Maastricht University

Evaluation of Oxford Nanopore Sequencing data classification beyond canonical nucleotides

Bachelor Thesis - Data Science and Artificial Intelligence

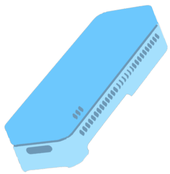
Adèle Imparato - 21st of June 2022

Plan

1. Introduction
2. Description of the technology
3. RQ1
   - Experiment
   - Results
   - Conclusion
4. RQ2
   - Experiment
   - Results
   - Conclusion

Problem statement

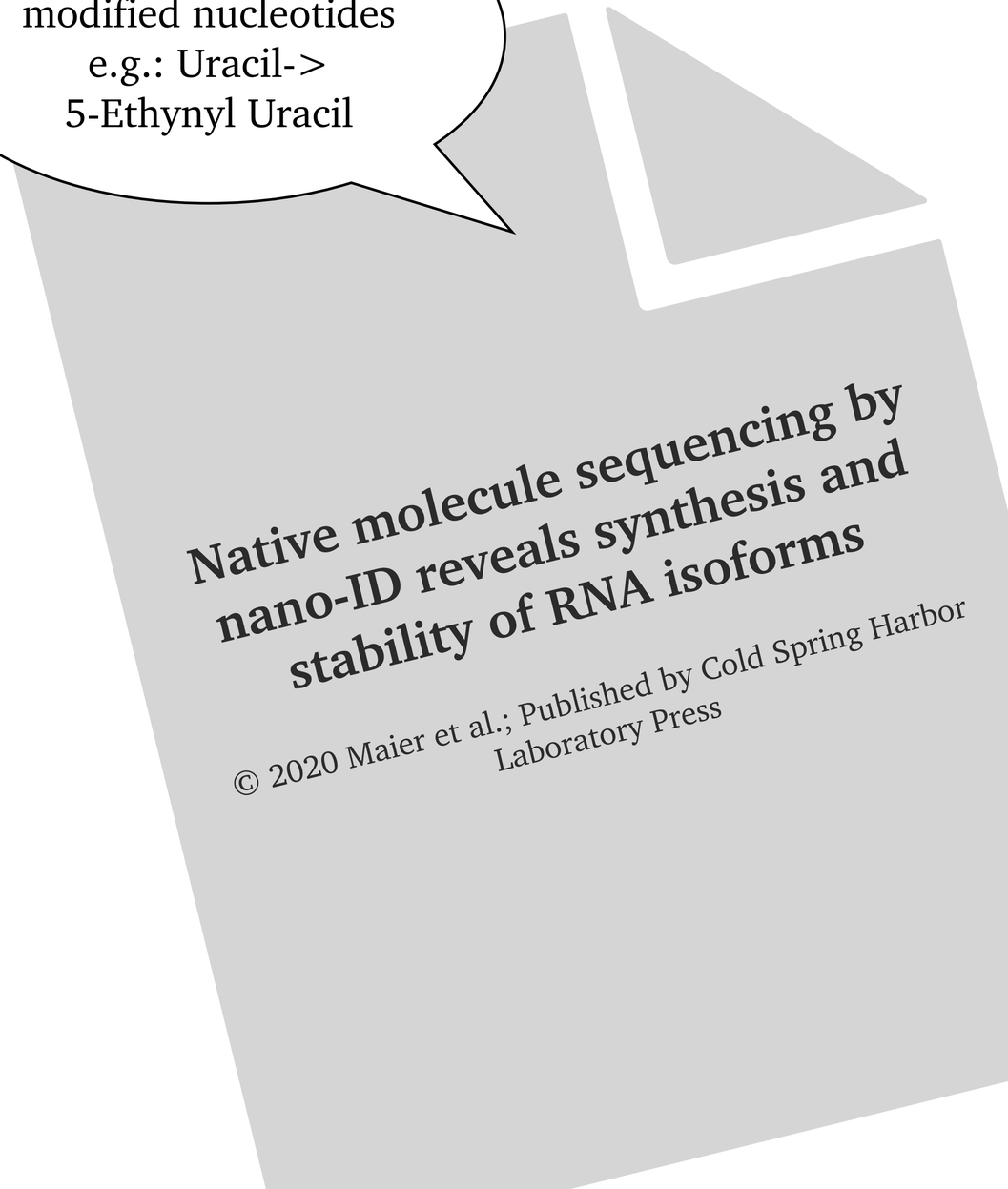
Context:
Sequence RNA or DNA to discover interesting genes that code for a disease for instance



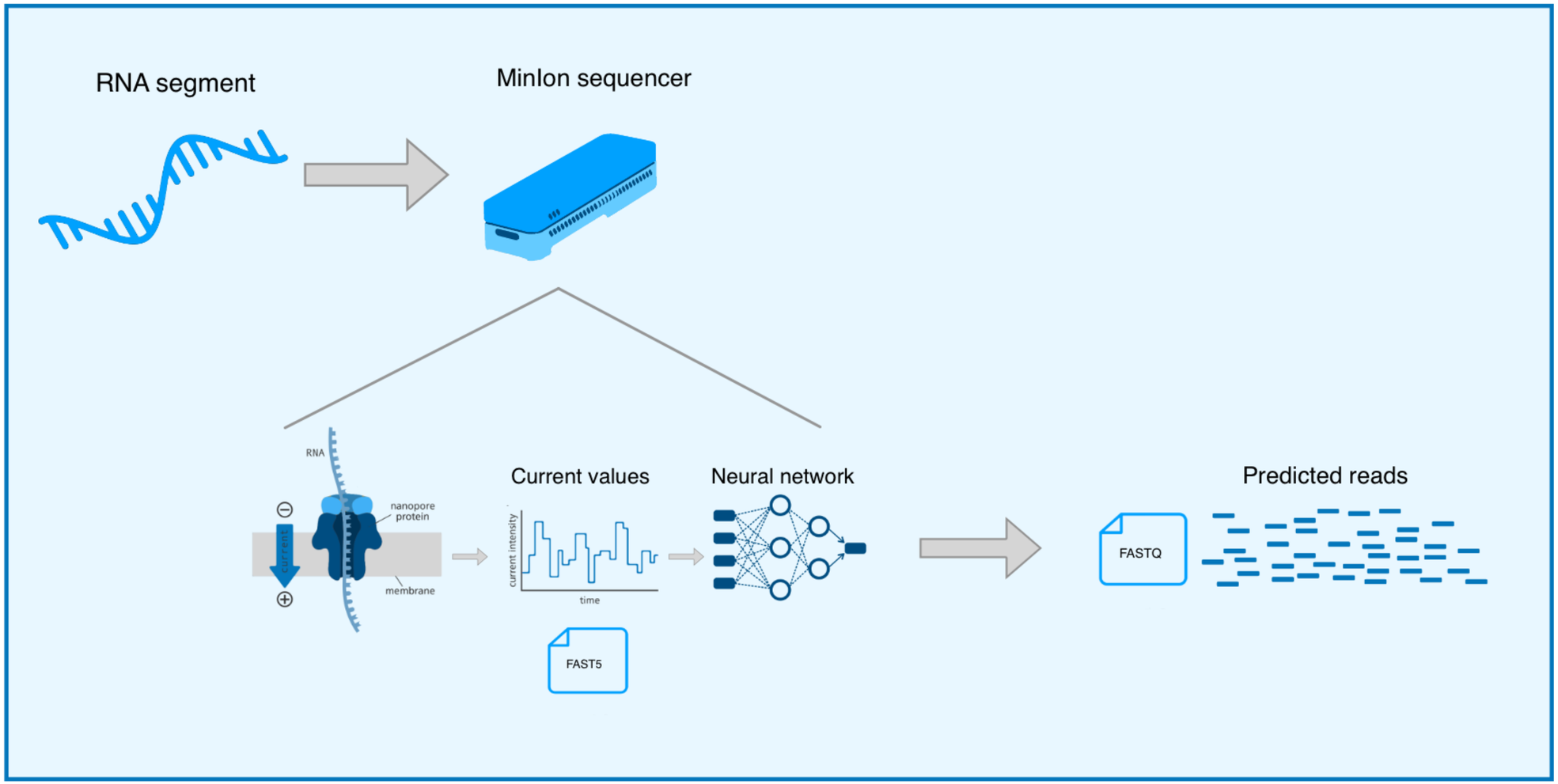
State-of-the-art:
Nanopore Sequencing developed by Oxford Nanopore Technologies (ONT)

Goals:
1. Test the accessibility of data and reusability of the nano-ID software
2. Evaluate the NN's performance to detect the presence of RNA isoforms

RNA isoforms ?
modified nucleotides
e.g.: Uracil->
5-Ethynyl Uracil



The technology - nanopore sequencing



RQ1

- **RQ1 (a)** Can the data used in Maier et al.'s research be called 'FAIR' ?

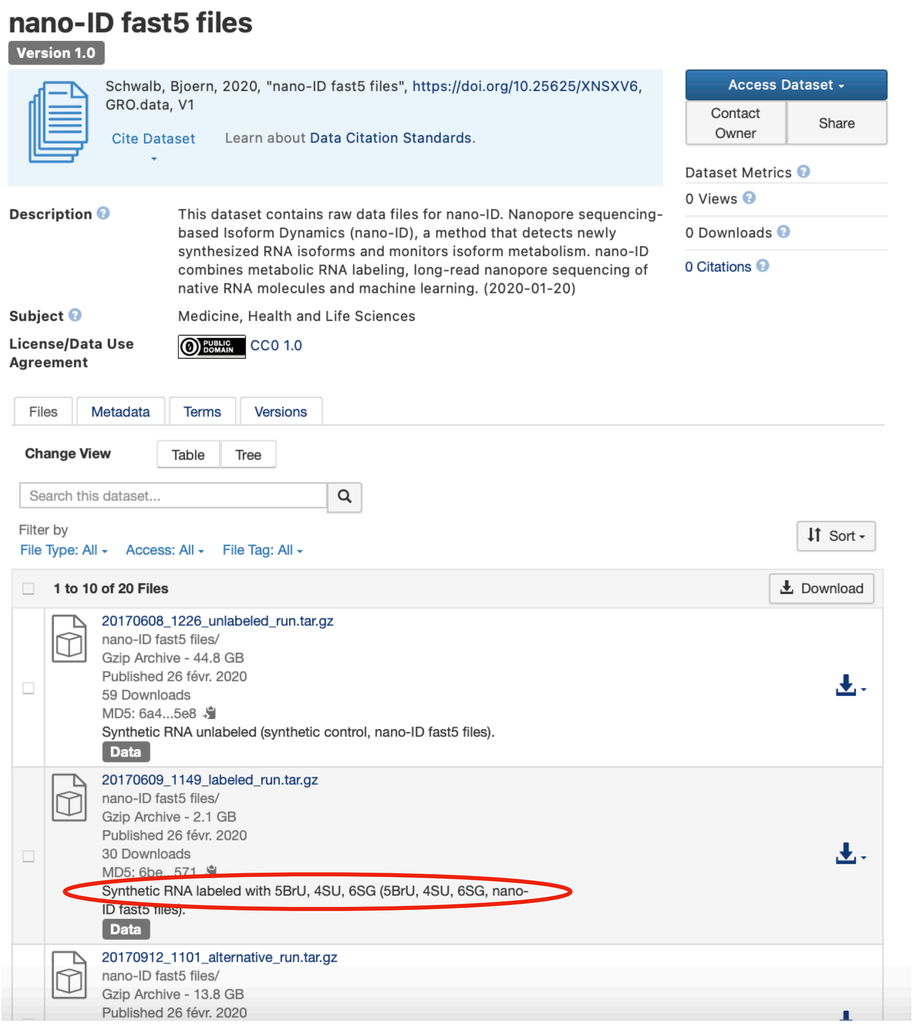
- **RQ1 (b)** If yes, can the results be reproduced when running the nano-ID neural network over these nanopore data, meaning that this research respects the principles of Open Science and Fair Data ?

Hardware and software should be **accessible** and **reusable**

Data should be :
- **Findable**
- **Accessible**
- **Interoperable**
- **Reusable**

= 'FAIR'

RQ1 (a) - Are data ‘FAIR’ ?



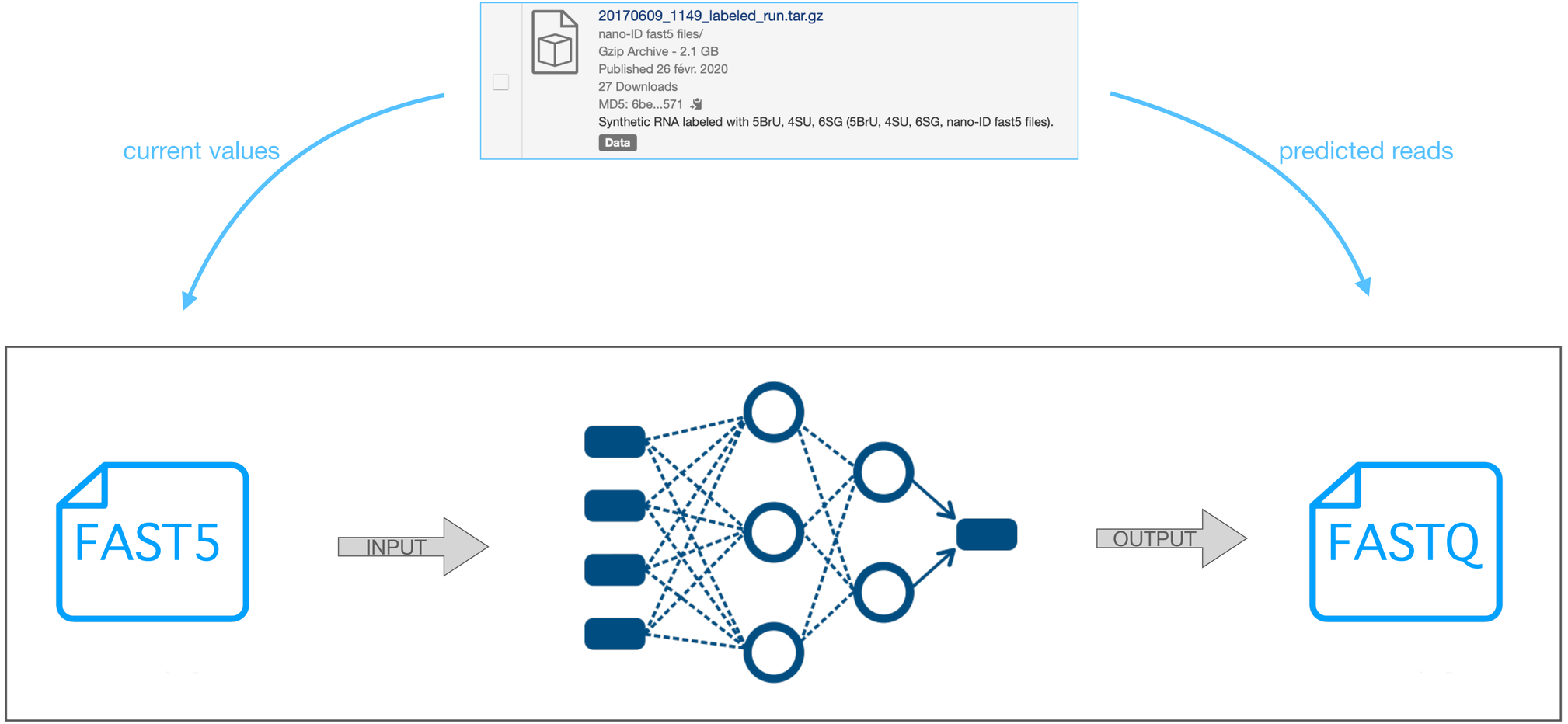
Göttingen Research Online / Data. (2020, February 26). *nano-ID fast5 files*. https://data.goettingen-research-online.de/dataset.xhtml?persistentId=doi:10.25625/XNSXV6

- ➢ stable connection
- ➢ wget
- ➢ powerful computer

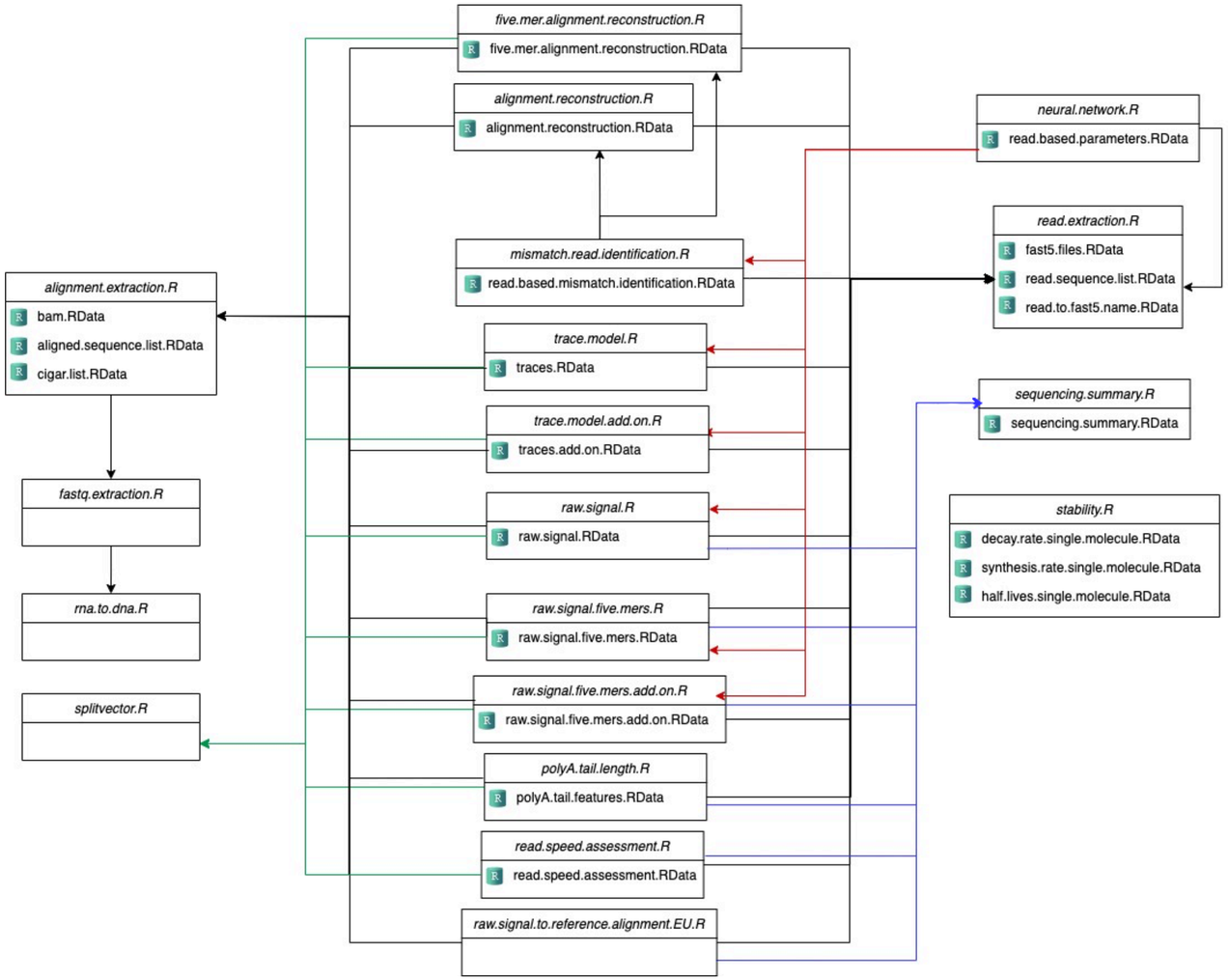
Findable ✓
Accessible ✗
Interoperable ✗
Reusable ✗

≠ ‘FAIR’

RQ1 (b) - Can the neural network be run ?



RQ1 (b) - Can the neural network be run ?



- ➢ poorly documented
- ➢ NN cannot be run

RQ1

- **RQ1 (a)** Can the data used in Maier et al.'s research be called 'FAIR' ?

- **RQ1 (b)** If yes, can the results be reproduced when running the nano-ID neural network over these nanopore data, meaning that this research respects the principles of Open Science and Fair Data ?

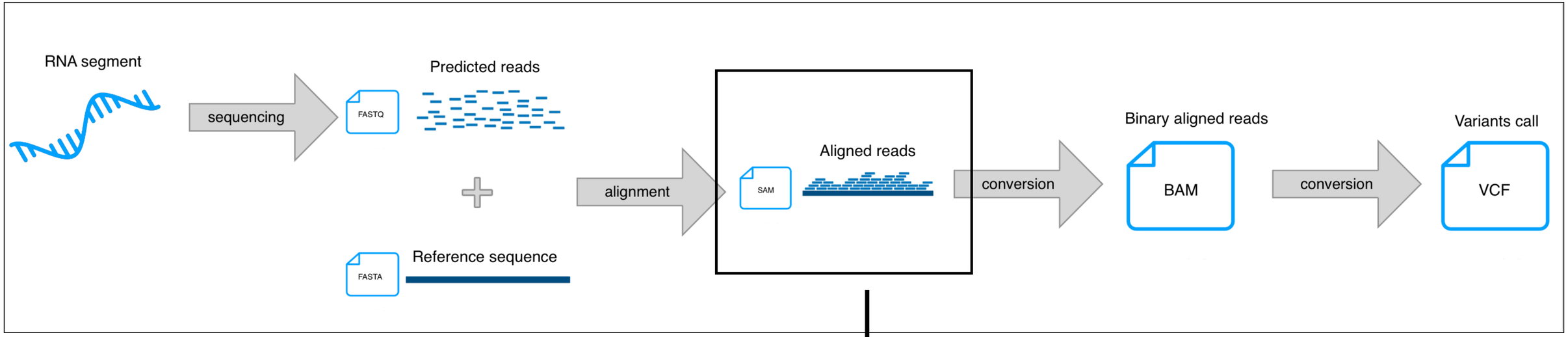
- **RQ1 (a)** The data in Maier et al.'s research are not 'FAIR'.

- **RQ1 (b)** The results of the neural network cannot be reproduced and thus the overall research does not respect the principles of Fair Data and Open Science.

RQ2

- **RQ2** What is the performance of the nano-ID neural network in terms of misclassification ? Does it enable one to detect RNA isoforms ?

How do we evaluate the performance ?



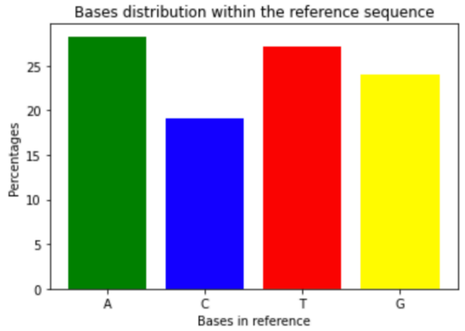
Interactive tool
for visualization

Results

synthetic RNA labeled with 5BrU, 4SU and 6SG

Complete alignment info

PRED \ REF	A	C	T	G
A	0.94	0.01	0.03	0.03
C	0.01	0.93	0.09	0.01
T	0.02	0.05	0.85	0.01
G	0.02	0.01	0.02	0.95



only variants

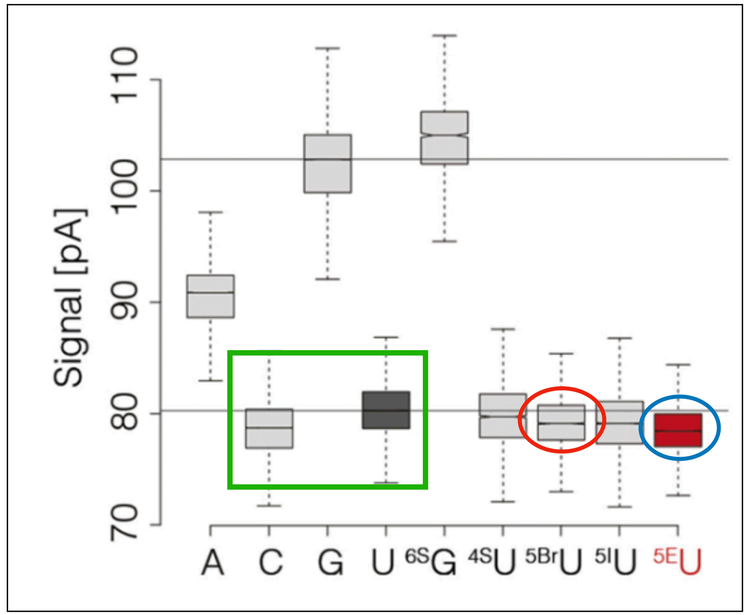
VARIANTS	A	C	T	G
A	/	0	0.05	1
C	0.14	/	0.92	0
T	0	0	/	0
G	0.86	1	0.03	/
proportion in REF	0.14	0.04	0.74	0.06

human K562 cells cultured in the presence of 5EU for 60min

only variants

VARIANTS	A	C	T	G
A	/	0.60	0.71	0.82
C	0.11	/	0.22	0.05
T	0.45	0.35	/	0.12
G	0.43	0.05	0.06	/
proportion in REF	0.17	0.24	0.35	0.24

Results



PRED \ REF	A	C	T	G
A	0.94	0.01	0.03	0.03
C	0.01	0.93	0.09	0.01
T	0.02	0.05	0.85	0.01
G	0.02	0.01	0.02	0.95

VARIANTS	A	C	T	G
A	/	0	0.05	1
C	0.14	/	0.92	0
T	0	0	/	0
G	0.86	1	0.03	/
proportion in REF	0.14	0.04	0.74	0.06

VARIANTS	A	C	T	G
A	/	0.60	0.71	0.82
C	0.11	/	0.22	0.05
T	0.45	0.35	/	0.12
G	0.43	0.05	0.06	/
proportion in REF	0.17	0.24	0.35	0.24

RQ2

- **RQ2** What is the performance of the nano-ID neural network in terms of misclassification ? Does it enable one to detect RNA isoforms ?

- **RQ2** The nano-ID neural network seems to perform well enough in order to detect the presence of 5BrU and 5EU despite the reduced amount of conducted experiments.

Thank you for your attention !