A novel paradigm for Precision Medicine: sparse non-linear Neural Networks for end-to-end Genome Interpretation

A Data Management Plan created using DMPonline.be

Creators: Daniele Raimondi, n.n. n.n.

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / **URL:** 12Y5623N

ID: 199290

Start date: 01-10-2022

End date: 30-09-2025

Project abstract:

In the last decade, the surge of high-throughput sequencing and big data technologies seemed almost unstoppable, to the point that even finally uncovering all aspects of our genome and enabling precision medicine seemed at reach. Ten years later, notwithstanding many advances in genetics, our genome is still hiding most of its secrets. When it comes to disorders with complex aetiology, the bottleneck has indeed only shifted from a data availability to a data interpretation problem, since the classical computational approaches have been largely unsuccessful in explaining these diseases.
Here we address this problem by building the foundations of a conceptually novel end-to-end Genome Interpretation (EtEGI) framework which aims at directly modeling the genotype-to-phenotype relationship using cutting-edge Machine Learning (ML) approaches, such as Neural Networks (NNs). EtEGI has just become feasible due to the critical mass of genomics data and the popularization of flexible NN libraries. We propose a fully differentiable framework of NN tailored solutions to deal with these unique kinds of genomics and phenotypic input and output data, addressing each NN modeling challenge in a biologically meaningful manner. EtEGI will be a fundamental step towards posing the basis for true precision medicine, and to do so we will address both ambitious ML and biological challenges related to modeling, accountability, scalability, interpretability and explainability of ML methods.

Last modified: 08-05-2023

A novel paradigm for Precision Medicine: sparse non-linear Neural Networks for end-to-end Genome Interpretation
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)

In this project, we will use genomic and phenotypic data produced by third party institutions that collected it. We will not collect the data ourselves, and we will not generate them. We will work with personal data such as Whole Genome and Whole Exome sequencing (WGS, WES) and phenotypic data. The data are provided in anonimized form upon signing a Data Access Agreement. Sequencing data are generally in Variant Calling Format (VCF) and phenotypic data are CSV or Excel files. The file size depends on the dataset, but it can be 1Gb per WGS sample. WES samples are few hundred Mb each. Phenotypic data have negligible size.

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)

The data are not public, they are provided by third party institutions upon signing a Data Access Agreement. The third party institutions take care of the indefinite permanence of the data. During the research, we will store the data within the ESAT-STADIUS LAN, on a specific hard disk of a specific machine we purposely bought. Since the data come from third party institutions and the Data Access Agreement prevents us from sharing it further, we do not need to provide long term storage. The persistence of the data is guaranteed by the third party institutions that distribute it. We cannot share the data, per Data Access Agreements terms.

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)

Depending on the Data Access Agreements (DAA) stipulated with the third party data providers, we might have to delete the data after the end of the project. Nonetheless, any other institution can request the same data from the data providers, upon DAA signature.

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)

The data will be stored an elaborated on a machine that is physically in the ESAT-STADIUS server room. The data will never leave a specific hard disk in that machine. Access to the machine and the hard disk is restricted to the allowed users using UNIX ACLs.

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

N/A

A novel paradigm for Precision Medicine: sparse non-linear Neural Networks for end-to-end Genome Interpretation
DPIA

DPIA

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

- Not applicable

A novel paradigm for Precision Medicine: sparse non-linear Neural Networks for end-to-end Genome Interpretation
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

A novel paradigm for Precision Medicine: sparse non-linear Neural Networks for end-to-end Genome Interpretation
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, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
DDD		reuse	digital	experimental	VCF, CSV, TXT	<1TB	
UK10k		reuse	digital	experimental	VCF, CSV, TXT	<1TB	

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:

The data are permanently stored in the EGA https://ega-archive.org/ . They are produced by third party institutions like https://www.uk10k.org/ and https://www.ddduk.org/

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.

- No

We have already obtained a Data Access Agreement with the third party data providers. Ethical issues related to data collection have been dealt with by the data providers before collecting the data. We will use the data only for the research purposed specified and approved in the Data Access Agreement.

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.

- Yes

In this project, we will use genomic and phenotypic data produced by third party institutions that collected it. We will not collect the data ourselves, and we will not generate them. We will work with personal data such as Whole Genome and Whole Exome sequencing (WGS, WES) and phenotypic data. The data are provided in anonimized form upon signing a Data Access Agreement. Sequencing data are generally in Variant Calling Format (VCF) and phenotypic data are CSV or Excel files. The file size depends on the dataset, but it can be 1Gb per WGS sample. WES samples are few hundred Mb each. Phenotypic data have negligible size.

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

The project is fundamental research. At the moment we are not focusing on commercialization of any kind, since the research is at an early stage and therefore the future possibilities are unclear.

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.

- Yes

Yes, all the datasets we mentioned have been obtained through Data Access Agreements with third party data providers. The DAA restricts our ability to share further the 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

We will follow what has been stipulated in the Data Access Agreement.

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).

I am working on this data alone and I cannot redistribute it.

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.

- No

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

Where will the data be stored?

The data will be stored an elaborated on a machine that is physically in the ESAT-STADIUS server room. The data will never leave a specific hard disk in that machine. Access to the machine and the hard disk is restricted to the allowed users.

How will the data be backed up?

The data are permanently available from our third party data providers. We use a copy of the data to prototype locally. We are not backing up the data, since we cannot redistribute them and anyone can access them from the third party data providers websites, upon stipulating a Data Access Agreement.

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

Currently we have enough hard disk space. We will add more dedicated hard disks if necessary. Access to these disks will be restricted to the allowed user.

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

The data will be stored and elaborated on a machine that is physically in the ESAT-STADIUS server room. The data will never leave a specific hard disk in that machine. Access to the machine and the hard disk is restricted to the allowed users.

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

No additional costs are required, the data are locally stored for model prototyping purposes on a machine we commonly use to develop deep learning models.

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...).

We cannot re distribute the data, since we are bound by the Data Access Agreement terms we stipulated to obtain the data in the first place. DAA impose to delete the data once the project is finished.

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

The data are permanently available, upon DAA stipulation, from the third party data providers institutions. We will not store the data ourselves.

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

There are no costs, we will not store the data.

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.

- No (closed access)

We cannot redistribute or share the data, due to the Data Access Agreement terms we signed to obtain the data.

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

The people mentioned in the Data Access Agreements can access the data. I am working on this project alone.

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, Other

The data access agreement we stipulated to access the data prevent us from sharing it or re distribute it.

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

The data will not be made available.

When will the data be made available?

Never.

Which data usage licenses are you going to provide? If none, please explain why.

The data will not be made available, since it would infringe the DAA we stipulated to obtain them.

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.

- No

The data will not be made available, since it would infringe the DAA we stipulated to obtain them.

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

The data will not be made available, since it would infringe the DAA we stipulated to obtain them.

6. Responsibilities

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

Daniele Raimondi

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

Daniele Raimondi

Who will manage data preservation and sharing?

Daniele Raimondi

Who will update and implement this DMP?

Daniele Raimondi