PhD PORTFOLIO

Rafsan Ahmed 2024

Cell Death, Lysosomes and AI Lab

Main supervisor: Sonja Aits

This document contains the PhD portfolio addressing the 12 point criteria to be met during the PhD studies.

Email: rafsan.ahmed@med.lu.se

GitHub link to this portfolio:
https://github.com/rafsanahmed/phd_portfolio/blob/main/Rafsan_ahmed_portfolio_2024.pdf

Table of Contents

01. Research Process 5

Start of PhD and Onboarding | Sept 2020 5

Project Plan - First half of PhD | Sept 2020 5

Project Progress by Halftime | January 2023 6

Project Plan - Second Half of PhD (planned before halftime review) | January 2023 6

Halftime Review | February 14th, 2023 7

Project Progress and Challenges till August 2024 7

Image Annotation | June 2024 8

The HALRIC Grant Incident | July 2023 8

02. Research Methodology 10

Methods Applied (Natural Language Processing and Deep Learning) - First Half of PhD 10

Timeline and Plan | Aug 2024 10

Language Models 11

Knowledge graphs and biological pathways 11

03. Subject Expertise 13

Courses 13

Cell Biology and Cell Death 14

Natural Language Processing (NLP) 14

Machine Learning / Deep Learning 14

AI in Medicine and Life Sciences Journal Club 14

04. Publications 16

Manuscript - English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19 16

Manuscript – EasyNER: A Deep Learning-based Named Entity Recognition Pipeline for Customizable Information Extraction from Medical Texts 16

Manuscript - An information mined cell death knowledge graph 17

Manuscript - An insight into the landscape of COVID19 using deep learning and dictionary based text mining approaches 17

Manuscript - An annotated high content fluorescent microscopy dataset with EGFP-Galactin-3 stained cells and manually labelled outlines 18

Scientific Communication and Writing Course 18

05. Teachers Training and Experience 19

Supervision and Teaching..... 19

COMPUTE, AI Lund and Hub AI python workshop 19

Teacher Training Course – Not taken..... 20

06. Conferences and Seminars 21

A summary of the seminars and conferences for the PhD defense opponents..... 21

Mini Conference – "Future Biomaterials: Data-Driven Insight to Cell-Biomaterial Interaction" | May 29, 2024 21

My LUBI Seminar on "reading" millions of research articles | Oct 6, 2023 21

COMPUTE WINTER MEETING with focus on Computatioal scince and sustainability | March 22, 2024..... 22

8th Biomedical Linked Annotation Hackathon and Symposium (BLAH8) Tokyo/Kashiwa, Jan 15-19, 2024..... 22

LUBI seminar "Spacial Omics, Image analysis and data integration in mantle cell lymphoma"| Nov 10, 2023 22

My LUBI Seminar on "reading" millions of research articles | Oct 6, 2023 22

PhD defense of Sakshi Vats in Malmo | Nov 10, 2023 23

Large Language Models and the Art of Using ChatGPT in Academia invited by Hub AI (Sept 21, 2023) 23

ISMB/ICCB Conference Lyon, July 23-27, 2023 23

My Halftime Review Seminar (Feb 14th, 2023) 24

NLP Seminar series organized by AI Sweden (multiple, online) 25

AI Lund Natural and Artificial Cognition (Oct 27, 2022) 25

LUBI Seminar on Single Cell gene expression (Oct 26, 2022) 25

AI Lund Lunch Seminars (multiple, online)..... 26

Half-time seminar of Max Olsson (October 7th, 2022) 26

COMPUTE Summer Retreat: Scientific Communication - from article writing to public outreach (August 23-24, 2022)..... 26

Half-time seminar of Iran Augustos de-Silva (June 29, 2022) 27

34th Swedish Artificial Intelligence Society (SAIS) Workshop 2022 (June 13-14)..... 27

ELLIT focus period 2022 (April 19 - May 20, 2022) 27

CIPA Expo 2022 (March 15)..... 28

COMPUTE winter meeting - Crossing scientific borders (March 9th, 2022) 28

Synapse and HUB AI Presents: Life Science in AI (March 8th, 2022) 28
Swedish Bioinformatics Workshop (SBW) 2021 (October 20-21) 28
07. National and International Cooperation with the Research Community 30
Clinical Chart Extraction – Johanna | 2021 30
Natural and Artificial Cognition Knowledge Exchange | Feb 19-22, 2024 30
HALRIC grand application | July 2023 30
COVID 19 Full Text Extraction - Peter Berck (continuing) 31
08. Cooperation with Wider Society 32
Engagement with Hub AI | 2020-2024 32
Cooperation with organizations 32
Use of my Research by the Wider Society 32
09. Ethical Issues 33
Learning about Ethical Issues - Research Ethics course 2021 33
Addressing AI ethics questions and AI regulations in 2024 33
10. Career Development 35
Career Control for Researchers 2020 35
Career Goals 35
Mentlife 2023-2024 - Helen Sjogren 35
11. Supervision 36
Mentor for Computational Expertise 37
Supervision from Peter Berck (indirect) 37
My Supervision of Masters and Project Students 37
12. Administration, Organization and Leadership 38
Career Control for Researchers 202 Involvement in the Medical Doctoral Council (MDR) 38
Administrative Duties During Events Organized by My Supervisor 39
Appendix 41
A summary of the seminars and conferences for the PhD defense opponents 41
Project Proposal (Draft) for the HALRIC Project 47

01. Research Process

Start of PhD and Onboarding | Sept 2020

The start of my PhD was quite interesting, if not utterly chaotic. I am what the kids these days call a "Pandemic PhD" student. Back in 2020 when I was finishing up my Masters in Turkey, I got the offer to start the PhD in Lund. 2020 was also the year when the lockdown and quarantine were ordinary measures in most countries, except Sweden. Skipping all the details on how difficult it was to arrive in Sweden, our department was under the guideline that those who did not need to be in the office, should not be there. This resulted in a situation where for the first 1.5 years I couldn't:

1. Meet my Supervisor/colleagues in-person
2. Efficiently attend and network in a conference or any physical event
3. Carry out my research with 100% efficiency.

Also, despite taking all courses and passing them successfully, they were not as impactful over zoom. For example, the course on Deep Learning would be more efficient if I had the chance to at least attend live lectures over zoom. But even those were recorded. The situation made me struggle quite a lot with my research topics despite the best support from my supervisor and colleagues.

Even though it was a rough start, I was able to make the best out of the situation. I regularly attended online events and courses to increase my knowledge. I finished all the compulsory courses of PhD education at LU within the first year. I had also actively been a part of the local AI community and took initiative to learn and hold talks/workshops during this time. However, the rough start affected the project and it took me longer than normal to produce impactful results.

Project Plan - First half of PhD | Sept 2020

Artificial Intelligence (AI) methods and especially deep learning show much promise for medical and clinical research use but there are few established methods and workflows. Natural Language Processing (NLP) and image analysis are excellent research domains within AI that has taken a huge leap over the past 10 years. The PhD projects, focused on AI in Medicine, are divided into three separate parts:

1. Development and evaluation of artificial intelligence-based analysis methods for text-based, biomedical image-based and tabular "big data".

2. Studying COVID19 with AI
3. Studying cell death and lysosomes with AI

As a start, I have developed a Natural Language Processing (NLP) pipeline with multiple deep learning based models for scientific literature mining. The results from this pipeline can be initially used to develop knowledge graphs to establish relationships between different biological entities that incorporates both covid19 and cell death & lysosomes research. These results can be eventually used for topic modelling, drug design, concept development and experiment design as well as finding popular interests and density of research topics.

The fundamental project plan was developed by my Supervisor - Sonja Aits. During the Individual Study Plan (ISP) updates we have gone over the plan several times and updated the plan accordingly. The research definitely has a global impact and provides ease of access to biomedical researchers who do not have NLP expertise.

Project Progress by Halftime | January 2023

The project has progressed well so far. We were able to develop an initial pipeline to process scientific text and detect 5 different types of domain specific entities, i.e. cell lines, chemical, disease, gene/protein and species. We have trained and developed multiple Named Entity Recognition (NER) models (several times with different configurations) that accurately detect biological entities from PubMed and PMC articles.

Project Plan - Second Half of PhD (planned before halftime review) | January 2023

By the time of my halfway review, an end-to-end information extraction pipeline has been developed. The pipeline manages to extract information from large biomedical "Big Data". For the second part of my PhD I plan to develop relation extraction models that can detect relationships between these extracted information (entities) from the text. This information will be further used to develop knowledge graphs to access different levels of information from the text and aid researcher in multiple aspects of their research.

Of course, there are several attributes on which the pipeline itself can be improved. One of the major updates are incorporating different NLP models and comparing their performance. Flexibility, portability, scalability and performance of the model are all areas which can see improvements.

I also plan to incorporate image analysis with NLP and use combined information within a pipeline to strengthen research findings.

Halftime Review | February 14th, 2023

The halftime review for the PhD studies involved two opponents:

1. Lars Juhl Jensen from University of Copenhagen and
2. Dag Ahren from Lund University

The halftime review was incredibly useful for the progression of my research. After presenting my work with the NLP pipeline and the COVID19 dictionary, I received many comments and criticisms for my work (along with encouragements). Some of the questions were:

1. Do your models differentiate between proteins and protein complexes?
2. Are the models easily deployable?
3. How did you do model selection?
4. How did you decide on an inter annotator agreement?

While I addressed most of the questions asked, the ones above in particular were quite insightful. They made me re-assess my work and improve upon several weak points. Such as, model deployment and data quality. For such challenges I had several discussions with my supervisor and improved the overall tools. For example, to address easier deployment, we hosted the models on an online portal (huggingface).

The interaction during the halftime shaped the future of the project to some extent. My supervisor was also quite satisfied with the outcome.

Project Progress and Challenges till August 2024

The PhD projects have taken an accelerated jump after the halftime. I used two supercomputing clusters to process around 40 million research abstracts and collect that data. This was planned before. But the amount of time, skill and computation power required were certainly underestimated. Following were the major contributions/impacts until August 2024.

1. Finalizing EasyNER paper, toolkit and submission.
2. Processing entire PubMED and running several models on it,
3. Addressing challenges with sentence level processing vs paragraph level processing.
4. Developing knowledge graph on COVID19
5. Developing knowledge graph on cell death

Also, one of the major challenges of my research was and still is to tie everything together. With large scale data the result can be quite scattered. But I learned, with the help of my

supervisor and collaborators as well as other experts that I met along the way, to deal with these issues. A lot of people in the world, especially after the explosion of CharGPT, are working in the same field as us and we have received help and guidance in understanding and addressing these challenges.

At this stage, I was not only involved with running the ML models, but also aimed towards finding good analytical solutions for the sum of large datasets that we were using. At the same time, I invested myself into self-learning different tools such as neo4j (a graph database tool) and chroma (a vector database tool) in order to run the analysis we had intended to run. For the image project I also had to quickly learn CVAT, an image annotation tool, to run the intended tasks for the projects.

Overall, the second phase was much more accelerated and challenging.

Image Annotation | June 2024

The NLP project for my PhD was always intended to be compared and tested against cell images in order to gain biological insights. To dive a bit more into the image project, I annotated microscopic cell images. These images are meant to be training data or "true" data for image models that are trained to detect cell boundaries. It was a long and tedious work to annotate 10 slides. But it was done within a week. During which I learned a new program named CVAT for cell annotation.

The research aspect of it was pretty straightforward. Annotate the images following inter annotator guidelines we devised and compare to make sure no human errors were made. I had to quickly learn an image annotation tool named CVAT. It was a good experience into the imaging part of the larger project. I learned quite a bit about cell boundaries and cell morphology, which was fantastic.

The HALRIC Grant Incident | July 2023

During my ISMB/ECCB 2023 conference visit, I met the head of research of European Bioinformatics Institute (EBI). They were interested in processing the large collection of research articles that they possessed, which is similar to the work we aimed to do during the PhD. I was also aware that, the HALRIC initiative at Lund University was quite new and were accepting applications. I put two and two together, had a few meetings with them and put together a research proposal including my supervisor. The proposal was to run similar pipelines that we were working on on EBI's repository, Europe PMC. It was particularly interesting for two reasons:

1. They had a large collection of text that were publicly unavailable.
2. They also had abstracts, which PubMed or public repositories did not have.

With our EasyNER tools, we could process the articles in parallel and gain deeper insights. The idea was to build networks/knowledge graphs from the insights we gained and look at similarities and differences between our original work and EBI papers.

Once I notified my supervisor, she seemed skeptic at first. This was a surprise since she asked me to put together short project plans countless time before and helped me learn more about writing project proposals. We attended several meetings with HALRIC coordinators and contacted the economist for collaboration overview, management and finances. However, in the end my supervisor informed me that we cannot proceed with this project, even though HALRIC was ready to give the grant. The reason stated was the overhead percentage.

To be honest, this was a crushing blow to my work and I felt like I wasted months of work and networking. It seemed like I missed out on a great opportunity where I could expand my research more and helped answer the questions we had more broadly.

However, I did learn a lot from all parties involved, learned how to write descriptive project proposals and to lead a project. Overall, it was a bittersweet experience, more bitter than sweet.

Following is the summary of the project proposal:

Duration: 6 months at the beginning, with possibility for extension

Collaborators: HALRIC, EBI and University of Oslo (supercomputer)

Economic contributions: 40% of salary + 15% overhead + 6% travel

02. Research Methodology

Methods Applied (Natural Language Processing and Deep Learning) - First Half of PhD

For the PhD projects I have been primarily working with Natural Language Processing (NLP) and Deep Learning to extract information from datasets. Before starting the PhD I had some Idea about the area and worked on some projects with specific applications on methods within the deep learning domain. However, for my current projects I had to extensively learn these topics, especially NLP. The way I learned was through application, since the medical faculty did not have courses to support computational degrees, unlike LTH.

So far in the first half of my PhD I have learned and applied concepts in NLP like Named Entity Recognition (NER), Relationship Extraction (RE), graph traversal, attention and transformers, structured document building and many more. Contemporary research shows that Transformers and Attention based approaches for NER and subsequent tasks have worked quite well for information extraction. I have also delved into image processing and image analysis as our lab has several open image projects. There are plans to combine NLP and image analysis to get insight into multi-domain data.

The NLP methods are excellent in making predictions (e.g. HeLa is a cell-line), understanding context (e.g. HeLa cell-lines and Hela the Goddess of death in Norse Mythology aren't the same) and extract specific information (e.g. HeLa cell-lines and cancer research) . However, generalization still remains a key question that needs to be solved. Many transformer-based approaches are currently swarming the world (GPT3, chat-GPT etc) and these methods need to be explored in my research area.

Timeline and Plan | Aug 2024

Unfortunately, the time initially proposed for my PhD was not maintained. Which may be common for a lot of PhDs, but in my case it is one step further. The initial goal was to publish the EasyNER paper by 2022. Similar for the Gold Standard paper. However, they got postponed due to several factors, which I do not want to elaborate further. The time delay also meant that upcoming publications and experimentations were halted. By August 2024 I have applied for an extension for 2 months (from end of September until December 6th), which the publication time would exceed by a large margin.

In August 2024, the timeline is as follows:

- Submission and rejection of Gold standard paper: 2022
- Submission of EasyNER paper: Done in February 2024
- Submission of image annotation paper: Done in July 2024
- Resubmission for EasyNER paper: August 2024
- Experimentation of Cell death paper: August 2024
- Submission of cell death paper: September/October 2024
- Experimentation of COVID paper: October/November 2024
- Submission of COVID paper: Unknown
- Submission of thesis: preliminary date September 6th, 2024, however that seems to be an unattainable goal.

The timeline now, which has been agreed upon by both my supervisor and the department, is not only highly ambitious, but impossible to attain. It seems that a further 6 months minimum is necessary to finish the experiments, tie in the loose ends and submit the thesis.

Language Models

The strength of my research is in the language models. These are special AI foundation models that we train on a specific domain of data to develop domain specific models. Such as, a chemical model to detect chemical entities within a given document. During my research language models (LM) research exploded and models like ChatGPT gained popularity. Of course the strength of these models are that, they can detect entities based on context. However, they are not perfect.

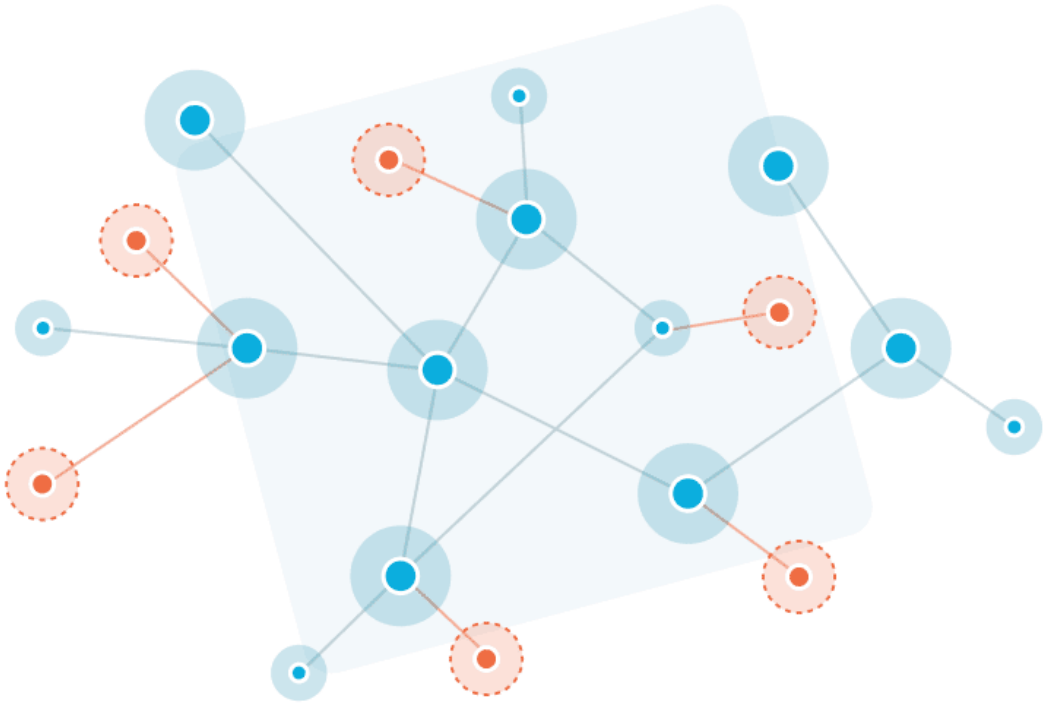
Their imperfections were demonstrated by my research finding. Particularly the cell model did not perform well. Other models also often had overlaps i.e. one entity (word) was categorized as both gene and disease by two different models. Such issues were addressed in post processing, but still the outcomes remained unpredictable. Which is one of the weaknesses of this method. At the same time, language models are the best approach by far for such experiments. So we took them with a grain of salt.

Knowledge graphs and biological pathways

Another direction I have taken was knowledge graphs or networks. The initial research question was, "can we draw a landscape of the knowledge of a particular disease, say COVID?". In order to answer such questions, we needed interconnection between entities. A disease is linked to one or several genes for example. We aimed to address these connections to find new knowledge, like biological pathways, neighbors and much

more. For example, if HeLa is a cell line, what kind of diseases can we connect with that cell line? The challenge with such graphs is that we need to evaluate each and every link. Therefore, in terms of computation, it can be quite expensive. We aimed to build one graph for cell death and another one for COVID and ask specific research questions that can be extracted from the graphs.

These interconnected graphs, like language models, is a great method for finding patterns. There are more traditional machine learning approaches to find these links, for example, agent based learni ng. But graphs are simpler and we can work with them on different layers. Therefore, they were the prime choice for the PhD projects.



03. Subject Expertise

Courses

I have taken the mandatory courses in the medical department and taken a Deep Learning course provided by Mattias Ohlsson. The courses in the department are given in 1 week stretches, except the Statistics II course which spans 2 weeks. Since my PhD is computational, all of these courses except the elective and statistics were not useful to me, even though I completed all of them in the first year. The courses (including electives) are designed for students with medical/clinical/biology/biomedicine background. For someone with computational background, a course in qualitative methods is not worth the time.

On the other hand I could benefit a lot from more computational courses, especially in the first years. Courses on containerization, parallel computing and more importantly, computational and method oriented NLP would benefit me tremendously. I did attend one or two online MOOC on NLP by my own and learned mostly through application. But I believe they were not sufficient for overall knowledge development. At least a couple technical classroom courses were necessary.

Taking additional courses, while I thought could help me gain a deeper knowledge for my research, was deemed unnecessary by my supervisor. Instead she advised to gain knowledge on the projects and learn by doing. Which was not my initial preference, but I took her advice. We also had a few one-on-one session and peer learning groups for sharing knowledge, advised by my supervisor. That worked quite well and I am happy with the outcome.

My supervisor has also organized several local events and talks for me to be a part of (for example, COMPUTE events). These events were also quite helpful in getting a broader perspective into the field. I got a chance to both learn from and speak to a wider spectrum of researchers which helped my understanding a lot.

Out of all the additional courses I have taken over the PhD, the writing course was particularly notable. Not only did I learn about the publication process thoroughly, but I also wrote drafts manuscripts and cover letters and had them evaluated. Reflecting back on it, that approach of learning was great and suited me quite well. It would have been great if I was given opportunities to attend more relevant courses as this.

Cell Biology and Cell Death

Since I am not a biologist, this was the most important skill to acquire. I have attended many seminars focused on cell biology, pathways and drug development which helped me a lot to gain more knowledge in this area. My supervisor also advised me to read more papers within cell biology. Particularly in the domain of cell death. I can sense that my level of understanding has improved over time, but there is still a lot to learn.

A major subject expertise to gain with my research is expertise in functional pathways. It was interesting to learn more about how functional pathways work and dive into how NLP and knowledge graphs can help find novel links or gaps within research

Natural Language Processing (NLP)

NLP was a fresh topic for me when I joined. My main way of learning about NLP was through countless medium and towardsdatascience blog posts, machinelearningmastery.com website and the Stanford NLU course - which I still haven't completed in full. But I also mostly learned by coding. For example, I have coded to understand popular NLP packages like NLTK and spaCy and transformer models like BERT and BioBERT to understand transformers and attention. I have also regularly attended the NLP seminar series from AI Sweden.

After the popularity of ChatGPT, the field of NLP exploded. It is difficult to keep track of all the novel papers coming out every day. The BLAH8 hackathon/symposium I attended in Tokyo 2024 also helped me a lot to gain a deeper understanding of NLP research and Life Sciences. I continue learning and exploring more within this area.

Machine Learning / Deep Learning

The machine learning/deep learning community in LU is not the strongest, but it is not the weakest either. AI Lund and COMPUTE research school and student organization Hub AI regularly organizes talks, seminars and other events where people with different expertise can participate. I am also in touch with several labs working with machine learning in Lund and try to stay on top of things.

AI in Medicine and Life Sciences Journal Club

I had organized and hosted the AI in Medicine and Life Sciences Journal Club under COMPUTE research school from 2021 to 2023. It was a monthly event where a presenter would present one paper relevant to AI and life sciences. It was great to read interesting

papers every month both from the medical faculty and internationally. I believe I gained a lot of contemporary knowledge on AI, machine learning and life sciences as well as gained some leadership skills by hosting the journal club.

04. Publications

Manuscript - English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19

Mini Abstract:

For dictionary- and deep learning-based Natural Language Processing (NLP) models to detect entities from given text, they require annotated datasets. We have annotated 10 articles abstracts on COVID19 and build several dictionaries which can be used for information extraction

Contribution:

The initial phase of this project was already done by my supervisor and my colleagues. I came into this project as my first project for the PhD. I have enhanced the dictionaries, evaluated the Gold Standard annotations and fixed the scripts. From a publication viewpoint, I have added my contributions to the manuscript and been involved in writing/revising the final draft. Submission and

all additional matters were handled by my supervisor.

DOI/link: https://doi.org/10.48550/arXiv.2003.09865

Manuscript – EasyNER: A Deep Learning-based Named Entity Recognition Pipeline for Customizable Information Extraction from Medical Texts

Mini abstract:

In this paper, we developed an end-to-end, dictionary- and deep learning- based pipeline that can process a large volume of article abstracts (in millions) or plain text and can return insights into the data. The paper uploaded on a preprint server has already been cited multiple times.

Contribution:

I am the sole first author on this paper. I have rebuilt the entire pipeline from scratch and contributed to most of the technical work. In terms of publication, I had written the first draft, made corrections, generated images/created tables, written a draft of the cover

letter and currently running experiments to address the comments from reviewers. We have also had open discussions as a group about choice of journal, choice of reviewer, addressing the comments etc.

My supervisor took the responsibility for submission and writing the rebuttal letter, as well as draft corrections and rewriting the manuscript. We are currently in the process of resubmitting this paper.

Arxiv: https://arxiv.org/abs/2304.07805

Manuscript - An information mined cell death knowledge graph

Mini Abstract:

An information mined collection of entities that are obtained using text mining over the entire PubMed collection of articles are used to generate knowledge graphs in order to gain insights into cell death and associated biological processes.

Contribution:

I am the sole first author of this paper as well. I have run all experimentation and generated visuals. Publication wise, I have written the first draft. The journal is already decided by my supervisor who will revise the draft, rewrite and submit.

The publication process for this paper can take long. Which may exceed my current extension application ending in December 6, 2024.

Manuscript - An insight into the landscape of COVID19 using deep learning and dictionary based text mining approaches

Mini Abstract

This project is intended to use full text biomedical articles and extract information on COVID-19 and associated biological processes using text mining.

Contribution:

I am the first author of this paper. We are currently in the process of running the experiments and text mining. Publication wise, I have the responsibility to write the first draft in collaboration with Peter Berck, one of our collaborators. The submission process will be collaborative.

The publication process for this paper can take long, which may exceed my current extension application ending in December 6, 2024. This paper will most likely be a draft submission for the thesis.

Manuscript - An annotated high content fluorescent microscopy dataset with EGFP-Galactin-3 stained cells and manually labelled outlines

Mini Abstract:

This paper is an annotated dataset paper that contains rich, manually labelled EGFP-galactin-3 stained cells. This dataset can be used as a training or benchmark dataset for image segmentation models.

Contributions:

As second author, I have manually annotated stained cells following inter-annotator agreements. Publication wise, I reviewed the first draft and added my comments/contributions. Submission and additional responsibilities were taken by my supervisor.

Scientific Communication and Writing Course

During the writing process the scientific communication and the writing course helped me a lot. Especially the writing course, where I got to learn about the different processes of writing and submission. Even though I have published articles previously, learning about different methods and analysing different articles was a necessary skill for the PhD.

05. Teachers Training and Experience

Supervision and Teaching

I have co-supervised several groups of students with their Masters projects in NLP in our group. That included programming support, technical support and review of thesis. I have also been a teaching assistant in most of Sonja's courses, where I helped the students answer questions or lead discussions.

Nils Broman: Nils did his Masters project with us where he explored normalization of chemical and disease entities. My supervision duties included sharing data, showing how the scripts work and guide him through the process. He was quite independent and did not need much of my help. I spent around 1/2 hour per week with him for a span of 3 months. He was also pretty quiet and I wish our interactions more dynamic.

Jacob Krucinski: He also did his masters project with us on named entity recognition. He was working as a distant student. I helped him set up the pipeline I have been working on so that he could work on downstream tasks. Jacob and I had regular meetings and spent around 2 hours a week for 4 months. He was quite happy with my supervision and we still stay in touch.

Carl Olivik Aasa: He is currently doing his masters project with us on environmental entities (Aug 2024). He is quite independent as well so he doesn't need much input from myself. Since he is also a medical doctor, we got to learn a lot from each other and it was quite a pleasant summer. I spent around 2/3 hours with him per week.

Supervision so far has worked pretty well. I usually don't micromanage and let students figure out things by themselves. But I also find that I can occasionally involve myself more in their projects and lead brainstorm sessions.

COMPUTE, AI Lund and Hub AI python workshop

I gave a crash course on python and launched peer learning groups combining three of the largest AI organizations in Lund. The crash course went great with large student participation. I held weekly follow up sessions for 3 months. The initial event went great. A large number of students (and researchers) were keen on learning more about python. I felt like I had a strong impact. However, the follow ups and peer learning groups, in my opinion, were not successful. The reason behind it was mainly

communication channel and lack of group activities. But it was a learning experience for future events.

Teacher Training Course – Not taken

Since I didn't teach any classroom courses, I did not take the teacher training course. I was informally and briefly supervised on teaching duties by my supervisor.

06. Conferences and Seminars

A summary of the seminars and conferences for the PhD defense opponents

A table including the summary of seminars is attached at the **end** of this portfolio

Mini Conference – "Future Biomaterials: Data-Driven Insight to Cell-Biomaterial Interaction" | May 29, 2024

This conference focused on biomaterials and future biomaterials, which isn't exactly my research area. However, the conference ended up quite relevant. I was able to interact with a lot of researchers working with AI and knowledge graphs but in different applications. I also learned a lot about explainable AI and practices from an open discussion session that was held. It was a great mix of technology experts and biology/biomaterials experts. If I was to attend the same conference next time, I would do a little bit more background research on biomaterials. I often felt out of place and had to ask a lot of questions. Our lab jointly presented a poster of all the research work done here.

My LUBI Seminar on "reading" millions of research articles | Oct 6, 2023

I gave a seminar titled "EasyNER: Using artificial intelligence to "read" millions of articles in life sciences" through the Lund University Bioinformatics Infrastructure (LUBI) organized by Karin Engstrom. The talk was about using my EasyNER pipeline to text mine information for the life scientist.

My talk was well received by 20-25 attendees at BMC. I received several questions on Language Models and applications of AI. I felt that the talk was meaningful and a couple attendees later reached out to me to talk more about how they can apply the tools for their own research.

Reflecting back on the oral communication course, it was helpful to organize my thoughts into slides. However, the course didn't include much that I already didn't know about communication and presentation. Yet, the practice was helpful for talks like these.

COMPUTE WINTER MEETING with focus on Computatioal scince and sustainability | March 22, 2024

Since our LLM training takes a lot of computational resources, it was nice to think about sustainability more in our contributions. In this seminar, there was a lot of discussion on optimizations and sustainable approaches towards computation. I really liked one talk about utilizing GPU resources and minimizing runtime by Lund University HPC architecture (LUNARC). Since we use the resources regularly, it was good to learn their process. We also had a lot of open discussions on best approaches towards sustainable AI.

8th Biomedical Linked Annotation Hackathon and Symposium (BLAH8) Tokyo/Kashiwa, Jan 15-19, 2024

BLAH8 is an annual linked annotation hackathon and symposium in Japan where researchers working with NLP and biomedical datasets gather in Japan, present their research and solve hackathon problems. I met my halftime opponent again in Lyon ISMB/ECCB conference in 2023 where he invited me to this symposium. I received a fully funded trip as well as a great arsenal of researchers who could help me solve my research questions. Not only was I able to lead my own project and problems there, but I could also learn and share ideas collaboratively. I also had the opportunity to have a lunch with Larry Hunter, a star in the field, and discuss my phd projects with him. I also met a lot of people I knew only through their research. I am hoping to invite one of them to be my PhD opponent.

The biggest take away from this conference was to share my work and learn from the best minds in my field. It was one of the best experiences during my PhD

LUBI seminar "Spacial Omics, Image analysis and data integration in mantle cell lymphoma"| Nov 10, 2023

This talk by Lavanya Lokhande included an insight into immunotechnology and data analysis. She integrated omics data analysis and image analysis which was interesting to learn about. A lot of the topics were still difficult to comprehend. But I was mostly interested in how she analysed her data and what insights she could derive from it.

My LUBI Seminar on "reading" millions of research articles | Oct 6, 2023

I gave a seminar titled "EasyNER: Using artificial intelligence to "read" millions of articles in life sciences" through the Lund University Bioinformatics Infrastructure (LUBI) organized by Karin Engstrom. The talk was about using my EasyNER pipeline to text mine information for the life scientist.

My talk was well received by 20-25 attendees at BMC. I received several questions on Language Models and applications of AI. I felt that the talk was meaningful and a couple attendees later reached out to me to talk more about how they can apply the tools for their own research.

Reflecting back on the oral communication course, it was helpful to organize my thoughts into slides. However, the course didn't include much that I already didn't know about communication and presentation. Yet, the practice was helpful for talks like these.

PhD defense of Sakshi Vats in Malmo | Nov 10, 2023

Sakshi was a PhD student at Malmo CRC and the President of MDR, the doctoral student union. Her research was about a particular disease called Abdominal Aeortic Aneurism (AAA) with a high mortality rate. She primarily investigated oxidative stress related factors, case studies and mapped the diagnostic landscape of AAA.

The talk was quite interesting, from a scientific perspective yes, but more from a procedural perspective of the PhD defense. I learned more about the process of the PhD defense and how it should be conducted. I also learned a lot about oxidative stress and free radicals. In my opinion, she defended her thesis very well.

Large Language Models and the Art of Using ChatGPT in Academia invited by Hub AI (Sept 21, 2023)

Along with the study director of LUSEM (LU), Bjorn Svensson, I gave a talk on Large Language in Academia and ChatGPT. I also briefly touched upon my own research. The talk was mostly well received with a dash of scepticism towards an AI driven world. But overall it was a great discussion that involved students and researchers but also teachers and students.

ISMB/ICCB Conference Lyon, July 23-27, 2023

The ISMB/ ICCB conference is an annual conference for bioinformatics, data scientists and technology enthusiasts within life sciences. The ISMB/ICCB in Lyon, 2023 involved around 4000 participants. It was incredible to meet people whose papers I have read and cited. Especially Robert Lehman who was leading the text mining cohort of the conference and was kind enough to discuss my project in detail and compare with his own similar one. I learned a lot and alongside my poster presentation, I also got to give a flash talk in front of hundreds of participants. The conference also paved its way for me to write the HALRIC project proposal (mentioned in section 1) as well as to visit Japan for the BLAH Hackathon in 2024.

This was an amazing experience for me. Not only did I get to meet the best researchers in my field, but I also got to put my research out their and be appreciated for the work I have done. It also gave me a lot of practice on presenting my research in front of experts.

Only negative experience was the food. It was awful.

My Halftime Review Seminar (Feb 14th, 2023)

With opponents Lars Juhl Jensen from Copenhagen university and Dag Ahren from LU, my halftime review was an amazing experience. I have known Lars for years, because of his work within network biology and text mining. Therefore, I was slightly nervous. My talk included details from my research projects and a balanced mix of pros and cons. Both Lars and Dag asked a lot of questions and scrutinized my work. However, I was able to answer their questions and have a detailed discussion. I also receive a lot of ideas (mentioned in section 1, research process) that I have implemented in the months to come. I have also had the opportunity to meet Lars on a few other occasions and

discussed our project broadly. My supervisor also commented that the talk and the defence/discussions were satisfactory to her.

NLP Seminar series organized by AI Sweden (multiple, online)

Natural Language Processing (NLP) and Natural Language Understanding (NLU) is a key area that AI Sweden has increased their focus since 2021. They have also been holding hybrid NLP seminars Wednesdays every odd week since 2021. I have attended several of these 45 minute seminars for better understanding of the Swedish (and global) NLP landscape. They have discussed several aspects of NLP and NLU research, including ethical and decision-making aspects that were quite useful for a broader understanding of NLP.

The seminars included top NLP voices in the domain from all across the world. I have attended 12 such seminar online between 2021 and 2023 that included topics from foundation models, scaling, ML and collaboration, context in life sciences, ethics, misinformation etc. These seminars were incredibly useful for my research as they were directly relevant to my work. I also gotten in touch with several of the speaker like Sampo Pyysalo, Barbara Plank and Magnus Sahlgren to ask specific questions about their work in the field. I have also met Sampo in several other events (e.g. Tokyo) and discussed his talks with him.

https://www.ai.se/en/events/nlp-seminar-series

AI Lund Natural and Artificial Cognition (Oct 27, 2022)

This waws seminar with various AI and machine learning talks. Most were incredibly interesting, such as controlling animal actions through nano probes, gestures and linguistics, mosquito cognition etc. Though most talks were not about NLP or Medicine (there were a couple), the workshop was useful in the broader machine learning perspective.

LUBI Seminar on Single Cell gene expression (Oct 26, 2022)

The talk was about single Cell transctiptomics TrusTEr and the speaker Yogita Sharma gave a case study on brain trauma. I wanted to learn a bit more about single cell data and how to train models to process such data, however this talk was not too relevant and the focus was primarily on the experiments rather than application.

LINXS workshop on biomedical imaging (Oct 19, 2022)

A workshop on biomedical imaging that gave basically a crash course on normal biomedical imaging as well as synchrotron. It was a good intro to imaging. We had several attendees from AstraZeneca. 6/10 for a workshop.

AI Lund Lunch Seminars (multiple, online)

AI Lund organizes periodical online lunch seminars where there is usually a short presentation and people can ask questions on various topics. I have attended several of those spanning topics such as image processing, NLP, transformers for image generation, NLP and psychology, AI ethics etc. These are complex topics but the seminar allows open discussions and networking to help anyone in their work. I have found several of these seminars quite useful for understanding topics of AI better.

One notable seminar, as example, was the "Fine-Grained Image Classification of Groceries for Assisting Visually Impaired People" seminar on Oct 12th, 2022 given by Marcus Klasson. Since a part of my PhD project is on imaging, it was interesting to learn a bit more about how automatic classification pipelines are developed for real time applications

Half-time seminar of Max Olsson (October 7th, 2022)

Max Olsson has been researching on and epidemiological and data driven study on breathlessness. It was a very interesting seminar where I not only learned a lot about the topic but also learned half-time process from a different faculty. Max's research on breathlessness aims to use the Swedish National Register to find trends and better treatment options.

COMPUTE Summer Retreat: Scientific Communication - from article writing to public outreach (August 23-24, 2022)

The first physical COMPUTE Summer Retreat was a great way to get away from the office research environment to the the lovely village of Arlid. Alongside sightseeing, jacuzzi and networking, we also had talks and group work on effective communication. The invited speaker Dan Csontos held workshops over the two days on communication scientific findings to both academic and non academic participants as well as managing time.

I also presented my poster on NLP pipeline on the poster session held on the first day.

Half-time seminar of Iran Augustos de-Silva (June 29, 2022)

Iran has been working on histopathological analyses of tissues to score lung damages. In his project there was application of machine learning on images to automatically score images. In his half-time seminar he explained the methods he used to obtain the pig lung tissues and how he captured the images. He also gave a description of the variability of scores among human scorers/pathologists and why it is difficult for a model to train on such scores. It was interesting since, a talk on the SAIS conference I attended in just a couple weeks before also addressed this issue. It was nice to see that some problems are not specific to one single type of research and helpful solutions can aid researchers across multiple domains.

34th Swedish Artificial Intelligence Society (SAIS) Workshop 2022 (June 13-14)

The SAIS workshop was one of the first national conferences that I could physically attend. It was also one which was purely AI focused. The first keynote talk from Barbara Plank on human label variation was very interesting and quite relevant to our projects. It was also fantastic to present my poster on my Natural Language Processing pipeline and get great feedback from some of the local experts.

https://www.ri.se/en/sais-2022/programme-for-sais-2022

ELLIT focus period 2022 (April 19 - May 20, 2022)

ELLIT is a strategic research environment funded by the Swedish government in 2010, with four partner universities. I was able to participate in the 5 week long focus period where I presented my current research, attended a hackathon and joined scholars from different countries for a focused learning period. The theme for 2022 was "Data-driven modelling and learning for cancer immunotherapy"

As a local I was also part of the organizing committee: organizing social events, building teams and communicating with attendees throughout.

Hackathon: The hackathon had four different project options, among which I worked on parts of three of them:

1. IMVIGOR: Analyzing clinical data from bladder cancer patients (provided by Roche) to develop a prediction model.
2. Copula GAN: Using clinical data to generate synthetic data using copula GAN.

3. Building knowledge graphs from genomics, PPI and gene expression data.
4. Overall, the experience was fantastic. Not only did I get to work with leading scholars in the field, but I also got a chance to work with real data. I'm looking forward to attending such events in the future.

https://elliit.se/news-and-events/focus-period-lund-2022/

https://elliit.se/rafsan-ahmed/

CIPA Expo 2022 (March 15)

Correlative Image Processing and Analysis (CIPA) had its first annual expo in 2022. Several interesting talks took place about image processing and visualizations. One talk in particular from Alexandros Sopasakis was quite thought provoking. He talked about adding domain knowledge and expertise to model training. He has already found some great results with this process and that can be very useful for our image analysis and nlp projects.

COMPUTE winter meeting - Crossing scientific borders (March 9th, 2022)

The COMPUTE research school at Lund University is a great hub for researchers who use computing in their research. But it is also a place for learning and collaboration. I have been actively involved with compute during my PhD and attended several events. The winter meeting at LTH was the first physical COMPUTE meet for me since the pandemic. There were several interesting talks, including one from Nikolay Oskolkov about communication between scientists in different areas (e.g. biology and physics).

Synapse and HUB AI Presents: Life Science in AI (March 8th, 2022)

I organized and co-hosted a presentation and panel discussion on life sciences in AI: a collaboration between Synapse Sweden and my own student organization Hub AI. We had three speakers on AI Sweden, AstraZeneca and NBIS/Scilife Lab to give talks on their own work within medical and AI domains. It was a great learning experience not only because of the great talks but also about organizing collaborative events like this.

Swedish Bioinformatics Workshop (SBW) 2021 (October 20-21)

The SBW 2021 occurred during 20-21 October, 2021 (online). I presented the following poster with a 3 minute flash talk. Since it was the first poster I presented during my PhD,

it was a great experience to summarize what I had done so far and put it to writing . Following is the poster I have presented.

A Deep Learning Pipeline for COVID19-related Information Extraction from Scientific Literature

LUND UNIVERSITY

RAFSAN AHMED[1], PETTER BERNTSSON[1], ALEXANDER SKAFTE[1], ADAM BARVESTEN[1], OLA OLDE[1], SALMA KAZEMI RASHED[1], ANTTON LAMARCAARRIZABALAGA[1], MARCUS KLANG[3], PIERRE NUGUES[3] AND SONJA AITS*[1,2]

Main Points

- Identifying domain specific entities is a challenging task, particularly with large datasets.
- An end to end, information extraction pipeline is developed which is trained on several Named Entity Recognition (NER) corpora.
- It can process large text collections and can detect domain specific entities such as cell lines, disease etc.
- The pipeline was deployed on a collection of over 400000 articles on COVID19.

Introduction

- The COVID-19 pandemic generated an extensive amount of research articles.
- Such a large dataset can provide us with key insights, however it greatly surpasses human processing abilities.
- Deep Learning-based Natural Language Processing (NLP) can help extract and connect such information

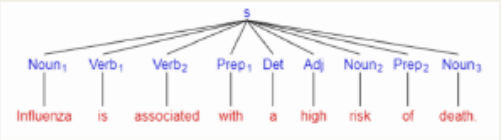


Fig 1: A sentence is broken down into parts of speech using NLP

Dataset and Models

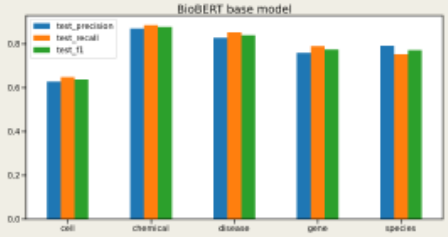
- Training Data: HunFlair Corpora integrated on the following entities:
  - Cell Lines
  - Chemical
  - Disease
  - Gene/Protein and
  - Species.
- Model: BioBERT base cased fine tuned on each entity above
- Main Dataset: CORD19 – a collection of 768929 research article abstracts and 454230 full-text articles on COVID 19

Methods

- Each HunFlair Corpus is pre-processed and fine-tuned on different BioBERT models.
- The models are then evaluated on a test set.
- The entire CORD19 dataset is converted into Docria document collection: a structured document format.

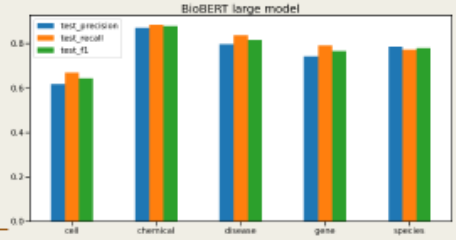
Fig 2: CORD19 dataset is converted into a structured Docria Collection. The collection contains main text, sections, sentences and entities

Evaluation Results



	Cell	Chemical	Disease	Gene	Species
Precision	0.628	0.871	0.828	0.759	0.792
Recall	0.648	0.886	0.854	0.789	0.752
F1 Score	0.638	0.878	0.841	0.774	0.772

Fig 3: Evaluation results of fine tuned BioBERT base models on each corpora



	Cell	Chemical	Disease	Gene	Species
Precision	0.618	0.870	0.797	0.743	0.787
Recall	0.670	0.884	0.838	0.792	0.773
F1 Score	0.643	0.879	0.817	0.767	0.780

Fig 4: Evaluation results of fine tuned BioBERT large models on each corpora

References

1. L. Weber, J. Münchmeyer, T. Rocktäschel, M. Habibi, U. Leser, "HUNER: improving biomedical NER with pretraining", Bioinformatics, Volume 36, Issue 1, 1 January 2020, Pages 295–302.
2. M. Klang and P. Nugues, "Docria: Processing and Storing Linguistic Data with Wikipedia", Proceedings of the 22nd Nordic Conference on Computational Linguistics, 2019, pages 400-405
3. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", Bioinformatics, Volume 36, Issue 4, 2020, Pages 1234–1240
4. L. W. L, L. K, C. Y, et al. "CORD-19: The Covid-19 Open Research Dataset." Preprint. ArXiv. 2020;arXiv:2004.10706v2. Published 2020 Apr 22.

Affiliations

1. Cell Death, Lysosomes and Artificial Intelligence Group, Department of Experimental Medical Science, Faculty of Medicine, Lund, Sweden
2. Lund Institute of Advanced Neutron and X-ray Science, Lund, Sweden
3. Department of Computer Science, Lund University, Lund, Sweden
*Corresponding Author

07. National and International Cooperation with the Research Community

Clinical Chart Extraction – Johanna | 2021

I developed a script to extract text from clinical charts for a collaborator. The scripts were used to extract patient information from PDF documents. Due to GDPR rules I couldn't access the real patient data, but a couple "fake" self generated ones. The script extracted information quite well, such as blood pressure, temperature and such. The final product, “Openchart” was used to generate real-looking patient documents that can help an AI model train.

Natural and Artificial Cognition Knowledge Exchange | Feb 19-22, 2024

As part of knowledge exchange of the Natural and Artificial Cognition Cohort, I spent 1 week in the Humanities Lab led by Marianne Gulberg and supervised by our existing collaborator Peter Berck. This week was supposed to be a an exchange of ideas with the Humlab researchers as well as to work on my own research with Peter.

The experience was amazing. I had no idea initially about the amount of ML and linguistics research done at Humlab. I got to learn a lot about applications of the same architectures I use on similar and different types of data. Different researchers like Joost, Johan Frid, Jan etc gave me individual demonstrations on their research and we had discussions about potential applications of NLP in their work. I also had a lengthy talk with the director about future goals and how collaborative work may be carried out by our departments. I also reported my experience to my supervisor. We both reflected that, in the future this kind of interaction can be a first step towards any project collaboration.

HALRIC grand application | July 2023

The details on the HALRIC grant application is mentioned in section 1 of this portfolio. In summary, a collaborative initiative between EBI, University of Oslo and Lund University

was proposed and all collaborative steps were taken by me for a grant application. This involved several meetings and discussions with each party.

Unfortunately, my supervisor decided not to allow submission of the project proposal. Therefore, this funded and collaborative project of 6 months was scrapped.

COVID 19 Full Text Extraction - Peter Berck (continuing)

We started working with Peter ever since he joined the Humanities Lab (HumLab) in February 2023. It was beneficial in multiple ways. Him being a linguist and Natural Language Processing (NLP) researcher, he wanted to work more with different data models. Whereas, I could use technical expertise and guidance in my NLP projects. My supervisor initiated the collaboration and we have been working on COVID19 papers ever since. We aim to build a knowledge graph to map the landscape of COVID related entities.

My role in this collaboration has been very active and we have regular weekly meetings. So far this collaboration not only strengthened the project but also increased the pace. My supervisor maintains the collaboration with Peter and the HumLab.

08. Cooperation with Wider Society

Engagement with Hub AI | 2020-2024

Through my 4 year term at Hub AI, I have had the opportunity to talk about my research in layman terms with people from a wider range of background. While it has been mostly with students, I have also had audiences who were high school educators, lawyers, developers etc. For example, my Hub AI talk on Sept 21st, 2023 on large language models and ChatGPT and my panel discussion on AI and COVID19 on Nov 11th, 2020.

Through these popular science talks I reached a wider range of audience and was able to discuss various interesting questions that I may not face on a day to day basis. For example, one Swedish teacher was completely opposed to allowing students to use AI in the classroom and would comment that "AI would make kids dumber". In that particular interaction, we had a discussion about similar events in history, such as introduction of the calculator etc. These interactions were not only insightful but they also led to answers to important societal questions..

Cooperation with organizations

I have been involved with several organizations in Lund: Hub AI, AI Lund, AirLund and COMPUTE. Workshop and seminars from these organizations are usually open for public. For example, material from the python workshops are available in github for anyone to access.

Use of my Research by the Wider Society

My research in AI is open source and available to the public over github (https://github.com/Aitslab/EasyNER, https://github.com/Aitslab/corona etc). The toold that I have developed can be used by anyone with a computer and a little command line skill. The tools can also be used for drug repurposing, finding missing/hidden links between disease, drugs and potential therapies etc. The impact can be endless. However, with the new AI regulations in place and the potential for mishandling, results obtained using such tools need to be carefully inspected before applying to medical/clinical/experimental use.

09. Ethical Issues

Learning about Ethical Issues - Research Ethics course 2021

I have taken a research ethics course from the faculty to learn about academic integrity and ethical assessment. With the GDPR rules and AI regulations, it is especially important for an AI developer working in the medical domain to know about ethical issues. I learned a lot about data management, duality, ethical review and good ethical practices. It was one of the courses that I liked quite a bit.

Addressing AI ethics questions and AI regulations in 2024

Working with AI has its advantages but also its drawbacks. With the explosion of chatgpt in 2022 onwards, the world was conjointly concerned about ethics and safety, both in the present and in future.

Since I predominantly use and apply AI through my research, the ethics questions needed to be particularly addressed. Since I work with large scale, publicly available data, and with publicly available AI models we made the process of using the tools we develop as transparent as possible. This involved making the code, the data, the models and the results all publicly available. Not only that, we have made the results and output as transparent as possible to make sure people who use the tools can go back and refer to the original data. We also do not use any sensitive data such as patient data in our research.

One potential risk of research misuse that may occur with our results, is that people may take advantage of the findings to find potential treatments without further investigating them. For example, the link between hydroxychloroquine and COVID may be strong due to the amount of research done on the two. However, it is not a drug to treat COVID, as misinformed people found out during the pandemic. The research has such limitations and should not be taken at surface value.

A lot of these ethical concerns were already addressed in the project plan and my role in that was limited. However, further investigations need to be conducted to ensure long term risks are avoided or mitigated.

My research did not involve the use of personal data, humans or biological samples. Therefore, an ethical approval by the Swedish Ethical Committee was not required. However, I learned about the importance and severity of the ethical permit in Sweden first hand. One of our associated COVID projects, in which I was briefly involved with, required approval for sensitive disease related information that the patient shared. While my

supervisor applied for the permit and gotten an approval, the hospital claimed that if the patient mentions a third party (family member, friends etc) their consent cannot be taken. In the end, our project could not start and it had to be scrapped.

This shows how different the ethical regulations are in different countries. In terms of research, it showed me that I needed to be extra careful of the regulations and make sure to apply for an ethical approval. Although I did not apply for one myself, during the research ethics course I learned about the projects of other PhD students and almost everyone needed to apply for an approval, since the majority were biologists or clinicians. In my lab, my supervisor is the one handling ethical questions.

I learned that I have to be particularly careful about GDPR issues and personal data in my research. If we intended to use medical datasets instead of publicly available research papers, we would have needed a permit, like the project that was abandoned.

Although I was not personally involved in research ethics approval applications, I learned that in Sweden the first application should be done through the research committee by filling up their online form that involves questions about the project, institution, responsible parties, type of research, methodology, timeline, data collection and handling, potential risks, risk-benefit analysis, participants, information and consent, finances and more. Also, tying up with the new AI regulations, wider implications and effects of the projects must also be in consideration. With these points in mind, I believe I can write an ethics permit application myself.

10. Career Development

Career Control for Researchers 2020

Right at the beginning of my PhD studies, I took this career development course in October 2020 (online). It was a perfect introduction to career development and self reflection on my imagined career. We had cases and assignments on our career choices. I also met some interesting students during the course and networked.

Career Goals

My career goals have always been oriented towards working in the industry. During my PhD I tried my best to improve upon the skills that are well sought in the AI and software development industry. I have also attended several career events to help me understand and pursue my goals.

Mentlife 2023-2024 - Helen Sjogren

The best thing for my career development during the PhD studies was joining the Mentlife program conducted by Pernilla Carlsson at LU. I received a year long mentorship from Helen Sjogren, Senior Scientist at Ferring Pharmaceuticals. She has been on both sides, academia and industry, and was the perfect guide to prepare me for my future career in industry. With our interactions, I prepared myself for interviews and talked about different aspects of the pharmaceutical industry. We med once every month. The mentorship included but was not limited to: guidance on interview preparation, CV/cover letter writing, focusing on strengths and weaknesses as well as further networking and answering a lot of questions I had about working in the industry. She was also an incredibly pleasant and simultaneously positively critical person and I had a great time exploring different career options with her guidance.

This helped me prepare well for the next step after PhD and I believe I gained a lifelong mentor/friend.

A lot of these ethical concerns were already addressed in the project plan and my role in that was limited. However, further investigations need to be conducted to ensure long term risks are avoided or mitigated.

11. Supervision

Supervision from Sonja

Sonja is a good supervisor. She has created quite a non-hierarchical, supportive and positive environment in the lab and always goes out of her way to solve our issues. There is also a strong sense of collaboration and open source publication in the lab. She has built the base structure of the projects I work in and allowed me to research independently. This approach comes with its own pros and cons. While this provides me a higher flexibility with the methods and research direction, I also need to self-learn a lot and set my own deadlines. So if I run out of motivation or am stuck somewhere I need to find a way to get out of it. Sonja has been helpful and if she couldn't provide me the help I need, she figured out a way for me to get help.

For my work I receive inputs for application, implementation and biomedical domain expertise from Sonja. I also got regular feedback and help from her for my writing. We ideally hold weekly meetings, but as we are a small lab, these are sometimes replaced with meetings based on availability. I personally could benefit from a regular schedule for meetings and research inputs, but the difference doesn't impact our work much.

On the other hand, during the second half of my PhD her lack of availability affected my output quite a bit, since she was away without notice for long periods of time. A lot of the work I had done during 2022-24 was done independently. She also didn't allow us to take any additional online courses to strengthen our knowledge in specific topics where I had to learn everything from ground zero without guidance. This made some project progress slower than I anticipated. She also rejected my (pre-approved funding) application to the HALRIC project grant (mentioned previously) which I believe would have been incredibly beneficial for my research and the lab. That would have also allowed me at least a further 6 months extension for my PhD.

Overall, it was mostly a good relationship with Sonja and in most cases when she was available, I received proper supervision from her. If I reflect back on the 4 years of PhD studies and the supervision I received from Sonja, since I was already in the research space and published two journal papers before starting the PhD, I believe I already knew quite a bit about research and publication and was sufficiently independent. During the first year, I was more reliant on Sonja, but in the latter years, I was working independently on my projects. For example, while in my first year I needed a lot of help understanding the core of the projects. I not only needed help from her, but also from ex-co-supervisor Marcus Klang. Alternatively, in 2024 I devised my own short project plan and received a fully funded invite to Tokyo, without any input from Sonja or other supervisors.

Unfortunately in the lab, my ideas about the projects were not usually well received and most were shut down, even though they were well received during the conferences I attended and the experts I spoke to. I believe in that aspect I haven't grown a lot in terms of creativity and haven't had a chance to work on my own ideas.

Mentor for Computational Expertise

Due to working in the medical faculty and my supervisor being a biomedical expert, I feel the need for a computational mentor, particularly within NLP. There aren't many people working in the field and in this kind of interdisciplinary research I often feel the need for expert help. For the second half of my PhD I intended to actively seek out mentorship.

The Mentlife program also helped me indirectly with research. I learned a lot from my mentor who guided me through the research process in pharmaceutical industries. That was a great help outside of my current domain.

Supervision from Peter Berck (indirect)

After the disinvolvement of Marcus Klang (co-supervisor on paper), I needed some help from a domain expert in my projects. Although he is not an official co-supervisor, Peter's involvement in the project significantly increased the pace of the output. Not only I received his linguistics and NLP expertise but we also benefitted from mutual discussions. Overall, this unofficial supervision was the exact boost I needed with my NLP projects.

My Supervision of Masters and Project Students

I co-supervised three masters students actively and two passively so far. I wasn't 100% actively involved in the projects, but I helped with programming, technical knowledge and manuscript editing. I think I did a fairly average job. I didn't have any exceptional contribution, but I was available for support whenever necessary. With one of my student's (Jacob) projects, I was more hands on involved.

I asked for feedback from all of them. Their reflection was positive and said my approach towards guided independent learning worked for them. They did hope for more social interaction among the people in the lab including my supervisor, but unfortunately that was not in my hands.

12. Administration, Organization and Leadership

Career Control for Researchers 202 Involvement in the Medical Doctoral Council (MDR)

I have been actively involved with the MDR at LU and participated in many of the meetings and social events. MDR holds many activities for PhD and in a great support structure for PhD students. I have learned a lot from the group and intend to stay active during my PhD.

Between 2022-2024 I held the position of representative for the Internationalization Board at the medical faculty through MDR. For this position I attended regular monthly meetings with the faculty board to discuss internationalization matters. It was great to have a voice for the students at this board and be part of important decision making.

At MDR I also organized the PhD day in 2024 where I was involved with many duties including contacting sponsors and vendors for the event and management of event registration etc. It was both an exciting opportunity and a learning experience.

Administrative Duties During Events Organized by My Supervisor

I have also performed administrative duties during multiple events organized by my supervisor, Sonja. I have been a co-organizer and local contact for the ELLIT Focus Period 2022. I have also welcomed Michel Dolsten and his children during his visit to Lund. Such tasks were done voluntarily as part of my PhD experience.

P.S. Thank you Olga for your support and comments to help me improve this portfolio!!

Appendix

A summary of the seminars and conferences for the PhD defense opponents

Here is a table of the highlighted seminars I have attended during the last 3 years of the PhD for the assessment of the opponent. The detailed explanations of the seminars will follow:

Date	Location	Seminar type and title	Presenter	Key Learning	Implementation
2024-05-29	Lund, Sweden	Mini Conference – "Future Biomaterials: Data-Driven Insight to Cell-Biomaterial Interaction"	Multiple. I had poster presentation	Explainable AI and related questions in life sciences, different image segmentation techniques	Made me think more about explainable AI in my research.
2024-03-22	Lund, Sweden	Seminar – "COMPUTE WINTER MEETING with focus on Computatioal scince and sustainability"	Multiple	Sustainability and large language model training	Since our LLM training takes a lot of computational resources, it was nice to think about sustainability more in our contributions
2024-01-15	Tokyo, Japan	Hackathon and symposium – "8th Biomedical Linked Annotation Hackathon and Symposium (BLAH8)"	Multiple, I presented my project and a new concept	Vector databases, AI and drug discovery, context matching with life science terms	I lead my own mini project and received a lot of input from experts about my PhD project and implementation
2023-12-12	Lund, Sweden	LUBI Seminar talk – "Spacial Omics, Image analysis and data integration in mantle cell lymphoma"	Lavanya Lokhande, Phd	Immunotechnology and processing of omics data	She integrated omics data analysis and image analysis which was interesting to learn about. A lot of the topics were still difficult to comprehend.
2023-10-26	Lund, Sweden	My LUBI seminar talk - "EasyNER: Using artificial	Myself	It was fun to talk about my research in	The talk was a bit short. Next time I will keep in mind

		intelligence to "read" millions of articles in life sciences"		simpler terms for the wider audience. I received great feedback and got contacted later to help other researchers implement my tools.	to give a bit more focus to cover the fundamentals
2023-10-10	Malmo, Sweden	PhD defense – "Oxidative stress-related factors in abdominal aortic aneurysm: potential clinical implications"	Sakshi vats	Oxidative stress, free radicals, AAA as well as PhD defense procedures	It was nice to learn about a wider area of science and learn about the defense procedures
2023-10-26	Lyon, France	Conference – ISMB/ICCB conference, 2023	Multiple, I had presented my poster and gave a flash talk	Language models in life sciences, potential applications, text mining, future proof design of models	I learned a lot in this conference to summarize in this small space. But I implemented the knowledge I gained directly into my projects. For example, in the named entity recognition parts.
2023-09-21	Lund, Sweden	Seminar – "Large Language Models and the Art of Using ChatGPT in Academia"	Myself (and Bjorn Svensson, director of studies at LUSEM)	I gave a talk along with Bjorn on applying LLMs in academia and their impact. It was great to have discussions on the pros and cons of LLMs and ChatGPT like models	I learned the other side of academia and how LLMs can impact learning and how it may be difficult for teaching aspects. This gave me a better perspective for teaching and supervision.
2023-02-14	Lund, Sweden	My halftime review seminar	Myself Opponents: Lars Juhl Jensen and Dag Ahren	It went smoothly and received a lot of interesting comments from the opponents. For example, on model selection.	We ended up implementing the majority of the comments directly into my research and focus on larger scale article processing. I also ended up building a strong relationship with

					Lars (and consequently his team)
2023 - 2021 - XX-XX	Online	NLP seminar series organized by AI Sweden – 12 different events	Multiple speakers	AI Sweden has organized online seminar series of top NLP voices in the domain from all across the world. I have attended 12 such seminar online between 2021 and 2023 that included topics from foundation models, scaling, ML and collaboration, context in life sciences, ethics, misinformation etc.	These seminars were incredibly useful for my research as they were directly relevant to my work. I also gotten in touch with several of the speaker like Sampo Pyysalo, Barbara Plank and Magnus Sahlgren to ask specific questions about their work in the field.
2022 -10-27	Lund, Sweden	Seminar and workshop – AI Lund Natural and Artificial Cognition	Multiple	Understanding how animals and machines learn, gestures and linguistics patterns in animals etc	While the seminar was not directly linked to my research, learning about how cognition works and the different research potentials helped me gain a broader perspective into NLP and deep learning research
2022 -10-26	Lund, Sweden	LUBI Seminar – "Single cell RNA sequencing: Beyond gene expression"	Yogita Sharma	Single cell RNA sequencing	I wanted to learn a bit more about single cell data and how to train models to process such data, however this talk was not too relevant and the focus was primarily on the experiments rather than application

2022-10-19	Lund, Sweden	LINXS workshop – biomedical imaging	LINXS researchers	A crash course/workshop on biomedical image analysis included imaging modalities, segmentation etc	Since a part of my PhD project is on imaging, it was interesting to learn a bit more about how researchers implement different imaging techniques
2022-10-12	Lund, Sweden	AI Lund Lunch Seminar - "Fine-Grained Image Classification of Groceries for Assisting Visually Impaired People"	Marcus Klasson	Computer vision based assistive technologies for automatic image classification	Since a part of my PhD project is on imaging, it was interesting to learn a bit more about how automatic classification pipelines are developed for real time applications
2022-10-07	Lund, Sweden	PhD halftime seminar – "Prevalence and contributing factors to breathlessness in the population – a data driven approach"	Max Olsson	Breathlessness, continuous data and ML applications	Breathlessness is an interesting topic and the data Max obtained was continuous, time series data. It was interesting to learn about his findings and compare our relatively static approach with his dynamic approach to data analysis
2022-08-23	Arlid, Sweden	Seminar – "COMPUTE Summer Retreat: Scientific Communication - from article writing to public outreach"	Multiple, I presented my poster on EasyNER	Scientific effective communication, writing, publication	Learning about effective communication strategies and communicating scientific findings was beneficial.
2022-06-29	Lund, Sweden	PhD Half Time Seminar – "Histopathological analyses of tissues to score lung damages"	Iran Augustos De silva	Histopathological scoring, inter annotator agreements	In his half-time seminar he explained the methods he used to obtain the pig lung tissues and how he captured the images. He also gave a description of the variability of

					scores among human scorers/pathologists and why it is difficult for a model to train on such scores. This was good to know for my annotation project.
2022-06-13	Stockholm, Sweden	Conference – "34th Swedish Artificial Intelligence Society (SAIS) Workshop 2022"	Multiple, I presented my poster on EasyNER	Human annotation and validation, attention based multimodal NER etc	The SAIS workshop was one of the first national conferences that I could physically attend after the pandemic. It was also one which was purely AI focused. The first keynote talk from Barbara Plank on human label variation was very interesting and quite relevant to our projects. It was also fantastic to present my poster on my Natural Language Processing pipeline and get great feedback from some of the local experts.
2022-04-19	Lund, Sweden	Focus period of 5 weeks with talks, hackathon, a conference and collaboration – ELLIT focus period Lund 2022	Multiple, I gave a talk about my research, presented a poster and was a member of the organizing committee	Working on analysis of real patient data, generating fake patient data using GAN, developing a cancer prediction model etc	The experience of 5 weeks was fantastic. Not only did I get to work with leading scholars in the field, but I also got a chance to work with real data. I also got a lot of input on my research methods.
2022-03-15	Lund, Sweden	Expo/ symposium – "Correlative	Multiple	Language models and their use in	From one of the speakers, Alexandros

		Image Processing and Analysis (CIPA) Expo 2022”		imaging applications	Sopasakis, it was interesting to learn about domain knowledge and adding expertise to model training. I implemented that in my NLP projects later on
2022-03-09	Lund, Sweden	Seminar – “COMPUTE winter meeting - Crossing scientific borders”	Multiple	Doing interdisciplinary research and communication between different areas	This meeting was the first physical COMPUTE event I attended after COVID. It was interesting to learn about interdisciplinary research like mine and how effective communication is key to successful research projects
2022-03-08	Online	Seminar – “Synapse and HUB AI Presents: Life Science in AI”	Multiple, I co-hosted the event	Application of AI in life sciences and how researchers, and students may benefit from it. It was a mix of scientific and popular science talks	It was nice to organize such events to communicate our science to the wider society and share the knowledge.
2021-10-20	Online	Conference – “Swedish Bioinformatics Workshop 2021”	I presented my poster and gave a 3 minute flash talk	It was my first public talk during the Pandemic. It was nice to share the work I had done so far.	I didn’t receive much input during my talk or the poster. The pandemic lessened the interaction aspects of online presentations quite a lot. But still, it was good to present my work.

Project Proposal (Draft) for the HALRIC Project

Attached is the first draft for the proposed HALRIC project mentioned previously. The project proposal was not submitted due to the conflict of interest with my supervisor.

Title of the project:	Biomedical information extraction using language models		
Start date (yyyy-mm-dd):	2023-11-01	**End date (yyyy-mm-dd):**	2024-05-01

Abstract (maximum 1500 characters with spaces – will be public):

The exponential growth of biomedical literature every year makes extracting key information through manual reading more and more difficult for researchers. Meanwhile, large biomedical text collections such as PubMed and EuropePMC are unexplored to a great extent. With the advancements of artificial intelligence and language models in recent years, information extraction from such large text collections can be made more accessible to researchers who aim to explore specific research topics within different biomedical domains, in particular cell lines, gene/protein, disease, drugs/chemicals and species as well as the metabolic relationship among them. In this project we aim to explore such large text collections to extract domain specific information and relations from them. As a pilot, we aim to explore the large collection of preprint (unpublished) articles available at EuropePMC. We follow our findings by exploring the relationships among the domain specific entities including previously unknown/novel links. We aim to follow this with building a knowledge graph and map the metabolic relationships among different diseases, drugs, species, cellines etc. We expect to find novel relationships of drug-disease, drug-drug, species-disease etc.

Aim of the project (maximum 300 characters with spaces):

The aim of this pilot project is to develop an artificial intelligence- and language model- based toolkit to process large text collections, starting with EuropePMC. The primary goal of the toolkit will be to extract biomedical domain specific entities and the relations among them from literature.

Background (maximum 3500 characters with spaces):

Biomedical literature has grown exponentially over the last few decades. In academic and commercial research, we often perform literature review of a handful of papers in order to gain insights into a specific area of knowledge. However, for a single researcher it is extremely difficult, if not impossible, to access the vast knowledge hidden within the millions of research articles that are published every year. Deep learning -based, text mining models has solved this problem to a large extent. These models can be used to extract specific information from a body of text by extracting entities and connecting them [Citations on BERT, BioBERT, BioGPT and EasyNER]. In life sciences, this is particularly useful for collections like PubMed, where a researcher can use such tools to extract information and explore different ideas. While there has been a number of extensive research projects on PubMed article collection, a large collection of non-peer reviewed papers has been overlooked – the preprints. These are the papers that are hosted in a draft format, but are not peer-reviewed (yet). Due to the duration of the peer-review process, it is extremely important to look into the preprint collection to extract the latest available knowledge for the researcher. However, such an initiative has not been attempted yet.

Furthermore, an understanding of the knowledge space and interaction of the specific entities within this space is particularly interesting. Not only for finding information, but also to extract novel information and connections within the existing space. Therefore, an end-to-end solution for exploring this information is not only vital for our understanding of this space, but also important for researcher who do not have the expertise to develop such models themselves. For example, the drug-disease interactions or the metabolic landscape within this space of non-peer-reviewed articles and comparison with PubMed can help us compare and contrast between large text collections and understand specific patterns that we may not be aware of.

We aim to develop and use Large Language Models in order to extract and collect information from preprint servers such as EuropePMC and develop knowledge graphs to enhance our understanding of this space for this pilot project. From our understanding, we aim explore further into specific functions or interactions within different

entities. We will also make use of the different expertise of the researchers involved in this project to diversify our understanding of specific interactions and evaluate our findings on ground truth (ex. Gold standard dataset or known interactions). At the end of this project, we aim to publish 1 paper to summarize the findings.

Work plan (maximum 4000 characters with spaces):

Remember to think of the six criteria.

1. Exchange of knowledge with EuropePMC and EBI
2. Access to High Performance Computing infrastructure in Norway
3. Help from Kajsa and Sonja

Information about participant 1:	
Name:	Rafsan Ahmed
Affiliation:	Lund University
Role in the project:	Primary Investigator
% time in project:	100%
E-mail address:	rafsan.ahmed@med.lu.se
Information about participant 2:	
Name:	Sonja Aits
Affiliation:	Lund University
Role in project:	Supervision
% time in project:	10%
E-mail address:	Sonja.aits@med.lu.se
Information about participant 3 (insert more rows if more participants):	
Name:	
Affiliation:	
Role in project:	
% time in project:	
E-mail address:	
Information about the industry partner (if applicable):	
Name:	
Affiliation	
Role in the project:	
E-mail address:	

Motivation of the composition of the cross-border research team:

Information about the financial officers for each participating organization:

Name	Affiliation	E-mail address

Describe how and when to get access to the research infrastructure(s) (e.g. beam time):

Describe how you will work with the following sustainability goals (you need to contribute to minimum one criteria):	
Environmental sustainability (max. 400 characters with spaces):	
Equal opportunities and non-discrimination (max. 400 characters with spaces):	
Gender equality (max. 400 characters with spaces):	

For the management team:

Received by:		Date:	
Approved by:		Date:	