3rd Year Report

Susana Roman Garcia

September 8, 2023

Table of contents

Preface

This year has been mostly focused on building up knowledge on how to put into action the idea of making a more ethical and reproducible PhD. In order to understand the progress I have made this year, I describe in the following capters the importance of thinking about ethics and reproducibility, what are and how to use Data Hazards, as well as a breakdown of creating reproducible computer models.

At this point in time I am in a place where I have started writing my thesis in ways which allow me to keep good version control, which helps manage the source code and documents by keeping track of all the version modifications. In fact, I am using this year's annual report as an exercise to write an example book using Quarto, which allows for a dynamic implementation of markdown files and python code all in one project, which can be version controlled via Git.

This is a Quarto Book:

This document is written using a Quarto book. Quarto allows for a dynamic implementation of different types of files that can be version controlled, which is extremely helpful to create a pipeline of work all in one place and that is traceable too.

Quarto is an open-source scientific and technical publishing system built on Pandoc. It allows you to weave together narrative text and code to produce elegantly formatted output as documents, web pages, blog posts, books and more.

Quarto is at its core multi-language and multi-engine (supporting Knitr, Jupyter, and Observable today and potentially other engines in the future); where you can have all your code and narrative text in one. For a full breakdown and FAQs of how Quarto works, you can have a look here.

The good thing for me personally, is that, thanks to Quarto, I can write my thesis chapters using markdown, which is a lot more intuitive than LaTeX, in my opinion. Moreover, it allows for much better version control compared to Microsoft Word, and allows for helpful traceability of materials included in the document I am writing; which in itself has many benefits as it immensely helps to be able to go back to a previous version where everything was working before I broke the code (yet again).

Part of my work includes work which runs with different python scripts, and using Quarto means that I can embed python code if needed to explain how certain functionalities of the models work. It also allows for code using R to run within a Quarto document, as well

as Jupyter Notebooks, and includes HTML implementations, which can all help with functionalities such as creating tables, adding images, and rendering the document in different formats (website html, or pdf, for example).

Different modes to view this report

- **In an internet browser**: through this url.
- **In `.pdf` format**: If you are viewing this document in your browser through a url, you can click on the top left icons of the document to download it.
- **Source code in GitHub**: You can also click on the top left icons of the document to access the source code.

1 What is my PhD research question?

The biological question that started this PhD was how do Calcium/calmodulin-dependent protein kinase II (CaMKII) and N-methyl-D-aspartate receptor (NMDARs) in the postsynaptic neuron interact and enable the processes of learning and memory?

But as time went by, I realised more and more how important looking at the ethics of the research that we do is. How biased science actually is, and how we continue to carry these if we don't look at them in the face. Additionally, how much research time and money is wasted by doing experiments which cannot be reproduced or replicated later on. Hence my emphasis on these topics.

An additional, and important, aim of this PhD project to highlight and talk about some of the things I care most in research: making it transparent, inclusive, and accessible.

When studying learning and memory at the molecular level, in health and disease, it has been shown that NMDAR and CaMKII together with their interactions with other proteins within neuronal spines can influence their shape and size (Fink and Meyer 2002). Long-term modifications of synaptic strength, such as LTD (Long Term Depression) and LTP (Long Term Potentiation), involve diverse chemical pathways and have been the primary mechanisms used to study the molecular basis of learning and memory (Blundon and Zakharenko 2008). So what exactly is happening at the cellular and molecular level during memory formation?

These are the biological prompts that I look at when creating 3D models of the molecules in question. I use mainly MCell and python to do so. In order to give you a better overview of the aims, types of data used, methods and applications of this research, please see below Table 1.1. In addition to the biological aspects of this PhD, as mentioned above, I have made a big effort into making my PhD accessible, reproducible and more ethical. It has transformed into a case study example of how to establish procedures for more ethical and reproducible research, which means future researchers can efficiently re-use and build up on what I have created.

Table 1.1: PhD overview

Wide-view angle of this PhD project
Aims of this PhD: • Explain how specific molecules work together during memory. • Develop new ways of 3D modelling to look at time and space dynamics of molecular interactions. • Bring awareness of the importance of implementing ethics and reproducibility into a PhD.
Type of data used: | | - Kinetic rates of molecule interactions, molecular concentrations collected from literature and databases. | - Numbers obtained from either wet-lab experiments or mathematically calculated. |
Methods: • Models written with standardised open source languages: python, bionetgen Language. • Numbers obtained from either wet-lab experiments or mathematically calculated. • Run locally or in cluster if simulations are more computationally expensive.
Applications and significance • Other researchers can build from these models to create further predictions for potential pharmacological applications. • Dysregulation of the molecules I look at have been suggested to have a potential impact in Alzheimer's disease, as well associated with multiple forms of spineopathies (Ghosh and Giese 2015), (Robison 2014).

To give a better idea of what the modelling might look like, I drew Figure 1.1, which shows a somewhat simplified version of what the graphical user interface of CellBlender can look like. It includes molecules and reactions, as well as placement in a 3D cell.

In brief, I create 3D models which simulate interactions between CaMKII and NMDAR in the postsynaptic neuron, to understand how memory works in animal brains.

1.1 Why use Computational Modelling to study biological systems?

Some of the main reasons for using modelling are:

1. Biological systems are complex and multiscale. Models can help us to integrate experimental data, facilitating theoretical hypotheses, and addressing "what if" questions.
2. Models aim to make clear the current state of knowledge regarding a particular system, by attempting to be precise about the elements involved and the interactions between them. Doing this can be an effective way to highlight gaps in understanding.
3. Related to point one, models then serve to combine knowledge from different published research, and make biological predictions which can then serve as hypothesis to be tested empirically by experimentalists.
4. Computer-simulated experiments can help guide the wet-lab process by narrowing the experimental search space, enabling more cost, time-effective and waste-free research, as well as more ethical research too as we reduce animal suffering through reduction of animal research.

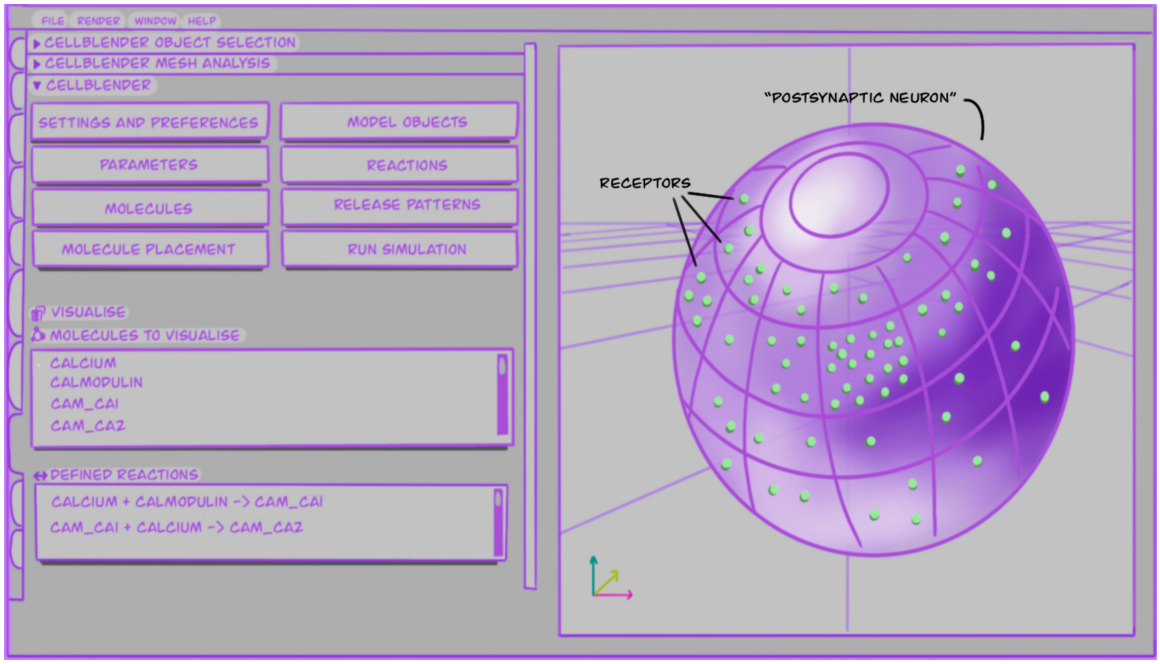

Figure 1.1: A 3D model of a postsynaptic dendritic head, in a schematic of the CellBlender interface.

2 Ethics and Reproducibility emphasis in this PhD and why it matters

The work I have carried out through this PhD so far, has had a very heavy focus on thinking about Ethics and Reproducibility as an embedded part of my research, not as a last minute add-on. This is because I believe it is of utmost importance that there is a shift in how we proceed with science. A science that currently works under a capitalistic mechanism which advocates for mass production of positivist, fast research that historically has benefitted some groups disproportionately (usually cis-male, white, able-bodied individuals [1]) (Webb, Etter, and Kwasa 2022), (Diogo et al. 2023), (Branch et al. 2022).

2.1 Science for the profit of whom?

There is a historical heritage of monetary incentivization to move towards drug discovery and the profit that comes from this, and how this has played a key role in biasing research towards drug discovery to "fix" individuals, without their wellbeing being necessarily at the forefront. Dosi et al. (2023) provide an long term review of Big Pharma and monopoly capitalism, but there are many examples of how companies move scientific research in a way that is driven by economic profit; and how there's a constant pull to publish more and publish first.

The book *Warp and Weft* by Fennen (2021), with a focus on psychiatry and neuroscience, looks at some of these sciences' history and examines the ways they have been, and continue to be used as a colonial force. Enforcing a global North science on to the world, as well as describing how many times this enforcement has been led by economic profit for only a few. One good example it cross-references is the "The Mega-Marketing of Depression in Japan" by GlaxoSmithKline, originally spoken about in the book *Crazy Like Us* by Watters (2010).

2.1.1 Historical oppressive biases:

It is because of these histories, that I want to attend to them in the work that I do. If we ignore thinking about the ethics, philosophy and history of the research that we do, we may

[1] This bias towards benefiting cis-male, white, able-bodied people does not mean they do not suffer, and does not negate the existence of the issues they may experience too. For more information on how society is biased in a way that provides privileges in a certain order/hierarchy, and how to handle it, see (DiAngelo 2018),or (Delgado 2022).

forget where certain ontologies [2] and basis of knowledge come from. Therefore continuing to pretend that these topics are not necessary to think about, whilst a privileged group continues to perpetuate oppressive biases towards historically marginalized groups.

In a presentation I gave in 2022, I give a few examples of biases that continue to happen in science, including examples of racism, sexism, ableism and speciesism (Garcia, Sterratt, and Stefan 2022). A good example of embedded biases in science is given by Branch et al. (2022) as they eloquently articulate how a desire to quantify and establish hierarchies among organisms was not purely for scientific interest, but that there is extensive evidence in the fact that the roots of evolutionary biology, which serves as a baseline for many other disciplines like neuroscience, are steeped in histories of white-supremacism, eugenics,and scientific racism. They discuss the definition of the "Not-So-Fit", and how this limits the diverse thought and investigative potential in biology. This is of importance for my PhD, as I use hierarchies and models of biology that are based on a historical context of how science has reached it's current status of knowledge.

2.2 Slowing down...

As a response to a fast-paced, profit-driven science, a few Slow Science Manifestos have been published, notably *Another Science is Possible: A Manifesto for Slow Science* by Stengers (2018), which maintains that in order to make higher quality education and science, it needs to serve society as a whole, and calls, among other things, for an "accountability in the knowledge society versus profitability in the knowledge economy".

Moreover, as long as we continue to create fast research without regard for reproducibility, we will continue to experience what some now call a "Reproducibility Crisis" (Baker 2016), (Treves 2022), where we find that, as work is done into trying to reproduce previous published results, this is not possible. The reproducibility or replicability crisis (more on these terms below) undermines the credibility of theories of scientific knowledge; as an essential part of the scientific method is to be able to repeat and reproduce or falsify empirical results and theories.

There is an argument to be made that making research more reproducible and ethical takes more time. This is precisely why slowing down can help in creating higher quality research that serves all in society.

2.3 Importance of reproducibility

During my PhD work so far, one aspect of the research that has been a challenge is to find accurate parametrization of values for protein dynamics. This is a known issue for most of us who create computational models of biological systems. Wieber and Hocquet (2020) call

[2] Ontologies meaning here "a set of concepts and categories in a subject area or domain that shows their properties and the relations between them." as defined by the Oxford Languages dictionary.

it an "epistemic opacity" when talking about lack of clarity in Computational Chemistry, where this opacity is entangled in methods and software alike.

This of course leads to reproducibility issues, and as this unfolds, it becomes clear that the "untrustworthiness" of research is also an issue for many other researchers. In fact, a survey of 1576 scientists published in Nature (Baker 2016) reported that over 70% of the participants failed to reproduce others' experiments and over 50% failed to reproduce their own results.

Interestingly, Tiwari et al. (2021), assessed the reproducibility of 455 mathematical models in systems biology and found that about 50% of published models were not reproducible either due to incorrect or missing information in the manuscript.

Making an effort into creating research that is reproducible can help to avoid wasting resources, including having to repeat the same experiment questions again and again because results from one study could not be reproduced or replicated by other groups.

2.3.1 Defining reproducibility vs replicability

These terms have been used interchangeably for a while, or their meanings being swapped depending on the field of study (Claerbout and Karrenbach 1992), (Ivie and Thain 2018), (Plesser 2018).

Here, we use the definition used by (Turing Way Community et al. 2019), where reproducible research is understood as work that can be independently recreated from the same data and the same code that the original team used. Reproducible is distinct from replicable, robust and generalisable as described in the table below (Figure 2.1).

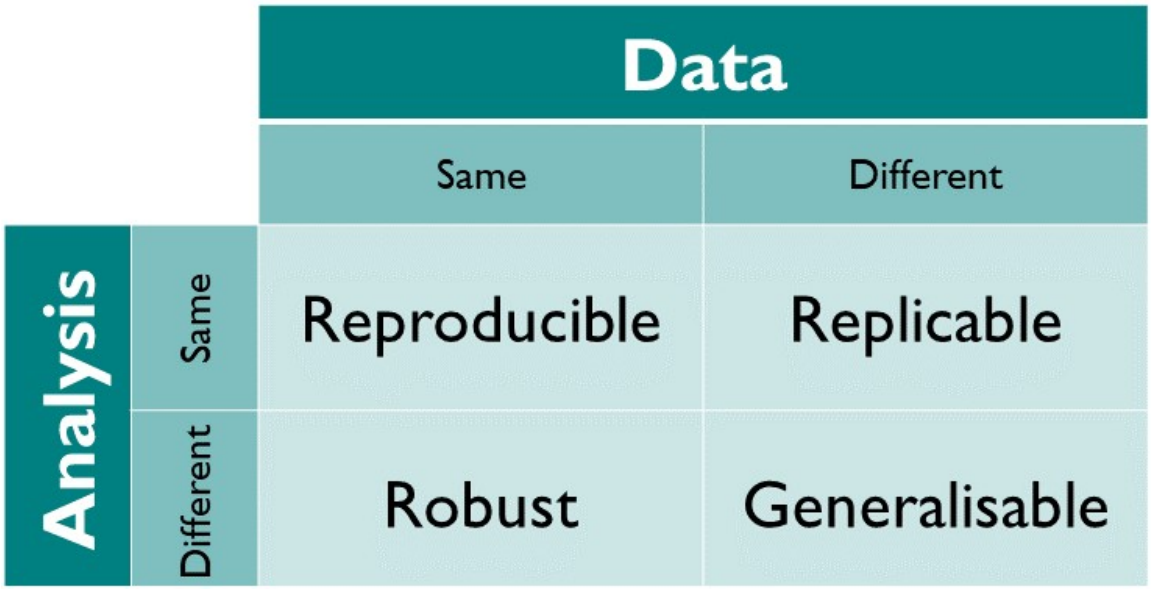

Figure 2.1: How the Turing Way defines reproducible research.

The different dimensions of reproducible research described in the matrix above have the following definitions, taken from the Turing Way booklet:

- **Reproducible**: A result is reproducible when the *same* analysis steps performed on the *same* dataset consistently produces the *same* answer.
- **Replicable**: A result is replicable when the *same* analysis performed on *different* datasets produces qualitatively similar answers.
- **Robust**: A result is robust when the *same* dataset is subjected to *different* analysis workflows to answer the *same* research question (for example one pipeline written in R and another written in Python) and a qualitatively similar or identical answer is produced. Robust results show that the work is not dependent on the specificities of the programming language chosen to perform the analysis.
- **Generalisable**: Combining replicable and robust findings allow us to form generalisable results. Note that running an analysis on a different software implementation and with a different dataset does not provide generalised results. There will be many more steps to know how well the work applies to all the different aspects of the research question. Generalisation is an important step towards understanding that the result is not dependent on a particular dataset nor a particular version of the analysis pipeline.

2.4 Ethics and Reproducibility go together

Entangled with reproducibility, is thinking about ethics. Because no matter how efficient and reproducible an outcome may be, if it's harming a group of individuals, how good really is this research? Likewise, if a project has considered and described potential bias and harms of their data, but then does not share enough material for their research to be reproduced by others, are we really advancing?

Thinking about reproducibility can in turn help to think how you will share your data, as well as where your own data has come from. Hence, reaching an increased awareness of how your data was sourced and its ethics and potential biases. In order to showcase how I see these topics as being interwoven, I presented a poster titled "Bias and reproducibility in a computational neurobiology PhD's journey" (Figure 2.2) at the International Conference on Systems Biology (ICSB) in October 2022. On the left side of the poster, I share how to think about the ethics and bias of your research, and on the right side I provide tools for reproducibility. I also wrote about this more in depth in this GitHub repository here.

I created this poster because I wanted to showcase, at a conference full of scientists at different stages of their research, how I take into account bias an reproducibility in my research, and how they could too.

Working with ethics, philosophy, reproducibility and an openness to discuss the wider context of where our research rests, may add a bit of time to the research timeline, but can very much enrich a fuller and more complex understanding of the shortcomings of our research and how to do better moving forward.

Following on the idea that scientists are great at selling the gains in efficiency and accuracy of their work, but less well-practiced in thinking about the ethical implications of our work, I present a framework developed to think about dangers or risks involved with your data and research: Data Hazard Labels, see following Chapter 3.

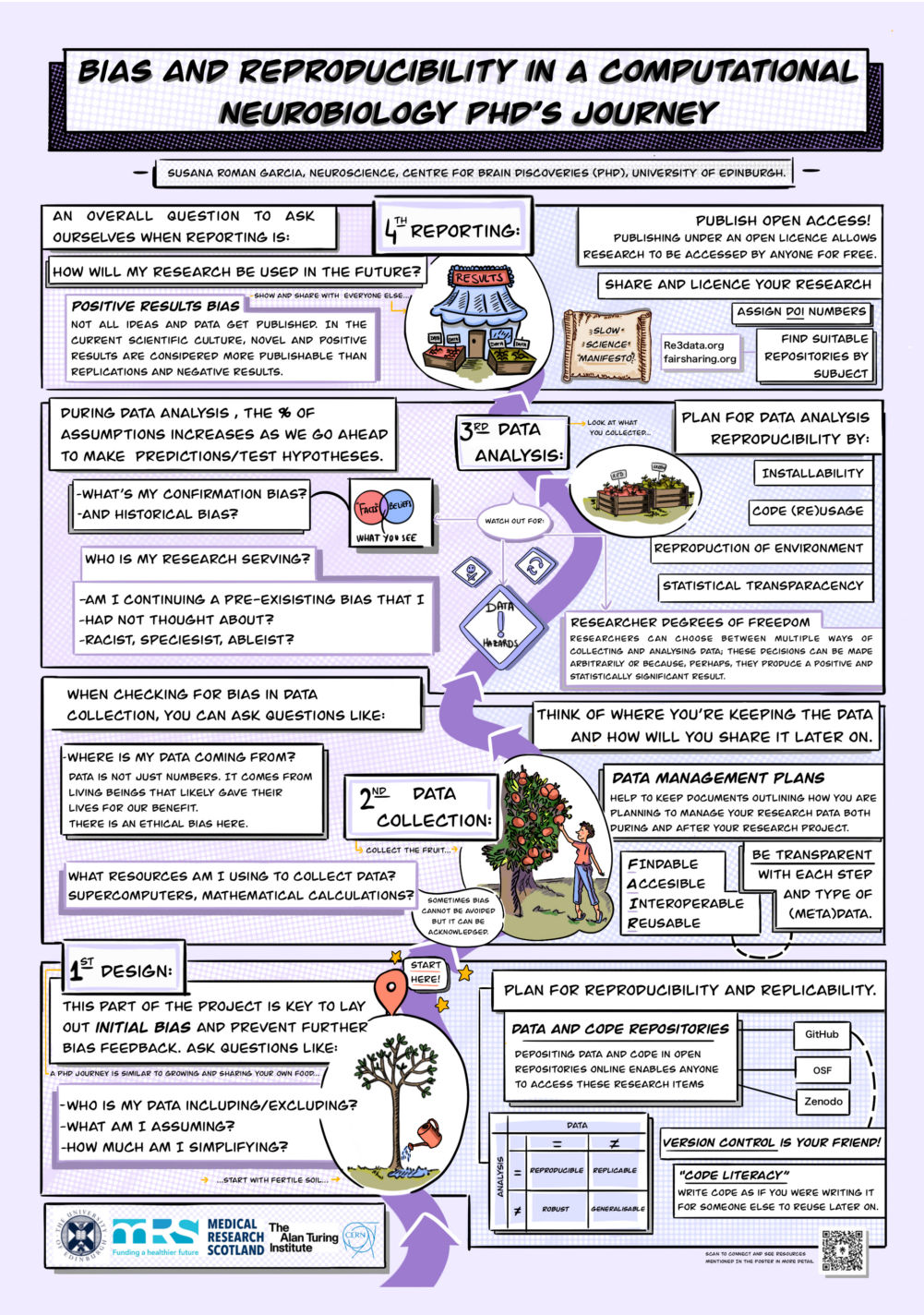

Figure 2.2: Poster about bias and reproducibility, showing research cycle as a journey which starts with design, then data collection, data analysis and final reporting, and compares this through images to growing an apple tree, collecting the apples and then selling them.

3 Data Hazards

The ethical implications that ought to be considered when doing research, usually go beyond what most ethics Institutional Review Boards propose; they should include questions about the wider societal impact of how data science and algorithms work. This is where a project like the Data Hazards Project comes in handy. Data Hazards is a project made to help us in thinking about worst-case scenarios and ways to mitigate these.

The Data Hazards Project has created a community-developed shared vocabulary of data science risks. The vocabulary presents data ethics concepts in the form of Data Hazard Labels, similarly to chemical hazard labels. This project exists to facilitate material for interdisciplinary discussions and self-reflection about all kinds of data ethics risks. How do these labels look like and how can they be implemeted? Let's go through some examples to show how.

3.1 Example label: high environmental cost

Figure 3.1: "High Environmental Impact" Data Hazard Label

3.1.1 Description

This hazard is appropriate where methodologies are energy-hungry, data-hungry (requiring more and more computation), or require special hardware that require rare materials.

3.1.2 Examples

- Example: Running computer models in super computers requires vast energy usage.

3.1.3 Safety Precautions

- Consider in what circumstances it is worthwhile to use this type of methodology.
  - To communicate the scale of the issue to other stakeholders, you may want to convert units of energy into more relatable units.
  - Find out if your cloud provider uses renewable energy.
  - Consider profiling your code, and rewriting it to use less energy.
- Consider future work that would reduce the need to use increasingly more resources.

3.2 How to use the Data Hazards Project

There are four steps to using the Data Hazard labels:

- **Learning**: familiarising yourself with the Data Hazard labels.
- **Applying**: deciding which Hazard labels are relevant to your project.
- **Reflecting**: on what to do differently and what mitigations to make.
- **Display**: displaying the labels alongside your work can help you to communicate that you've thought about these broad ethical issues and how you'd like others to use your work.

This spells LARD , which makes it pretty easy to remember! It is however an unfortunate word it shortens to, as lard comes from dead pigs, so I like to manifest it's a plant-based LARD .

As part of a Turing Way Book Dash hosted in May 2023, I worked together with a team to create a chapter on Data Hazards for the Turing Way Book. This chapter is still in draft form, as part of this experience we worked with an artist from Scriberia, to make an illustration of the Data Hazards application (Figure 3.2).

3.3 Application example into Research life-cycle

To help visualize where and when Data Hazards can be used in your workflow, below is an example assuming four main stages of workflow: design, data collection, data analysis and reporting. This is a generalised example, but something like this is what it looks like for me when I work on my PhD.

3.3.1 Design:

- Are you using data? Then doing some reflection on identity and positionality could help you think of what Data Hazards labels you might encounter as you design your project, for example "ranks of classifies people hazard" or "risk to privacy" could apply at this stage.

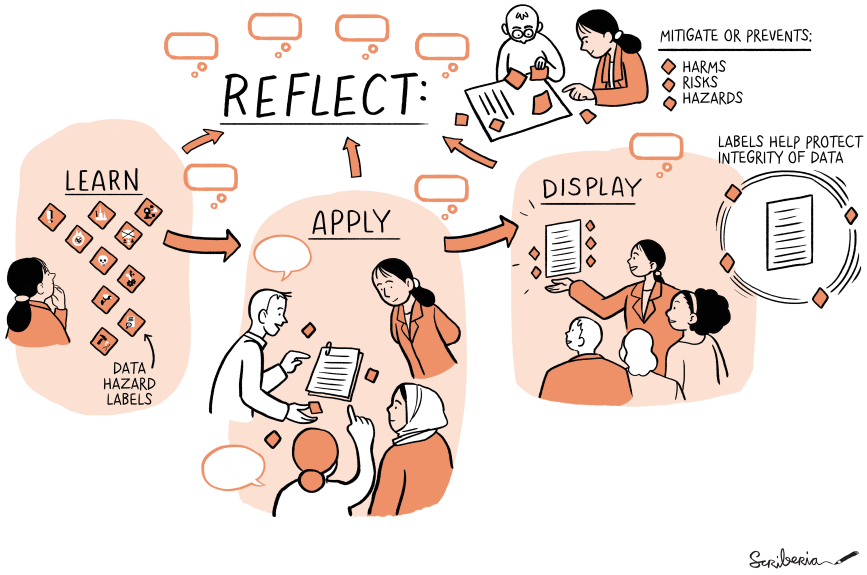

Figure 3.2: Data Hazards Application Cycle (Community and Scriberia 2023).

- In this part of the workflow, you might want to prepare to avoid certain Data Hazards if you can, and if you can't avoid them because of where your data has come from, you may want to acknowledge this. For example, if you a sensitive data project, what Data Hazard labels will apply, and/or what can you do to design your project in a way that avoids certain harms?

3.3.2 Data Collection and Analysis:

- As you are collecting and analyzing your data, you might want to iteratively think of the potential Data Hazards that exist in the information you are collecting. To then apply the labels as you perform the next step of the process: reporting.

3.3.3 Reporting:

- When reporting your results, you can think of applying and reporting the Data Hazard labels that are relevant for your project; examples of how I've done this can be found below in Table 3.1. Labeling your project with Data Hazards should also include considerations of mitigations to these risks. This would then be helpful for people who see your outputs in the future. They can be aware of potential risks as they proceed with the project, and continue to think of solutions to any issues related to the research topic.

3.4 Application into my PhD project: Presenting my PhD as a case study at AI UK conference

In order to showcase how Data Hazards can be reflected upon during a PhD, and taking the self-reflection described above into consideration, I have been implementing thinking about the vocabulary they provide into my own work. In line with this work, I made a poster that summarised aims of my PhD, for people to be able to say which labels they thought applied to my project. This poster was part of an exhibition stand with the Data Hazards Team, at AIUK 2023.When creating this poster (Figure 3.3), I was able to both do some self-reflection and collaborative reflection, as described below.

3.4.1 Self-reflection (*what is my project and how will it be used?*):

When making the poster, this kind of self-reflection questions are useful for oneself to think about, but also for external people who are not involved with your project to understand what potential data hazards it might have. The final poster can be seen below in (Figure 3.3). I followed the prompt questions available in the Data Hazards website for project owners who would like their projects to be discussed for data hazards:

– The overall objective of the project.

– Fairly detailed description of the variables in the dataset they are using (and what is not included).

– How and when the data was collected.

– Any statistical/algorithmic methods being used.

– Who has input on the project.

– What outputs are expected, and how these will be shared.

3.4.2 Collaborative reflection (*what data hazards may apply to my project?*):

During the poster presentation, people talked about the project, had a look at the poster, and decided by adding stickers to a list of hazards, to say which ones applied to it (Figure 3.4).

As can be seen in Figure 3.4 (before end of the day), people were adding stickers to record which data hazard labels they thought applied to my PhD project. At the end of the day, I recorded final numbers and the results can be seen in the barchart below Figure 3.5.

3.4.3 Reflections on Data Hazard Labels that apply to this PhD project:

Interestingly, not all labels were chosen as applicable to my project (Figure 3.5). Only 6 of the 11 current labels were chosen as relevant, with "difficult to understand" being the most prevalent one, chosen by 6 people. High environmental impact and danger of misuse follow in closely with 5 people having chosen these ones. Of course these numbers are small and hold, more than anything, illustrative value as to how and why people may think certain labels apply to a project. Difficult to understand" label was chosen the most, followed by "high environmental impact" and "danger of misuse". I will discuss the three most chosen labels and what mitigation strategies I am taking to make sure these risks are reduced.

- **Difficult to understand:**

This project includes niche topics, like knowledge about specific postsynaptic protein interactions, as well as specialised software to model such molecular interactions. This project is interdisciplinary and sits in between biology and computer science. This means that conveying varied, niche topics to different audiences, including a broad audience such as that of AIUK, requires a big effort to make the methods and results very clear to all. On the one hand, this project includes programming, which means that in order to make the models accessible and clear, code should be well documented and hopefully shared with appropriate licences. Likewise, there is a need for transparency as to how and why models were created the way they were, and publication of this data provenance is of upmost importance to mitigate the risk of "difficult to understand".

This is why I have spent the last year making the models as reproducible as possible to be able to build up from them in a way that allows other people to easily understand where things have come from in this project. Additionally, making sure biological interactions and relevance in this project are easily understandable and accessible is important as this should enable incorrect results to be more easily identifiable and that the models are more easily implemented by other researchers. The "difficult to understand" hazard is one that in order to be mitigated requires fine tuning and finding a balance between how the research is written and knowing that some of this work will inevitably require prior knowledge on some topics.

- **High environmental impact:**

The next most chosen label was "high environmental impact". The models created during my project have the potential to require a high degree of energy consumption, and therefore a potentially high environmental impact. As the models become too large to run in my own machine, running them in high performance computing clusters (HPCs) such as the University of Edinburgh's Eddie can mean I don't have to run the model in my own machine for 4+ hours, instead I can access the cluster and run it there. Using HPCs means a variety of environmental impacts: energy production, hardware manufacturing, long-term storage management, cooling, maintenance, and more. Calculating the exact environmental impact of the models I run can be difficult, as it is a challenge to find exact spefications of Eddie, and what are its energy sources. I could not find exact speficications of Eddie. However, I found that it uses "several thousand Intel Xeon cores [and a] significant number of NVIDIA GPUs" (see here), as well as the default storage space for research groups being 200GB.

With this restricted information, I used a free online calculator to do an estimation of carbon emissions of one model run Lannelongue, Grealey, and Inouye (2021).

The result estimate, with an assumption of 3500 CPUs Xeon E5-4620 and 3500 GPUs NVIDIA Titan V [1], based in the United Kingdom, is that a model which takes 4 hours to run would have a carbon footprint of 1.42 T CO2e, which is equivalent to a 8090km car jorney, or 61% of a flight trip from NYC to Melbourne, or 128.66 tree years [2].

These are just estimates, however they shed some light on the fact that this project, like many others, has an environmental impact that we may not have realised at first glance. Unless we stop and reflect, this kind od hazard may not have seemed as apparent as the loud computer tours with countless cables and noisy fans are tucked away in a data centre where we can more easily forget about these. This is why taking the time to write optimised code that runs faster with fewer resources can save both money and the planet. Likewise, thinking about what jobs *really* need to be run can also be a way to reduce waste during research.

- **Danger of misuse:**

I would say that this hazard label is so potentially broad that it interlinks with the two mentioned above. I would argue that research that is potentially difficult to understand, has the potential to therefore be misinterpreted. Statements of what and how this research can and should be applied should be made explicit and clear. Additionally, the fact that future models could be built from this one, means someone could make a model that has an even higher environmental impact. In discussing these, it does not mean they will necessarily happen, but if at least discussed, they could be prevented or circumvented. A potential danger of misuse that is even more likely, is *in vitro* animal testing. Animal testing brings the creatures being tested on stress and pain. Showcase examples of tests done about camkii and how language is used in ways that detaches us from the animals being abused. I have examples of this in the 2nd year report. clinical trials in people too, even.

[1] The specifications of the type of NVIDIA AND Xeon processors were assumed, as this was not publicly available information. This means that if an NVIDIA Jetson AGX Xacier was used in comparison to NVIDIA Titan, for example, there would be a drastic difference in their power draw. The former draws 30W, whereas the latter draws 250W. I have emailed the people who maintain Eddie to request this information.

[2] Tree years is the amount of CO2 sequestered by a tree in a year. The calculator uses it to measure how long it would take to a mature tree to absorb the CO2 emitted by an algorithm/model. The green-algorithm calculator use the value of 11 kg CO2/year.

Table 3.1: Three most chosen Data Hazard labels as applicable for my PhD project during collaborative reflections at AIUK.

Data Hazard Description	
	**Difficult to understand.** There is a danger that the technology is difficult to understand. This could be because of the technology itself is hard to interpret (e.g. neural nets), or problems with it's implementation (i.e. code is not provided, or not documented). Depending on the circumstances of its use, this could mean that incorrect results are hard to identify, or that the technology is inaccessible to people (difficult to implement or use).
	- Make research code Open Source with an appropriate software license where possible.
	- Ensure code is well documented with accompanying and/or inline documentation.
	- Ensure accessible descriptions of biological models to allow for future researchers to find potential errors and to be able to replicate results.

| **High environmental impact.** This hazard is appropriate where methodologies are energy-hungry, data-hungry (requiring more and more computation), or require special hardware that require rare materials. | - Write optimised code that runs faster with fewer resources. | | | | | - Think about what jobs really need to be run to help reduce waste. |

Danger of misuse. There is a danger of misusing the algorithm, technology, or data collected as part of this work. | - yet to add

3.5 Data Hazards Workshops

In order to showcase how to implement the Data Hazards, there is a template in the website which showcases a template on how to run workshops to learn about the project. I organised and facilitated two Data Hazards workshops during my third PhD year:

3.5.1 Workshop at COMBINE conference (Berlin, October 2022)

The COMBINE (Computational Modeling in Biology) conference was an in person event, where I proposed, ran and co-facilitated a Data Hazards workshop with my supervisor Melanie Stefan. Melanie presented a project study for the participants to then think what potential Data Hazard labels applied to it. We had 12 participants, who discussed and labelled the project during the workshop.

Topics of interest at COMBINE included discussion of data exchange, pipelines and discussing standardizing methods for computer modelling of systems biology. Data Hazards have the potential to become a standard practice for modelling systems biology research. This session was a good opportunity to discuss how to make Data Hazards assessments a standard part of the information shared with computational models. The materials used for this workshop can be found in this GitHub repository.

The workshops I have ran have had the following types of roles:

Facilitators:

– There to run the workshop and help everyone get the most out of it.

– This involves managing each of the breakout rooms and supporting the discussions.

Project Owners:

– There to have their project discussed by the audience members.

– They are seeking feedback, with a focus on Data Hazards, on an idea or project.

Audience Members:

– There to find out more about the projects and provide feedback.

– Combination of different types of people.

– Can be 'experts' on topic being presented by project owners or not (both can be interesting!).

3.5.2 Data Hazards, Ethics and Reproducibility Symposium (London, March 2023)

As another way to implement the Data Hazards framework, we decided to host a symposium that revolved around this topic. Together with Ceilidh Welsh, we co-organised a hybrid, one day symposium at the Alan Turing Institute (ATI) HQ in London: Data Hazards, Ethics and Reproducibility Symposium. This was possible thanks to the Enrichment scheme we were part of, and thanks to the grassroots funding we were granted by the ATI.

The event encouraged attendees to explore, discuss and reflect on the ethical implications and wider societal impact of specific data-intensive projects. It was an opportunity for attendees to appreciate that ethics is complex, situational and important to discuss in our own contexts. It aimed to shine a light on an event that promotes data ethics through the content of the event, and also its planning and delivery.

We are in the process of creating a chapter in the Turing Way Book to publish our experience organising an accesible event which aimed to discuss how people in different stages of their careers can embed thinking about ethics, reproducibility and data hazards as they go, not as an add-on at the end.

Below you can find a table with the day's programme (Table 3.2), and here you can find a draft for of the behind the scenes of how we organised and our reflections on making this event happen.

Table 3.2: DER symposium program for the day

Time (GMT)	Topic
10:00 - 10:15	Welcome and Introduction to the day
10:15 - 11:05	**Keynote Speakers:** • Anne Lee Steele - Talk title: `From culture to computation: mapping my open research journey.` • Paz Bernaldo - Talk title: `Am I in or am I out? Investigating who is in, in open science.`
11:05 - 11:15	Break
11:15 - 13:00	**Data Hazards Workshop** - Materials here
13:00 - 14:00	Lunch Break - Yellow Kitchen Catering

Time (GMT)	Topic
14:00 - 14:30	Networking discussion: why do you care about data ethics? (online only).
14:30 - 15:00	**Reproducibility in PhDs** • Ezra Herman - Talk Title: `A reproducible thesis - writing code and reports in one go with Snakemake and R Markdown`. • Natalie Zelenka - Talk title: `How I tricked myself into writing my thesis (by making it as ethical and reproducible as I could).`
15:00 - 15:15	Break
15:15 - 16:00	**Embedding Ethics and Reproducibility into your Research Career** • Alden Conner - Talk Title: `The Turing Way: A collaborative guide to data science and research`. • Melanie Stefan - Talk title: `The ethical lecture: looking at university teaching through a Data Hazards frame.` • Clau Fischer - Talk title: The Turing Commons, Training in AI ethics and responsible research.
16:00 - 16:15	Facilitated Discussion: Embedding ethics into your research projects - key takeaways
16:15 - 16:30	Closing remarks

All of these examples showcase how the Data Hazards framework can be dicussed, used and applied at different levels. At my own PhD project level, I have used this framework to think about dangers of my own data. When discussing with other researchers, it has sparked conversations about their own potential data hazards.

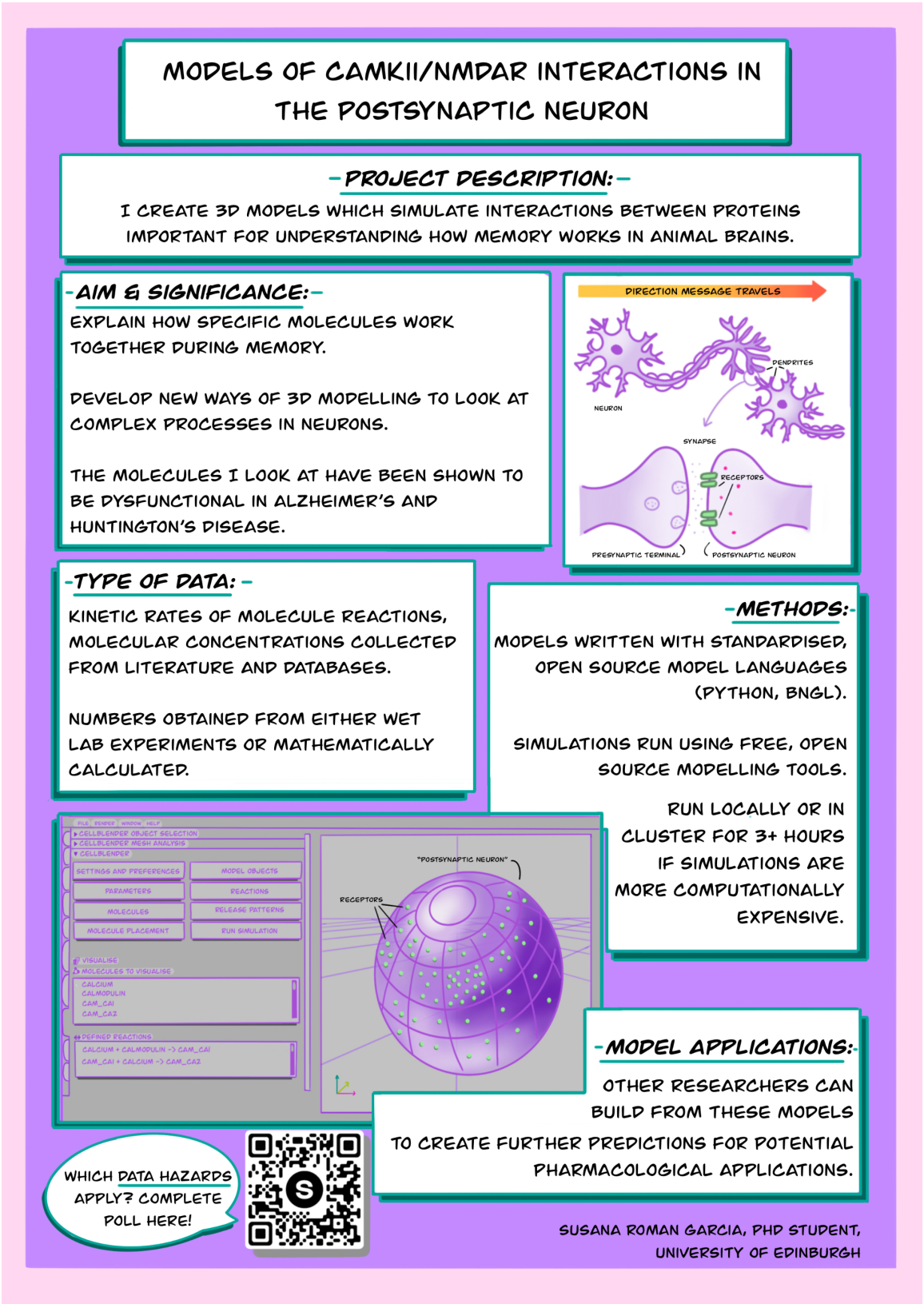

Figure 3.3: PhD Project decription - Case Study, to see GitHub repo, click on this figure.

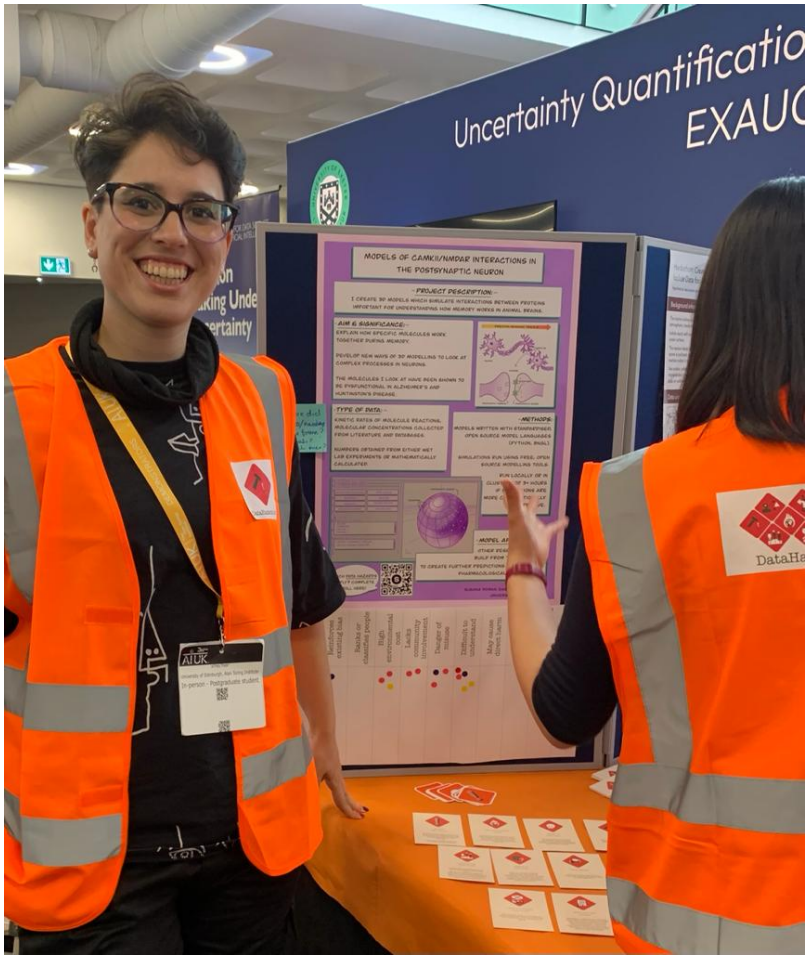

Figure 3.4: Data Hazards Case Study Poster at AI UK

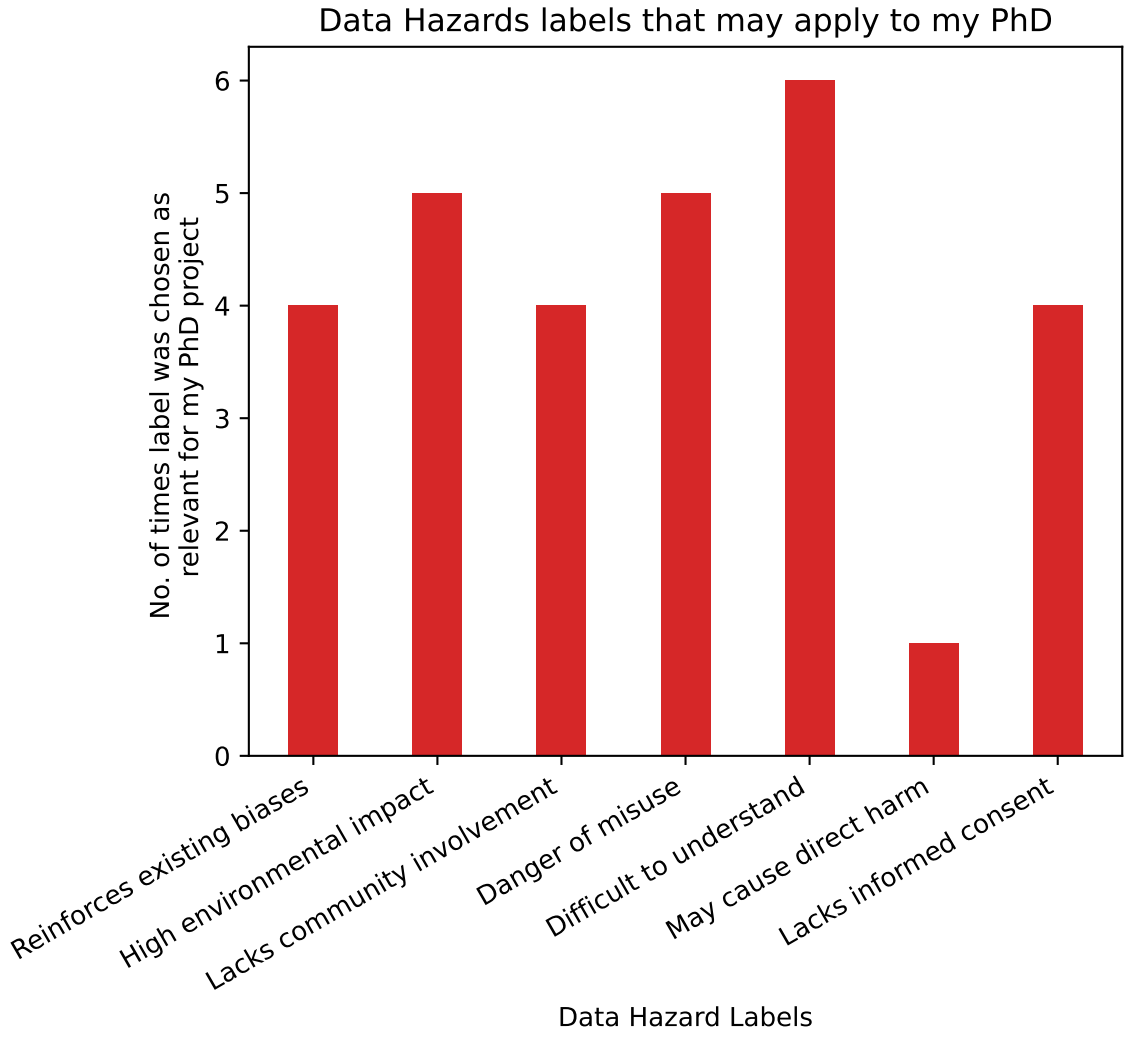

Figure 3.5: Data Hazards labels that may apply to my PhD.

4 Computer models of CaMKII/NMDAR interactions

As mentioned in the introduction Chapter 1 of this report, I use computer models to study interactions between specific postsynaptic molecules; below I describe what these models look like and what software I use. Likewise I also explain how I have been working on creating a reproducible model and the validation steps taken to achieve this.

4.1 BioNetGen and how rule-based modelling can help with combinatorial complexity.

BioNetGen is a set of software tools which facilitate a rule-based approach to modelling biochemical reaction kinetics, where we can largely overcome the problem of combinatorial complexity that arises when modelling CaMKII. It has been calculated that CaMKII as a dodecamer can approximately have 10^{20} possible states (Pharris et al. 2019); this, together with the potential of a full reaction network for each simulation (an added factor of combinatorial complexity), can render the model computationally intractable. BioNetGen can help us deal with this combinatorial complexity thanks to its rule-based modelling (RBM) "don't care, don't write" capabilities. And as we will see later MCell can help with modelling network-free simulations.

BioNetGen language (BNGL) is a formal language which uses the BioNetGen software (Faeder, Blinov, and Hlavacek 2009). It allows for site-specific details of protein-protein interactions to be captured in models for the dynamics of these interactions in a systematic fashion, which also alleviates nomenclature and reusability issues.

Hence, using this RBM approach is notable as it facilitates writing of multi-state modelling and can significantly, reduce the number of reactions that need to be written due to its "don't write, don't care" characteristic. Thereby dramatically improving the ability to model CaMKII as a dodecamer; I can make a model with multistate molecules, and specify the states of the reactants that are relevant for a particular reaction, and leave the rest unspecified (see Figure 4.1)

To interact with this code, you can have a look and download a jupyter notebook I have created here, where I also describe some of the ways in which the model can be simulated, with stochastic simulation algorithms (SSAs) or ordinary differential equations (ODEs). See screenshot from notebook in Figure 4.2 below.

```
begin model
begin parameters
# Define initial number of molecules released
  A_i 150
  B_i 150
  C_i 100
#Define reaction rates
  kon 1e-2
  koff 1e-3
  k_P 1e1
end parameters
begin molecule types
# Here we define the molecules and the possible states and binding sites
they can have
# Molecule A has a binding site (a), and a Phosphorylation site which can
be unphosphorylated (~0) or phosphorylated (~P):
    A(a,T286~0~P)
# Molecule B has a binding site (b):
    B(b)
# Molecule C has no binding sites:
    C()
end molecule types
begin species
# Molecule A starts with binding site a free, and with phosphorylation site
unphosphorylated
    A(a,T286~0) A_i
# Molecule B starts with binding site b free
    B(b) B_i
# Molecule C has no binding sites so it starts as it is
    C() C_i
end species
begin reaction rules
# A_free and B_free can resversbly bind to give AB_complex
# Don't need to specify, if I'm not interested, status of phosphorylation
for molecule A. Note how it is not written in the rule definition (don't
care, don't write):
    A(a) + B(b) <-> A(a!1).B(b!1) kon, koff
# If A is unphosphorylated, it can become phoshorylated by the presence of
C
# Don't need to specify status of binding site 'a' (don't care, don't
write):
    A(T286~0) + C() -> A(T286~P) k_P
end reaction rules
begin observables
  Molecules AB_complex A(a!1).B(b!1)
  Molecules A_phosphorylated A(T286~P)
  Molecules A A(a)
  Molecules B B(b)
  Molecules C C()
end observables
end model
simulate({method=>"ssa",t_end=>10,n_steps=>100})
```

(3) Run ODE model and plot results

- Now let's run the ODE model.
- Notice the difference?

```
# change file name here (if running your own code):
file_to_run = "ode_simple_dcdw.bngl"


# don't change anything after this

# this command runs the model in a temporary folder which is removed after execution is done
r = pybng.run(file_to_run, suppress=True)[0]

print(r.dtype.names) # this will print the names of the observables
# now we can loop over each observable name and plot them
for name in r.dtype.names:
    # we don't want to plot time
    if name != "time":
        # plot the observable values over time
        plt.plot(r['time'], r[name], label=name)
# label everything
plt.xlabel("time")
plt.ylabel("counts")
_ = plt.legend(frameon=False)
```

```
('time', 'AB_complex', 'A_phosphorylated', 'A', 'B', 'C')
```

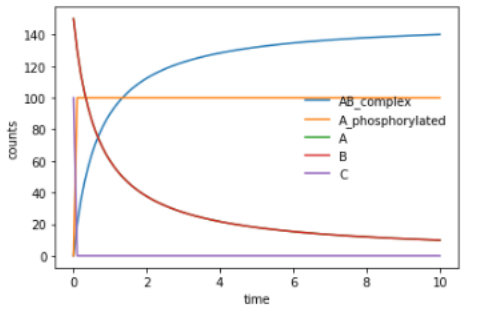

Figure 4.2: What an ODE output from the model above looks like, shown as output is created in jupyter notebook.

4.1.1 MCell (Monte Carlo Cell) and how it simulates reactions in 3D

MCell is a biochemistry simulation tool that uses spatially realistic 3D cellular models and stochastic Monte Carlo algorithms to simulate the movements and interactions of discrete molecules within and between cells, (Bartol and Stiles 2000), (Kerr et al. 2008), (Bartol et al. 2015). MCell is a particle-based simulator that represents molecules as point particles in 3D space. At every time step in an MCell simulation, each particle can move, collide with other particles or surfaces, and undergo bimolecular and unimolecular reactions. The basic elements of a simulation step are as seen in Figure 4.3 taken from Gupta et al. (2018).

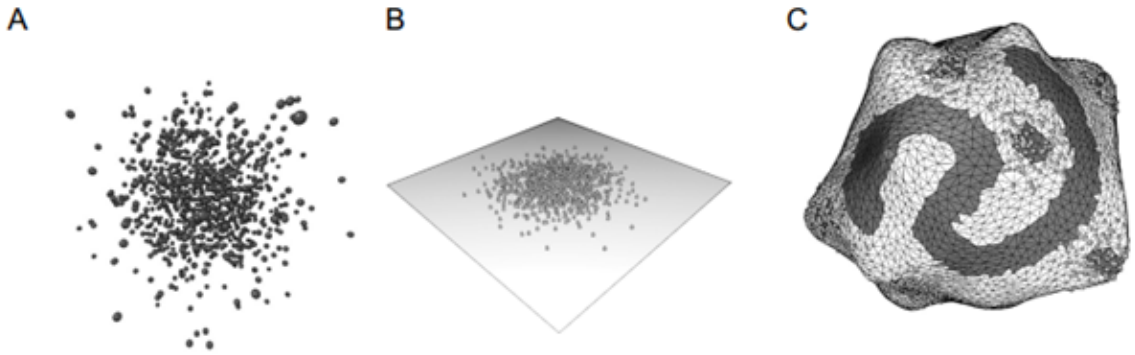

Figure 4.3: MCell Components. (A) Volume Molecules diffusing in free space. (B) Mesh Object defined by a Plane with Surface Molecules diffusing on it. (C) Mesh Object defined by a complex closed mesh with multiple defined Surface Regions, in which Surface Molecules have different diffusion constants, as defined by corresponding Surface Classes.

Briefly, MCell operates as follows: as a volume molecule diffuses, all molecules within a given radius along its trajectory, or at the point of collision on a surface, are considered for a reaction. For surface molecules (in membranes), the molecule first diffuses, and then its neighbours are evaluated for reaction.

There is no volume exclusion for molecules diffusing in 3D volumes, and molecules on surfaces occupy a fixed area. MCell allows defining arbitrary geometry Figure 4.3 (C), and complex models such as a 180 m3 3DEM reconstruction of hippocampal neuropil have been used to construct a geometrically-precise simulation of 100s of neuronal synapses at once (Bartol et al. 2015). A detailed description of mathematical foundations of MCell's algorithms can be found here: Bartol and Stiles (2000), Kerr et al. (2008), Bartol et al. (2015).

MCell4, version used for this project, provides a versatile Python interface, which is very useful for writing models with said interface and running mcell models this way. MCell4 provides two different user experiences, one through its visual interface as an add-on in Blender 2.93, known as CellBlender (see back at Figure 1.1), the other user experience one through a new Python interface. This provides users with the flexibility to change between both experiences, or to run the simulations using Python and visualize the simulations in Blender (Figure 4.4).

Significantly, in MCell4 the reaction language is BNGL; making MCell4 fully support rule-based reactions and allowing all models to use this feature. The support for BNGL and network-free simulations of MCell4 allows direct, agent-based evaluation of reaction rules and thus enables spatially-resolved network-free simulations of interactions between and among volume and surface molecules. The CaMKII models I developed during this PhD would not be possible without the spatial network-free simulation allowed by MCell4.

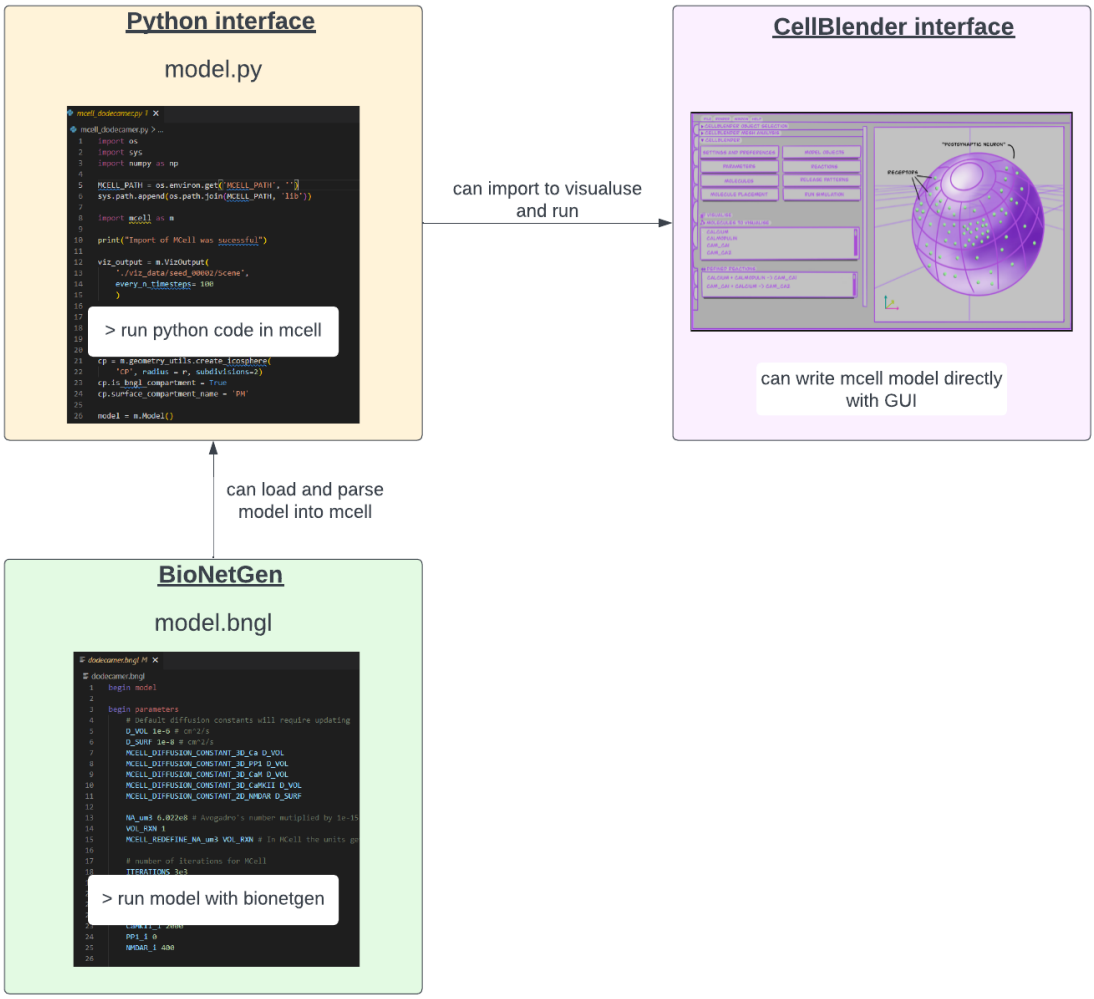

Figure 4.4: Diagram of what workflow of this project can look like, it is not exhaustive of all the ways in which these software can interact. Diagram was made with the softare Lucid.

4.2 Model description

I have constructed the models at different scales to validate CaMKII interactions with other molecules like calmodulin and NMDARs, at increasing levels of complexity. First I

re-created a model of CaMKII as a monomer that was previously completed in 2017. The model created uses cBNGL and represents CaMKII as monomers to serve as a proof of concept as well as a starting validation point, as dynamics of this model were previously shown to be within biologically accurate limits. Secondly, I created a model of CaMKII as a hexamer since modelling this molecule as a dodecamer gave rise to a combinatorial explosion due to the high number of possible states and the network of interactions generated. This was them resolved as I run the model using the network-free simulation capabalities using MCell. This has then resulted in being able to create a (still in the workd) model of CaMKII as a dodecamer. These simulations include only calcium binding to CaM, and CaM binding to CaMKII as a dodecamer, without further reactions added to avoid further complexity. Finally, I aim to validate this work against a model from Ordyan et al., 2020, where they successfully modelled CaMKII as a twelve subunit holoenzyme using BioNetGen simulations.

4.2.1 Model development and validation: CaMKII modelled as a dodecamer

CaMKII is a dodecameric molecule, meaning it's composed of twelve subunits. In the past I have modelled CaMKII as a monomer and as a hexamer, building up the models and validating with existing published research. Now, modelling it as a dodecamer allows us to infer more accurately any emergent behaviour of this protein.

In order to replicate and validate the results obtained from the CaMKII monomer in 2017, I have been re-writing all reactions into a cBNGL model using the same parameters used originally, then building CaMKII as a dodecamer instead of a monomer. I define a volume previously modelled of 0.5 m3, which is within ranges of spine volume of 0.004 to 0.6 m3 in hippocampal CA1 neurons (Harris and Stevens, 1989). Using cBNGL, I can specify the above mentioned 3D volume, with a 2D surface compartment acting as the cell membrane, where NMDARs can diffuse.

As well as modelling the multimeric properties of our molecules of interest, an aim of this project has been to further refine the spatial computational model to check if the prediction of the distribution of phosphorylated CaMKII is still valid, as described in 2017 model, and if so incorporate existing biological data. Previously, the head of dendritic spine was modelled as the cell, with a PSD shown in blue where there was a higher number of NMDARs (Figure 4.5). The prediction made was that if CaMKII binding to NMDA receptors is disrupted, we get more CaMKII activation overall, this was a counterintuitive result because NMDA receptor binding is known to activate CaMKII, so their disruption was assumed to have decreased CaMKII activation. It will be interesting to see if these results replicate with CaMKII as a hexamer, or if not then it might show new insights as to how CaMKII and NMDAR abolition results in the new models.

Parameters used in this project so far requires further analysis, as well as additional investigation of specific reaction parameters. As mentioned in the introduction section, finding parameters, if any, for computer models is a challenging as they tend to vary from publication to publication. For now, I used the validated parameters used for my thesis in 2017 (table 1). I hope to run sensitivity analysis to determine how much parameters used affect model behaviour.

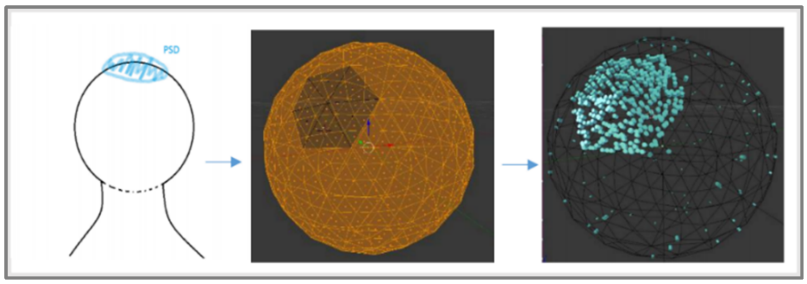

Figure 4.5: Head of dendritic spine modelled as the cell, with PSD shown in blue, modelled in MCell/CellBlender for the 2017 dissertation project.

An important part for the last part of this project has been to validate CaMKII holoenzyme activity against the model published by Ordyan et al. (2020). As well as an extended model published in 2022 (Husar et al. 2022) where they observe the effects of the geometry of the compartment on the simulation results thanks to MCell4. Their work adopts a RBM approach through the Monte Carlo method to study the effect of Ca2+ signals on the dynamics of CaMKII phosphorylation in the postsynaptic density. Their study looks at calcium surges in synaptic spines during an EPSP and back-propagating action potential due to the opening of NMDA receptors and voltage dependent calcium channels. Using agent-based models, they computationally investigate the dynamics of phosphorylation of CaMKII monomers and dodecameric holoenzymes. This model is very useful to have as a validation point, firstly to see if my results compare to what they observed, and hence enabling reproducibility of results. It is also helpful to have a model which looks at CaMKII but focuses on its interactions with other molecules, as this adds further to the narrative of how CaMKII functions.

One of the main questions of the research is to look at the dynamics of interactions between NMDARs and CaMKII in the postsynaptic neuron. The models created will state the predictions which will then enable us to generate hypothesis for system dynamics behaviour. It would not make sense to start asking questions about this if the basis of the model is not coherent and cannot replicate baseline molecular behaviour. Therefore, it is very important that validation of the model build up is carried out meticulously and reasoning as to why each step and decision was made is explained.

need to potentially make: Biologist friendly description of the model.

4.2.2 A reproducible model

With the idea of making reproducible research, it is good to build up step by step a model, and the same processes used in software development can also be applied to biological model development. Therefore, when developing the models in this project, four main points were considered throughout, as suggested by Husar et al. (2022):

1. Create incremental development where the model is built step by step, relying on solid foundations of modelling done and validated before,

2. Create a modularity that provides the capability to create self-contained, reusable libraries,
3. Perform unit testing and validation to verify that parts of the model behave as expected and,
4. Create human-readable and writable model code that can be stored using git or other code version control software which also allows code reviews so that other team members can inspect the latest changes to the model.

With this and robustness, replicability, reproducibility and generalisability (as defined in Chapter 2) in mind, I managed to make a model which was successfully reproduced in other people's machines. This means that the results of the model (Figure 4.6), where able to be reproduced by someone else using the same dataset which consistently produced the same answer.

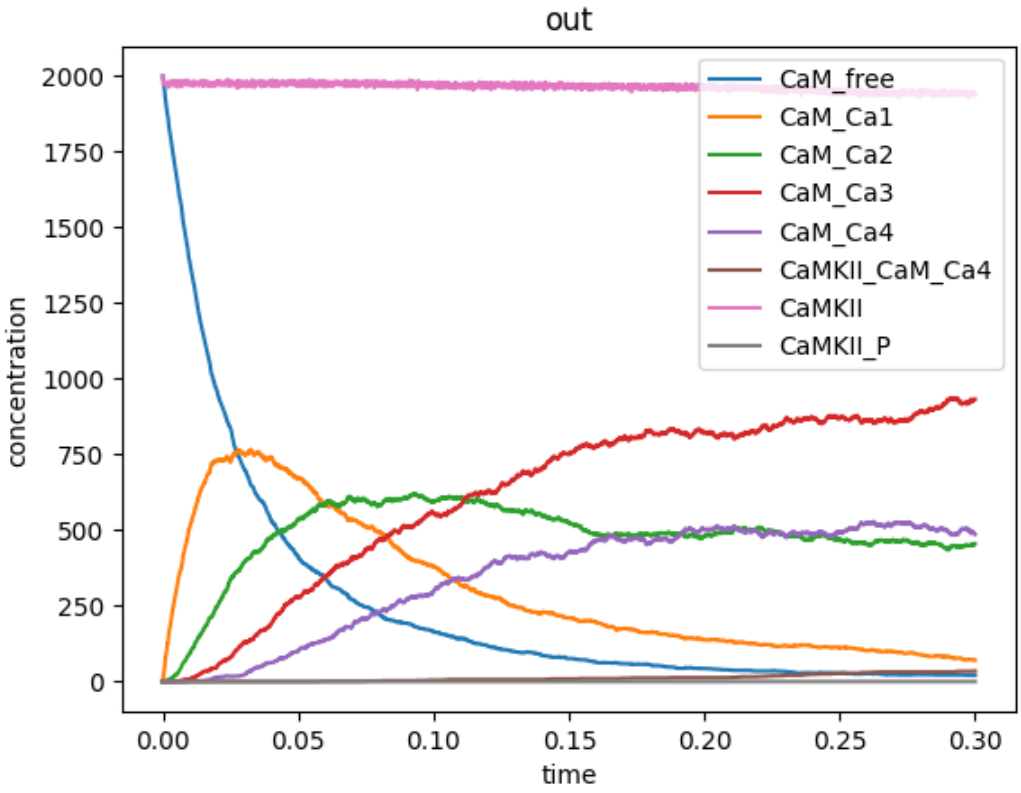

Figure 4.6: Sequence of modelled molecules interacting in a model of CaMKII as a hexamer (dodecamer should also reproduce as it's built off of this)

This was possible thanks to the efforts put into creating detailed step by step instructions on how to implement and run the model. These instructions currently live in a private repository in GitHub, which hopefully will evolve into something publishable. This repository includes a README.md file that allows you to know how to:

1. Install MCell/CellBlender v4.0.6 bundle with Blender 2.93.
2. Set up MCell4 Python in your machine and why this is necessary.
3. Set up a python environment with any potential dependencies required to run the code.
4. Run the code and what output is expected for comparison.

4.3 Advantages and limitations

This work has, therefore, this PhD project is eqquipped with strong advantages for future, trustworthy research done by other people, as well as for myself as I am able to keep cleaner, more neat track of what exactly I am developing. Some of these advantages include:

- **Ensuring continuity of work.** By following guidelines for reproducibility, I can easily communicate work with different stakeholders whether that's my supervisors, future funders, reviewers, students, and potential collaborators. This aspect of reproducibility increases the usefulness of this research by allowing others to easily build on our results, and re-use the research materials I created. This ensures the continuity of a research idea and can even, hopefully, find fresh applications in other contexts.
- **Being able to track a complete history of my research.** By storing a complete track-record of my work, I can ensure research sustainability, fair citation/acknowledgement, and usefulness of my and others' work in our research fields.
- **Facilitation of collaboration and review process.** Reproducible workflows have facilitated the peer review process tremendously by allowing other people not directly involved with the project to have access to the different parts of the projects that are necessary to validate the research outcomes.
- **Writing papers, thesis and reports efficiently.** Putting an effort in maintaining documentation of methods and analyses helps in maintaining easy access to all the results generated in the project that can then be written up efficiently. Additionally, by making available the underlying dataset and methods makes it more easy to comply with the highest-level journal guidelines.
- **Get credits for work fairly and more often.** Applying reproducibility practices separately on different parts of the project such as the data, code and scripts, and reports will enable other researchers to test and reuse my work in their research and therefore also allow for more fair recognition. Researchers who publish their work with the underlying information, get cited more often as their research outcome can be broadly replicated and trusted (NEED REFERENCE).

It has been challenging (and fulfilling) to learn and apply reproducibility practices. Learning to create reproducible research that also thinks about its ethical connotations takes time. This has meant that most of my third PhD year has been spent interacting with different communities (see chapter ON ACTIVITIES), creating contacts that can inspire and help me in creating quality research, and also getting hands on experience on creating bits of reproducible, accessible and open research - as exemplified throyghout this report. This also means that I have had to balance my time between creating reproducible models, as

well as the models having a significant biological outcome that is valid and can help further understand the mechanisms of how CaMKII and NMDARs interact in the postsynaptic neuron during memory. Now that this project has strong pillars to build up-on, I hope to be able to focus more time on looking at the biological aspects of these models.

5 Activities involved with 2022/23

In order to give an overview of what activities I have been involved during my 3rd year, please see a summary below.

I should make it more descriptive but haven't had the time to do so yet.

- **Presentations and workshops**
  - Presentation given: Thinking about Ethics in (Computer) Science. CMVM Good Science showcase, University of Edinburgh, August 2022.
  - Poster and lightning presentation: Bias and reproducibility in a computational neurobiology PhD's journey, at ICSB Conference, October 2022.
  - Co-facilitated a Data Hazards Workshop for COMBINE conference, October 2022.
  - Co-facilitated Open Science & Reproducibility: The Turing Way Workshop, February 2023.
- **Extracurricular activities**
  - AIUK Volunteer, March 2023.
  - Enrichment Community Champion, October 2022 - October 2023.
  - The Turing Way Book Dash participation, May 2023.
- **Awards and training**
  - Enrichment Community and Placement Award, October 2022 - October 2023.
  - Co-facilitated a Data Hazards Workshop for COMBINE conference, October 2022.
  - Participated in Open Life Sciences Cohort, under project "Ethical standards and reproducibility of computer models in Neurobiology", September 2022 - January 2023.
  - Turing Commons: AI Ethics and Governance course, November 2022.
  - Research Software Engineering with Python course, November 2022.
  - Awarded £2000 grant to organize a Data Hazards, Ethics and Reproducibility one-day Symposium, March 2023.
- **Publications**
  - **OSF preprint: Data Hazards v1.0: an open-source vocabulary of ethical hazards for data-intensive projects.** *Authors:* Natalie Zelenka, Nina Di Cara, Euan Bennet, Vanessa Aisyahsari Hanschke, Emma Kuwertz, Ismael Kherroubi Garcia, Susana Roman Garcia.

6 Thesis layout

1. BACKGROUND
    - Research question
    - Report style and philosophy
    - Ethics and Reproducibility
        - Science for the profit of whom?
        - Importance of reproducibility
        - Ethics and reproducibility go together
    - Data Hazards
    - Memory and Learning
        - Why study CaMKII and NMDAR interactions to study memory formation?
        - A brief history background on learning and memory research
        - Long Term Potentiation
        - NMDA receptors structure and functions
        - CaMKII structure and functions
        - Bringing it all together: LTP, CaMKII/NMDAR complex as a molecular memory and interactions within the postsynaptic neuron
    - Why use computational modelling to study biological systems?
        - How do we model biochemical systems networks?
        - Rule Based Modelling
        - BioNetGen
        - MCell
        - Biodynamo
2. METHODS
    - Model Description
    - Model development and validation
        - A reproducible model
3. RESULTS

References

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533 (7604, 7604): 452–54. https://doi.org/10.1038/533452a.

Bartol, Thomas M., Daniel X. Keller, Justin P. Kinney, Chandrajit L. Bajaj, Kristen M. Harris, Terrence J. Sejnowski, and Mary B. Kennedy. 2015. "Computational Reconstitution of Spine Calcium Transients from Individual Proteins." *Frontiers in Synaptic Neuroscience* 7. https://www.frontiersin.org/articles/10.3389/fnsyn.2015.00017.

Bartol, Thomas M., and Joel R Stiles. 2000. *Monte Carlo Methods for Simulating Realistic Synaptic Microphysiology Using MCell*. Vol. chapter 4. CRC Press. https://books.google.com?id=8TLpBwAAQBAJ.

Blundon, Jay A., and Stanislav S. Zakharenko. 2008. "Dissecting the Components of Long-Term Potentiation." *Neuroscientist* 14 (6): 598–608. https://doi.org/10.1177/1073858408320643.

Branch, Haley A., Amanda N. Klingler, Kelsey J. R. P. Byers, Aaron Panofsky, and Danielle Peers. 2022. "Discussions of the 'Not So Fit': How Ableism Limits Diverse Thought and Investigative Potential in Evolutionary Biology." *The American Naturalist* 200 (1): 101–13. https://doi.org/10.1086/720003.

Claerbout, Jon F., and Martin Karrenbach. 1992. "Electronic Documents Give Reproducible Research a New Meaning." In *SEG Technical Program Expanded Abstracts 1992*, 601–4. SEG Technical Program Expanded Abstracts. Society of Exploration Geophysicists. https://doi.org/10.1190/1.1822162.

Community, The Turing Way, and Scriberia. 2023. *Illustrations from The Turing Way: Shared Under CC-BY 4.0 for Reuse*. Zenodo. https://doi.org/10.5281/zenodo.8169292.

Delgado, Nick. 2022. "Owning Your Privilege: Leaving Guilt, Shame, and Blame Behind." Integrated Work. February 15, 2022. https://integratedwork.com/jedi/owning-your-privilege/.

DiAngelo, Dr Robin. 2018. *White Fragility: Why It's So Hard for White People to Talk About Racism*. Beacon Press. https://books.google.com?id=abZdDwAAQBAJ.

Diogo, Rui, Adeyemi Adesomo, Kimberly S. Farmer, Rachel J. Kim, and Fatimah Jackson. 2023. "Not Just in the Past: Racist and Sexist Biases Still Permeate Biology, Anthropology, Medicine, and Education." *Evolutionary Anthropology: Issues, News, and Reviews* 32 (2): 67–82. https://doi.org/10.1002/evan.21978.

Dosi, Giovanni, Luigi Marengo, Jacopo Staccioli, and Maria Enrica Virgillito. 2023. "Big Pharma and Monopoly Capitalism: A Long-Term View." *Structural Change and Economic Dynamics* 65 (June): 15–35. https://doi.org/10.1016/j.strueco.2023.01.004.

Faeder, James R., Michael L. Blinov, and William S. Hlavacek. 2009. "Rule-Based Modeling of Biochemical Systems with BioNetGen." *Methods Mol Biol* 500: 113–67. https://doi.org/10.1007/978-1-59745-525-1_5.

Fennen, Lisa. 2021. *Warp & Weft; Psycho-Emotional Health, Politics and Experiences.*

https://lisafannen.bandcamp.com/album/warp-weft.

Fink, Charles C., and Tobias Meyer. 2002. "Molecular Mechanisms of CaMKII Activation in Neuronal Plasticity." *Curr Opin Neurobiol* 12 (3): 293–99. https://doi.org/10.1016/s0959-4388(02)00327-6.

Garcia, Susana Roman, David Sterratt, and Melanie Stefan. 2022. "Thinking about Ethics in (Computer) Science." University of Edinburgh, Edinburgh, August 8. https://doi.org/10.5281/zenodo.6973796.

Ghosh, Anshua, and Karl Peter Giese. 2015. "Calcium/Calmodulin-Dependent Kinase II and Alzheimer's Disease." *Molecular Brain* 8 (1): 78. https://doi.org/10.1186/s13041-015-0166-2.

Gupta, Sanjana, Jacob Czech, Robert Kuczewski, Thomas M. Bartol, Terrence J. Sejnowski, Robin E. C. Lee, and James R. Faeder. 2018. "Spatial Stochastic Modeling with MCell and CellBlender." September 30, 2018. https://doi.org/10.48550/arXiv.1810.00499.

Husar, Adam, Mariam Ordyan, Guadalupe C. Garcia, Joel G. Yancey, Ali S. Saglam, James R. Faeder, Thomas M. Bartol, and Terrence J. Sejnowski. 2022. "MCell4 with BioNetGen: A Monte Carlo Simulator of Rule-Based Reaction-Diffusion Systems with Python Interface." May 19, 2022. https://doi.org/10.1101/2022.05.17.492333.

Ivie, Peter, and Douglas Thain. 2018. "Reproducibility in Scientific Computing." *ACM Comput. Surv.* 51 (3): 63:1–36. https://doi.org/10.1145/3186266.

Kerr, Rex A., Thomas M. Bartol, Boris Kaminsky, Markus Dittrich, Jen-Chien Jack Chang, Scott B. Baden, Terrence J. Sejnowski, and Joel R. Stiles. 2008. "FAST MONTE CARLO SIMULATION METHODS FOR BIOLOGICAL REACTION-DIFFUSION SYSTEMS IN SOLUTION AND ON SURFACES." *SIAM J Sci Comput* 30 (6): 3126. https://doi.org/10.1137/070692017.

Lannelongue, Loïc, Jason Grealey, and Michael Inouye. 2021. "Green Algorithms: Quantifying the Carbon Footprint of Computation." *Advanced Science* 8 (12): 2100707. https://doi.org/10.1002/advs.202100707.

Ordyan, Mariam, Tom Bartol, Mary Kennedy, Padmini Rangamani, and Terrence Sejnowski. 2020. "Interactions Between Calmodulin and Neurogranin Govern the Dynamics of CaMKII as a Leaky Integrator." *PLOS Computational Biology* 16 (7): e1008015. https://doi.org/10.1371/journal.pcbi.1008015.

Pharris, Matthew C., Neal M. Patel, Tyler G. VanDyk, Thomas M. Bartol, Terrence J. Sejnowski, Mary B. Kennedy, Melanie I. Stefan, and Tamara L. Kinzer-Ursem. 2019. "A Multi-State Model of the CaMKII Dodecamer Suggests a Role for Calmodulin in Maintenance of Autophosphorylation." *PLOS Computational Biology* 15 (12): e1006941. https://doi.org/10.1371/journal.pcbi.1006941.

Plesser, Hans E. 2018. "Reproducibility Vs. Replicability: A Brief History of a Confused Terminology." *Frontiers in Neuroinformatics* 11. https://www.frontiersin.org/articles/10.3389/fninf.2017.00076.

Robison, A. J. 2014. "Emerging Role of CaMKII in Neuropsychiatric Disease." *Trends in Neurosciences* 37 (11): 653–62. https://doi.org/10.1016/j.tins.2014.07.001.

Stengers, Isabelle. 2018. *Another Science Is Possible: A Manifesto for Slow Science*. John Wiley & Sons. https://books.google.com?id=oxJSDwAAQBAJ.

Tiwari, Krishna, Sarubini Kananathan, Matthew G. Roberts, Johannes P. Meyer, Mohammad Umer Sharif Shohan, Ashley Xavier, Matthieu Maire, et al. 2021. "Reproducibility in Systems Biology Modelling." *Mol Syst Biol* 17 (2): e9982. https:

//doi.org/10.15252/msb.20209982.

Treves, Adrian. 2022. "'Best Available Science' and the Reproducibility Crisis." *Frontiers in Ecology and the Environment* 20 (9): 495–95. https://doi.org/10.1002/fee.2568.

Turing Way Community, The, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, and Kirstie Whitaker. 2019. "The Turing Way: A Handbook for Reproducible Data Science." Zenodo. https://doi.org/10.5281/zenodo.3233986.

Watters, Ethan. 2010. *Crazy Like Us: The Globalization of the American Psyche*. New York: Free Pr.

Webb, E. Kate, J. Arthur Etter, and Jasmine A. Kwasa. 2022. "Addressing Racial and Phenotypic Bias in Human Neuroscience Methods." *Nat Neurosci* 25 (4, 4): 410–14. https://doi.org/10.1038/s41593-022-01046-0.

Wieber, Frederic, and Alexandre Hocquet. 2020. "Models, Parameterization, and Software: Epistemic Opacity in Computational Chemistry." Published Article or Volume. Perspectives on Science; MIT Press. October 2020. https://doi.org/10.1162/posc_a_00352.