1. Research statement: Major unsolved problems in Biodiversity research

1.1. Summary

We are in a enthralling scientific era. We have the computer power, the open-source tools, and the team capabilities to break down the disciplinarity barriers to integrate Earth science and Biodiversity research. We are in a period where novel analytical methods and data are being fussioned at an incredible speed for first time to decipher the complexity and feedbacks between the Earth system and the diversity of life. Yet, we are in a massive human-driven biodiversity extinction with large uncertain consequences for Earth climate, life conditions and the stability of Earth (Figure 1). This combination of an enthralling scientific era and rapid global change put us in an edge to take in science the necessary risks to reduce the uncertainty related to the consequences of feedbacks between the Earth system and Biodiversity. For this to happen, we must team up to take the necessary steps to 1) break down the disciplinarity barriers, and 2) fussioning data-analytics and process-based theory to create synergies between predictive and understanding power.

During my scientific career, I have pursued these two main goals: fussioning modern data analytics and theory in Biodiversity research, and fostering synthesis and interdisciplinarity. First, Biodiversity research has been sistematically studied at only one biological level and splitted in many temporal and spatial scales. While this has produced an immense gain in detailed knowledge at each of the levels and scales studied, it might be insufficent to understand the consequences of biodiversity decline in predicting the outcome of feedbacks between Earth system and the diversity of life.

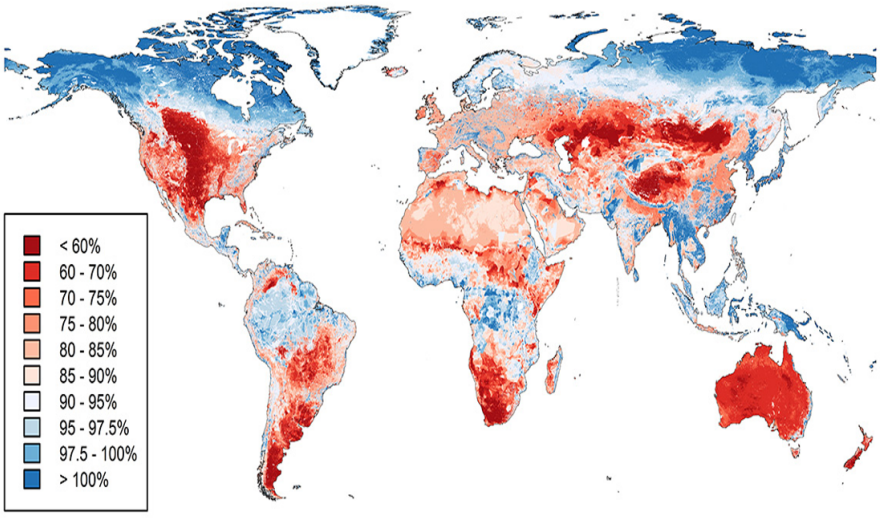

Figure 1: **Biodiversity is declining globally at unprecedented rates**. This map shows the remaining populations of native species across many taxa as a percentage of their original populations. Blue areas are within proposed safe limits, and red areas are beyond these limits (http://www.nhm.ac.uk/discover/news/2016/july/biodiversity-breaching-safe-limits-worldwide.html.)

We have recently developed a framework to facilitate data- and process-based integration to explore the interdependencies among levels and scales in ecological and evolutionary networks (Figure 2, Ref. #30 document *Publications.Funding.Melian.pdf* [1].) Second, interdisciplinary and synthesis engagement is needed to build teams with the skill set integrating predictive and understanding power beyond the boundaries of scientific disciplines (Figure 3). Below I provide my view of the major unsolved problems in merging data science and biodiversity research.

1.2. Towards deep process-based learning in Biodiversity research

Most methods in machine learning and ecology and evolution have been considered classically as distinct fields. However, the current scientific ecosystem is at the stage where merging methods from distinct fields is radically transforming the discipline boundaries, the reproducibility of science and our predicting-understanding power[2]. For example, recent approaches in ecology and evolution have introduced deep learning methods for labelled data, from which selection modes and demographic history can be jointly inferred[3]. The nature of biological data is large heterogeneity and a mixture of labelled but also unlabelled data. From one side, there are large databases with labelled DNA sequence or gene network expression data from which deep learning methods can be used to jointly infer selection modes, demographic histories and range dynamics. On the other side, there are many databases with unlabelled ecological data like the patchy distribution of many unidentified species ranges with the corresponding uncertainty associated to quantifying functions like CO_2 sources and sinks, for example. This creates many uncertainties and challenges the finding of sufficient enough labelled data for training a machine learning system from which species distributions, range dynamics, ecosystem functions and species interactions can be predicted for many species across broad spatiotemporal scales.

Many of the recent approaches applying deep learning methods in ecology and evolution have mostly focused at one level of biological organization. While this might produce additional gain in detailed knowledge at each level, it remains unknown how many layers are going to be needed for predicting and understanding the existing biodiversity patterns. Therefore, the one-level and one-scale approach might be insufficent to understand the consequences of biodiversity decline in predicting the outcome of feedbacks between Earth system and the diversity of life. To gain predictive and understanding power in ecology and evolution we are going to need to

[1]Melián, C. J.; Matthews, B.; de Andreazzi, C. S.; Rodríguez, J. P.; Harmon, L. J.; Fortuna, M. A. (2018) Deciphering the interdependence between ecological and evolutionary networks, *Trends in Ecology and Evolution*, 33:504-512

[2]Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*. 566:195-204

[3]Sheehan, S., Song, Y. S., (2016). Deep learning for population genetic inference. *PLoS Comput. Biol.* 12:e10048452

build hybrid deep process-based learning methods accounting for many layers and the topology of the interactions within and between the layers. Indeed, many methods from data science and biological systems share fundamental properties and the full potential of these shared properties have not yet been explored. Biological systems are composed by many layers (Figure 2), and they can contain interdependent hierarchies and feedbacks with interacting learning entities within and also between the layers. Both deeep learning networks and biological systems can be represented with nodes and links for each layer and the connections between layers can also be explored to analyze the dynamical and topological properties of these multilayer deep-process based networks.

In this setting, and in addition to the limited training set contained in many biological and ecological databases, biological and ecological multilayer networks can be trained or explored integrating datasets from many sources. This creates opportunities to infer how the real world ecological systems might be predicted by process-based interactions from complex traits to non-linear ecological models accounting for interdependencies and feedbacks between levels. There are going to be at least two big group of questions consequence of the fussion between deep learning and multilayer biological networks. Methods driven questions focused in the structural and dynamical properties integrating deep learning networks and multilayer networks. Applied driven questions like inferring future Biodiversity trends projections under different deep process-based learning networks scenarios. Both types of questions would require to explore gradients combining predictive and understanding power to jointly infer the processes and the patterns that can be interacting to produce specific dynamics and topologies. In the applied side, such outputs will produce likelihood scenarios for future biodiversity declines and its consequences for Earth climate and stability (Figure 3). In summary, integrating deep learning and multilayer biological networks accounting for processes within each of the layers, their interaction effects within and between the layers and the effects on biodiversity dynamics and ecosystem functions is full of open challenges and also opportunities to advance our understanding of multidisciplinary data science and biodiversity research.

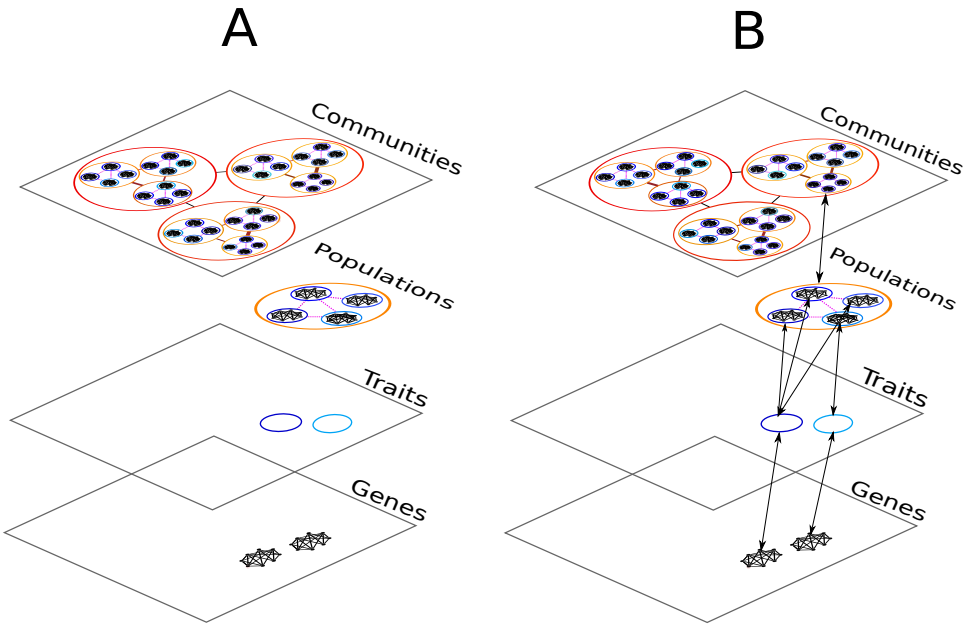

Figure 2: **Biodiversity is hierarchically structured** yet inferring interdependencies among the levels developing hybrid deep-process based learning approaches to predict the consequences of biodiversity decline remains poorly studied. **a)** Biodiversity has been studied mostly considering independent levels, from genes, traits and populations to communities and ecological networks. **b)** Biodiversity represented as interdependent levels accounting for feedbacks from genes and traits, and from traits and populations to communities. It remains unknown which of these two scenarios best predict current trends in Biodiversity decline and its consequences for Earth climate, life conditions and the stability of Earth.

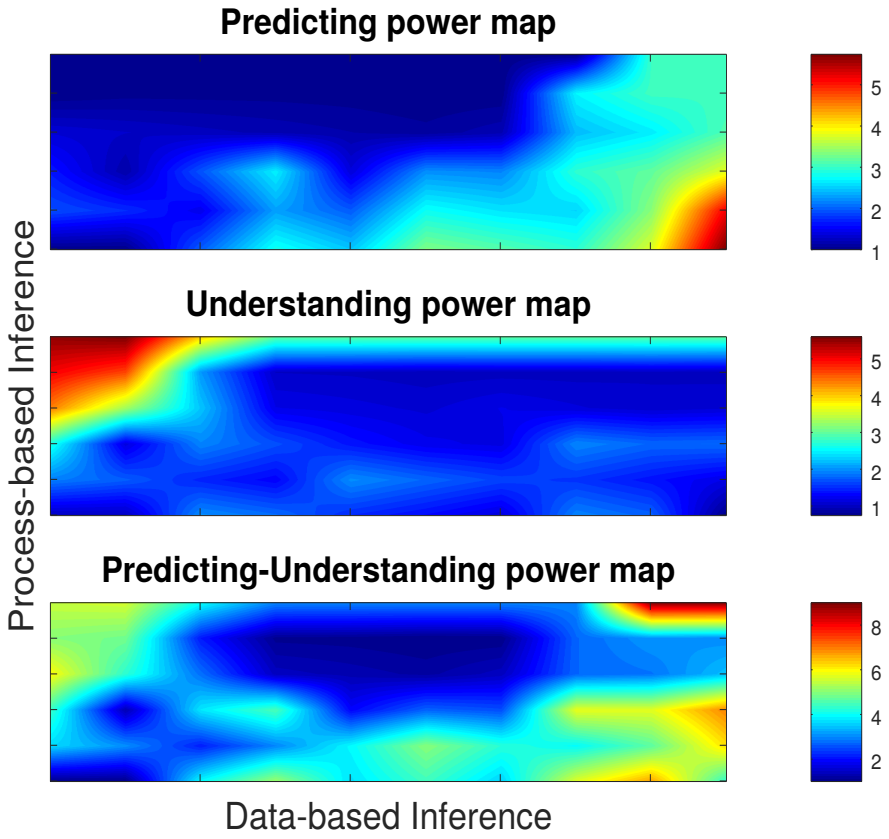

Figure 3: **Interdisciplinarity and synthesis in science** will be needed to join pedicting and understanding power in deep process-based learning networks. This figures shows a cartoon of a predicting power map (top), the understanding power map (middle), and the predicting-understanding power map (bottom). x- and y-axis represent data-based inference (i.e., gradient of AI methods from low (left) to high (right) predictive power) and process-based inference (i.e., gradient of process-based methods from low (bottom left) to high (top left) understanding power). The gradient of predicting power map (top) shows a hot spot red area in the bottom right highlighting the region where AI methods best predict the empirical data. The gradient of understanding power map (middle) shows a hot spot red area in the top left highlighting the region where the best mechanistic understanding occur. The predicting-understanding power map (bottom) shows the sum of the two previous maps highlighting a red hot spot where the best synthesis and interdisciplilnary research joining predicting and understanding power of the empirical data occur.