Why this topic?

Slide 1:

The idea for this project came from a gap I encountered in my Honours research, where I investigated the shared etiology between autism and migraine. I performed two complementary analyses.

First, I identified 63 overlapping genes from published GWAS studies, which is a method that identifies variants associated with a phenotype.

I explored their function annotations using **Gene Ontology or GO**— a widely used resource that organises gene functions in a hierarchical structure. GO terms are incredibly detailed and precise, describing processes like “hexose biosynthetic process” or “monosaccharide metabolic process.” While useful, they’re not always intuitive — especially when you want to quickly summarise biological meaning across multiple genes.

Separately, I ran a protein-protein network analysis on all genes associated with any of both conditions, and I filtered the results to only keep cross-condition interactions to be able to compare the results between these and interactions that were found in condition specific analyses. I also allowed the addition of intermediate proteins in the autism-only or migraine-only networks to ensure fair comparisons. Then I conducted an enrichment analysis to understand what biological processes were overrepresented in my network for each case. At this stage, my goal was to understand how condition-specific genes might influence broader pathways when present together — going beyond just overlap.

This is where I saw the gap. Even though GO is hierarchical, the terms remain highly granular — and researchers are still left interpreting crowded visualisations like *cnetplots* or relying on plotting just top functions to simplify it, losing relevant information. These tools don’t really help when you're working across conditions, with large gene sets, or exploring system-level patterns.

So I started wondering: what if we had a tool that could take these precise annotations and automatically group them into higher-level, intuitive category levels corresponding to groups such as “innate immune response” , or even a broader one “immune system”? Something that simplifies interpretation and helps others understand your conclusions without losing detail.

When analysing gene lists, I found that existing tools provide useful annotations and enrichment terms but fail to classify gene functions into intuitive, higher-level categories like “innate immune system” or even a wider level such as just “immune system”.

Researchers are often left manually sorting through complex outputs or relying on plots that display only the top-ranked terms, which risks losing important biological context. And when we try to visualise the broader picture (like in plot at top left of slide 1) the result can be both visually overwhelming and biologically too granular. This level of detail can make interpretation difficult, especially for scientists without a background in molecular biology. It also complicates direct comparisons across conditions or organisms, as researchers often need to interpret multiple dense plots with granular information side by side to identify differences, making it harder to understand what the results mean in a broader biological context.

Slide 3:

To me, this topic is exciting not just because of what it can achieve, but because it fully aligns with the kind of research I believe in.

First, it reflects my core research values. I strongly advocate for open science and FAIR principles — making research Findable, Accessible, Interoperable, and Reusable. This project embraces those values by aiming to be an open-source tool, with publicly available code and documentation. It's designed to work with any annotation dataset, even new or custom ones, making it adaptable and future-proof. And most importantly, it makes omics research more accessible — not just across technical systems, but across scientific backgrounds too.

Second, this project brings together the things I love most: bioinformatics, artificial intelligence, and biomedical research. It sits right at the intersection of these fields and gives me a chance to apply what I'm learning in a meaningful way.

And finally — it's fun. I genuinely enjoy working on this topic. It keeps me curious, motivated, and eager to keep learning.

So when I think about this PhD, I don't just see an academic opportunity — I see a chance to build something impactful, to grow as a researcher, and to contribute to science in a way that's open, collaborative, and driven by real-world needs.

Engagement with this topic

Slide 4:

As you've already seen, this idea came from a real gap I experienced in my own research. So once I realised the current tools couldn't meet that need, I started thinking seriously about how I could design a strategy that could do the job.

During my project, I also became familiar with functional annotation tools and learned to work with the Gene Ontology system. And my background in biomedical sciences gave me a strong foundation in understanding gene and protein function across multiple biological levels — from cellular pathways to whole-body systems. This gave me hands-on experience with how complex and granular these annotations can be, and how much interpretability may be difficult to scientists without this background. I additionally took courses on neuroscience to be better prepared for my Hons-based research.

At that point, I prototyped a manual strategy to group functions myself using chatGPT-4o to automate the classification, with manual review. It worked for my relatively small dataset and gave me a better visual summary for my conclusions, but it clearly wasn't scalable nor optimal.

"To develop this tool, I knew I needed to start building a solid basis in Artificial Intelligence. So I began by studying machine learning, which involves essentially computer computers that learn patterns from data and make predictions or decisions without being explicitly programmed for every scenario, also called models. I thought this approach was promising for my tool, since classifying biological functions involves identifying subtle patterns and relationships that aren't always obvious.

To build my understanding, I completed introductory courses in data science and machine learning from IBM, and I'm currently enrolled in a supervised ML course on Coursera. I'm also studying the book *An Introduction to Statistical Learning with R* to deepen my grasp of the statistical principles behind these models.

As I progressed, I realised that conventional machine learning might not be enough, especially since gene functions are described in complex biomedical language. That led me to explore deep learning, which is a subset of machine learning that uses networks of interconnected units, similar to how neurons work in the brain, to learn patterns in data. These *neural networks* can process many layers of information at once, allowing them to capture complex relationships that would be hard to define with simple rules, like how different biological processes relate to each other based on the way they're described. It is particularly effective in tasks involving text.

That's when I started learning about text tokenization, which is the process of breaking down sentences into smaller units, like words or subwords, so they can be analyzed by a model. And from there, I discovered BioBERT.

BioBERT is a deep learning model that's been trained specifically on biomedical literature. It takes biomedical text, tokenizes it, and transforms it into *embeddings*, which are numerical representations that preserve the context and meaning of the original text. These embeddings make it easier to detect functional similarities even when the wording is different.

To make sure I was on the right track, I met with professors in the field, shared my approach, and received feedback to make sure that this idea was realistic.

This approach opened up a whole new range of possibilities for the tool, allowing it to generalize better and recognize biological context, not just keywords.

I have also taken time to review other existing tools like DAVID and PANTHER, and browsed through GitHub projects with related aims. While many of these are useful, none provide the kind of self-learning, flexible classification system I'm hoping to build — one that can handle biological complexity while remaining easy to interpret. To refine my thinking, I've met with professors from different fields to get their input. Their feedback has been really constructive, and helped me clarify not only how this tool could work, but what I still need to learn to make it happen.

Next steps

Slide 5:

One of the key challenges in this area is the lack of labeled data. There's currently no dataset that organizes biological functions, meaning both gene functions and biological processes, into clear, high-level categories like 'Immune System' or 'Cell Cycle.' This makes supervised learning difficult. Manual labeling would be time-consuming and prone to individual biases. Implementing a system where users can confirm or correct the labels would be a way to address this issue, essentially making it a community-supported effort.

Another important step is choosing the right model. Traditional machine learning models, like Random Forests, are easier to interpret but don't capture the complexity of language. Deep learning models are more powerful but can be harder to set up and adjust properly. BioBERT sits in the middle, which is a model that has already been trained on biomedical texts, so it understands the kind of language we use to describe biological functions and processes. I can use it in two ways: either to convert those descriptions into meaningful numeric representations, which I then feed into another model to make the final classification, or I can adapt BioBERT to do the classification itself. I think it is best to test both options and compare their performance using reliable evaluation methods.

A third challenge is that many biological functions belong to multiple categories, and these categories can be hierarchical, like a process related to the same level categories innate immune response and acquired immune response, and then to the higher category immune system. One option would be to train separate models for different classification levels, such as cellular function and bodily system. But this approach could be less efficient and harder to maintain. Instead, using a model that supports

multi-label and hierarchical classification would allow it to learn those relationships more naturally and make more consistent predictions.

Generalizing to new or unseen descriptions is another major concern. Biomedical annotations are often phrased differently depending on the source. A model that can't handle new inputs wouldn't be practical. That's where BioBERT is especially helpful, because it's pretrained on biomedical language and it can generalize to varied inputs, which improves reliability.

Finally, interpretability is essential. We need to understand why a description was classified a certain way, especially in research contexts. It would be interesting to integrate explainable AI tools to make model decisions more transparent.

All of these challenges are critical because they shape whether the tool will be usable, reliable, and ultimately helpful to researchers navigating a sea of annotations.

PhD aspirations

Slide 6:

This research topic is at the core of what I hope to achieve in my PhD.

This is a gap I encountered in my own research, and it has clear relevance for bioinformatics, systems biology, and multi-omics analysis. But as we saw it also has many challenges. These are exactly the kinds of challenges that benefit greatly from collaboration — from feedback, mentorship, and shared expertise. That's why DiveIn is such an ideal place for this PhD: it brings together interdisciplinary training, community support, and hands-on learning that would make this tool possible.

Slide 7:

Through this PhD, I want to grow as a researcher and contribute to work that makes a difference. I'm excited to develop a tool that supports better interpretation and reproducibility in omics research, but also to deepen my expertise in AI for biomedical applications. What makes this opportunity even more meaningful is DiveIn collaborative environment: I really value being part of a research community where I can work closely with others, learn from different perspectives, and strengthen my skills in communication, mentoring, and inclusive teamwork.