Harnessing Human Intelligence for Biodiversity Informatics

Item 6. Research vision

I. Introduction

I believe that the two planetary-scale changes that humanity is facing in the 21st century are human-induced biodiversity loss and exponential growth of compute-capabilities. They may not seem linked at the first glance, but I think we can *develop human- and Earth-centered information technology* to help avert the ecological catastrophe.

Biodiversity informatics is the science that holds the key: only when we have described and cataloged the species on Earth and their interactions, can we create realistic models that help us understand the potential impact of protective measures that humanity may undertake. A successful approach in biodiversity informatics comprises of two facets: (a) a data aspect—creating and, more importantly, interlinking primary biodiversity data on species occurrences, interactions, genomic data, etc. and (b) a modeling aspect—using computational models to understand how biodiversity changes with time and interacts with its environment both in the past and in the future.

In the data aspect, the challenge is that the task of describing biodiversity is not finished, let alone databasing (cataloging) it and properly interlinking it. Ultimately, we cannot hope to do these steps manually and in this sequence on time to avert catastrophy, but should rather opt for an integrative approach that describes and databases and interlinks simultaneously, and uses as much automation as possible. In the modeling aspect, the challenge is to overcome the current crisis in reproducibility in science, which is evidenced by numerous studies indicating that a significant portion of scientific research is not replicable. This reproducibility crisis stems from various factors including poor data, but also pressure to publish positive results, which often lead scientists to try models that they do not fully grasp. The second factor can be combated by better modeling frameworks such as probabilistic programming languages.

II. Habsburg AI

The major scientific insight that I would like to highlight and inform this research proposal is known as the Habsburg AI. Philosophers of science since the 1980s have speculated that given exponential improvements in computing, machines of super-human level intelligence will appear, and once they appear they will start improving on themselves even faster leading to an event known as "singularity." In reality, however, we observe that even though AI systems of the current generation outperform humans on many tasks and show "sparks of general intelligence," they are not capable of self-improvement, and worse, they deteriorate in capacity if trained on self-generated data[1]. This technological "inbreeding" has been called Habsburg AI.

The implications of Habsburg AIs for the field are that collective *human intelligence* is what makes AI systems smart and that AI systems need to maintain their expertise by relying on input by humans who direct and curate large datastreams. So for now, *human expertise cannot be automated away*, but the focus of human researchers, even in fields such as biology will be shifting more towards human-computer interaction.

III. The many cultures

In my experience as an early-stage researcher I have been convinced that there are many cultures within the science of biology: traditional taxonomists, ecologists and evolutionary biologists, people with mathematical but not with computational backgrounds, computational and machine-learning and data-driven researchers. Representatives of the three cultures often work together but have considerable difficulties communicating and exchanging ideas because as C. P. Snow points out they each have their own climate of thought and intellectual approach. Due to my unusual background combining experience from all three positions, my intention is to bridge the many cultures gap and to find ways to encode taxonomic, ecological, and evolutionary expertise into AI systems.

For this reason, my research vision is to *harness human intelligence for biodiversity informatics* by (a) creating state-of-the-art intelligent (semantic) databases of biodiversity knowledge who index and understand vast amounts of biodiversity literature, and by (b) enabling biologists to use probabilistic programming languages (PPLs) to create cutting-edge statistical models of ecology and evolution.

IV. Research packages

As an early state scientist I have had two research focal points: (1) semantic biodiversity databases, the subject in which I wrote my dissertation and (2) probabilistic programming languages for statistical phylogenetics, the field in which I have spent the bulk of my postdoctoral research. On the basis of these starting points, I propose several research directions, encapsulated in two Work Packages (Gantt chart available under Fig. 1)—WP1: Semantic biodiversity databases and WP2: Probabilistic programming languages

[1] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk. Self-consuming generative models go MAD. arXiv:2307.01850, 2023.

for computational biology—to ultimately *harness human intelligence in the service of biodiversity science*.

WP 1: Automating biodiversity knowledge-base construction using large language models and probabilistic programming languages (Semantic biodiversity databases)

Semantic databases, such as OpenBiodiv, are "intelligent" databases which encapsulate complex knowledge about the world encoded in schemas called *ontologies*. The strongest suite of a semantic database is its ability to mash heterogeneous knowledge together and give fact-based answers to a complex queries such as e.g.: "What are the primary pollinators of plant species found in tropical rainforests that have seen a significant population decline in the last decade?" Unlike a chat-bot that may hallucinate an answer, the semantic database will always produce a "true" answer at least in the sense that its derivation will be traceable to meaningful objects in it (in the example—taxa, ecological relationships, geographic data, etc.). Therefore, the construction of such databases on quantities of data similar to those that chat-bots are trained on is paramount; unfortunately, the manual ontology creation is an extremely labor-intensive process and does not scale well. However, the same large language models (LLMs) that are used to power chat-bots can be leveraged to construct a *factual semantic database* in an automatic way—an early example is OntoGPT. Concurrently, another cutting-edge AI technique also showed promise in automatic knowledge-base construction, namely probabilistic programming languages (PPLs, also see WP 2)—an early example of such an effort being Project Alexandria. Thus, I see opportunities to revolutionize automatic knowledge base constructions in the domain of biodiversity informatics by leveraging one or both of these state-of-the-art technologies to *create a knowledge base of biodiversity information that would encompass nearly all of humanity's expertise*.

In this WP, the proposed research directions are (1) to develop foundational methods leveraging LLMs (similar to OntoGPT) and/or probabilistic programming languages (similar to Project Alexandria) for knowledge base construction and interrogation and (2) to involve biodiversity experts (taxonomists, ecologists, modelers) in evaluating the performance of these databases in a feedback loop leading to better foundational models.

In the Gantt chart on Fig. 1, I suggest to hire one PhD student to lead this effort, who will collaborate with two different postdoctoral researchers (one focusing on PPLs and another focusing on biological applications) during the course of her/his PhD study.

WP 2: Leveraging PPLs for efficiency gains in mathematical biodiversity modeling

PPLs are a revolutionary technique allowing to (a) separate phenomenological model description from statistical inference (which is provided automatically by a compiler) and to (b) account for uncertainty in data as well as in model via a principled Bayesian way. As the lead author of the novel probabilistic programming language TreePPL I am in a perfect position to spearhead the application of these methods in biodiversity modeling. The aim of TreePPL is to bring methods developments originating in the computer science and in the machine learning communities to the computational phylogenetics and to the broader biodiversity informatics communities, as well as to create new methods specific for our domain.

The first research direction within WP 2 is *excellent statistical methods for automated inference in computational biology*, abbreviated to "Advanced inference with TreePPL" in the Gantt chart. As a fairly advanced topic I intend to have a PPL-oriented postdoctoral research lead this direction. The goal is to create novel statistical procedures that could be incorporated into the automated repertoire of TreePPL and thus enable a larger space of previously inaccessible models to be attacked. Potential algorithmic breakthroughs may come from leveraging hybrid Monte Carlo methods, variational inference, or advances in compiler design, as well as from general machine learning. I propose to collaborate closely with lab of F. Ronquist at the Swedish Natual History Museum, who is interested in the application of such models to real-world problems in host-parasite evolution, diversification, online tree inference, and species circumscription.

The second research direction in WP 2 is the *democratization of automatic statistical biodiversity modeling using PPLs*. In the Gantt chart, these are tasks marked with orange and will be lead by the second PhD student with support from both postdoctoral researchers. The goal of this research direction is to develop high-end language features as well as PPL-based modeling workflows specific to biodiversity science (syntactic constructs, semantic editors, development tools and practices, publishing workflows, data standards), which enable practitioners without a background in computational statistics to write complex phylogenetic models leveraging large data.

Conclusion

The interplay between biodiversity science and advanced AI techniques offers possibilities for significant advancements in our understanding and preservation of Earth's diverse life forms. This five-year plan, through its focus on semantic databases and probabilistic programming languages, aims to automate knowledge-base construction and elevate mathematical modeling in biodiversity research. By collaborating closely with domain experts, the proposed AI-driven research remains rooted in biological realities. Ultimately, this approach seeks to merge humanity's vast knowledge pools about biodiversity with cutting-edge technologies, creating a foundation for proactive and informed conservation strategies in the future.

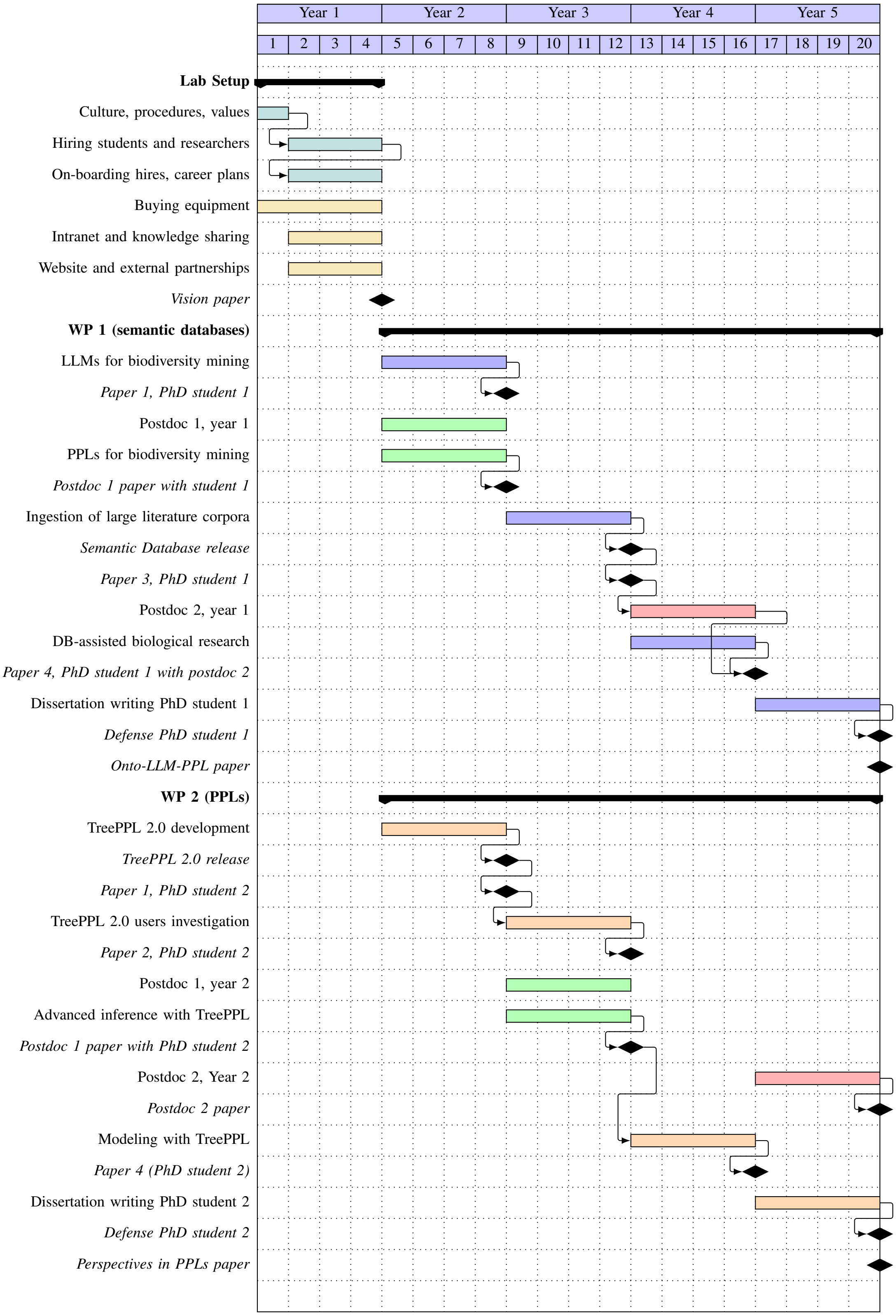

Fig. 1. Gantt Chart showing project timeline.