynck) with support from the research and technical staff involved in the project

Who will update and implement this DMP?

The researcher (Niklas Kempynck)