Title: Deep process-based learning networks in Biodiversity research

Summary

We are in a enthralling scientific era. We have the computer power, the open-source tools, the know-how in many highly specialized fields and the team capabilities to integrate Earth science and Biodiversity research in open and decentralized automated research platforms. We are in a period where novel analytical methods and data are being fussioned at an incredible speed 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 team up to go beyond our disciplinary boundaries to contrast scenarios accounting for feedbacks between the Earth system and Biodiversity (Figure 2). For this to happen we need to connect fundamental and applied science (Figure 3b) and one way to do it is throughout distributed open research platforms to provide informaton for management forums in applied conservation and sustainability centers. This proposal aims to develop a distributed open-source automated research platform to integrate multiple databases into Biodiversity dynamics and function scenarios taking into account the interdependencies among biological levels and scales in ecological and evolutionary networks (Box 1 and Figures 3 and 4).

Deep process-based learning networks in Biodiversity research

Biodiversity is declining globally at unprecedented rates (Figure 1). Despite the importance of biodiversity for persistence of all species, including humans (REF), we do not yet understand the interactions between biodiversity and Earth system - physical, biological, and chemical properties of the Earth as an integrated system. Biological systems are composed of many layers, and they can contain interdependent hierarchies and feedbacks with interacting learning entities within and between the layers (Figure 2). Understanding interactions between biodiversity and Earth system requires analysis of life at different levels, from genes to individuals to populations and to communities (Figure 2). There is an immense detailed knowledge at each of the levels and scales studied in biodiversity research. Yet, such data and knowledge is not sufficiently integrated to provide a holistic understanding of feedbacks between biodiversity and Earth system. We aim to creating an automated research platform that integrates different pieces of data, information, and knowledge of biodiversity research from different levels into a single framework.

Despite the rapid development of automated research platforms integrating different aspects of scientific cycle [7, 2, 1, 6, 4, 5, 3], distributed open-source automated research platforms in biodiversity research are still at an incipient stage (REF).

One reason of still being at an incipient stage is that most methods in biodiversity research have been considered classically as distinct fields. However, the current scientific ecosystem is at a stage where merging methods from distinct fields is radically transforming the discipline boundaries, the reproducibility of science, and our prediction/understanding power [11]. Many of the recent approaches applying deep learning methods in ecology and evolution have mostly focused at one level of biological organization [13]. While this might produce additional gain in detailed knowledge at each level, it remains unknown how many layers are needed for predicting the consequences of feedbacks between Earth system

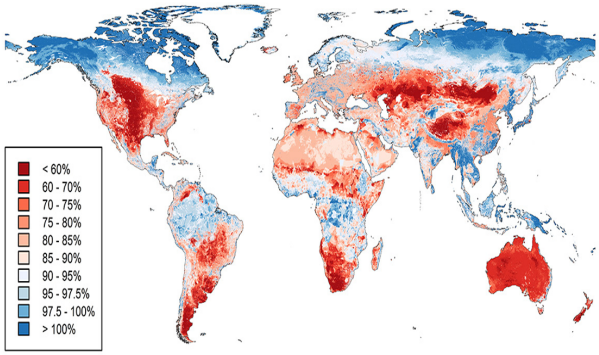

Figure 1: **Decline of biodiversity across the globe**. Map showing the remaining populations of native species across many taxa as a percentage of their original populations. Blue areas are within proposed safe limits, and red areas are beyond these limits. For further information please check the original work at `http://www.nhm.ac.uk/discover/news/2016/july/biodiversity-breaching-safe-limits-worldwide.html`.

and biodiversity.

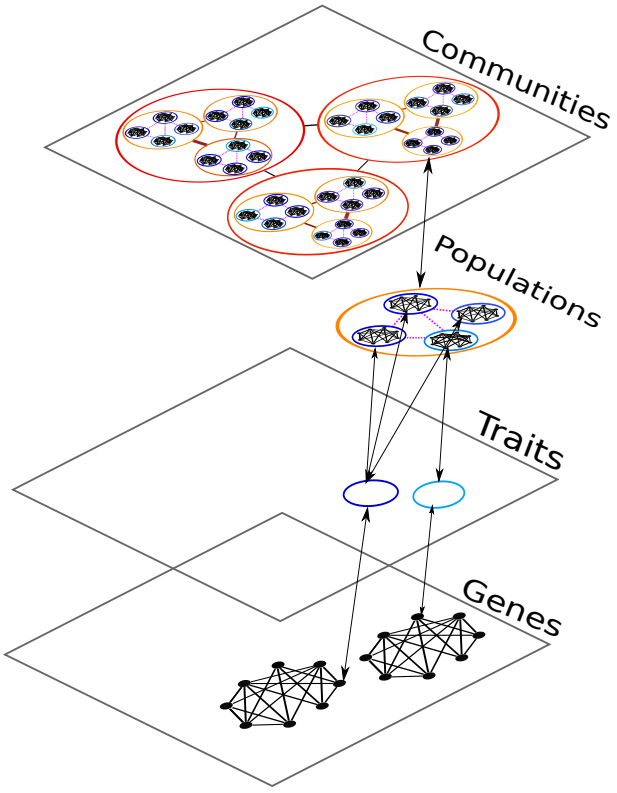

Figure 2: **Biodiversity is hierarchically structured** yet inferring interdependencies among the levels developing hybrid deep-process based learning approaches to predict the consequences of biodiversity decline remains poorly studied. A) Biodiversity has been studied mostly considering independent 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.

To gain predictive and understanding power in biodiversity research we are going to need to merge distinct databases into hybrid deep process-based learning methods accounting for many layers and the topology of the interactions within and between the layers [10]. Many methods from data science and biological systems share fundamental properties (i.e., network-like patterns, multiple layers, etc). Yet, the full potential of these shared properties have not been sufficiently explored. We will integrate different biological layers into a platform to explore contrasting scenarios of Biodiversity dynamics accounting for interdependencies and feedbacks within and between layers (Box 1 and Figures 3 and 4).

Box 1. Deep process-based learning networks in Biodiversity research

We will implement a multilayer approach to generate process-based species distribution maps accounting for interdependent biological networks. Each layer will be parametrized taking advantage from the integration of biodiversity datasets. Most data in biodiversity are collections of small data. In areas such as species ranges and species interactions, there is a large amount of data, but only a relatively small amount of data for each gene, phenotype, individual or trophic interaction. To customize predictions accounting for interdependent biological levels we will use a formalism considering the heterogeneity at individual level, with its inherent uncertainties, and to couple the individual level together in a hierarchy scaling from genes to phenotypes, populations, communities and species ranges, so that information can be borrowed from other similar levels across the landscape in the absence of empirical estimations. We will implement a multilayer approach using hierarchical Bayesian neural networks[9]. The outputs of the multilayer approach will generate a biodiversity distribution map for many interacting species that can be evaluated against the empirical patterns.

We will contrast two scenarios to explore the best one fitting the empirical patterns. The first scenario will simulate independent levels considering modularity within- and between-layers (i.e., a highly modular pleiotropy matrix determining the genotype-phenotype map and a highly modular within- and between-species interactions with most interactions weak or zero across the landscape.) Such scenario will produce a non- or weakly-interactive species biodiversity map. The second scenario will account for feedbacks among layers. We will explore a range of topologies from bidirectional recurrent neural networks (BRNN) to feedforward neural networks (FNN) and reinforcement learning in unknown and fluctuating environments (RL) [12]. Such scenario will produce an (strongly)-interactive species biodiversity map. We will disturb both scenarios following random and non-random disturbance regimes (i.e., removing specific interactions, abundances and habitats) and will quantify responses to disturbances using a variety of metrics, from biodiversity to functional metrics [10].

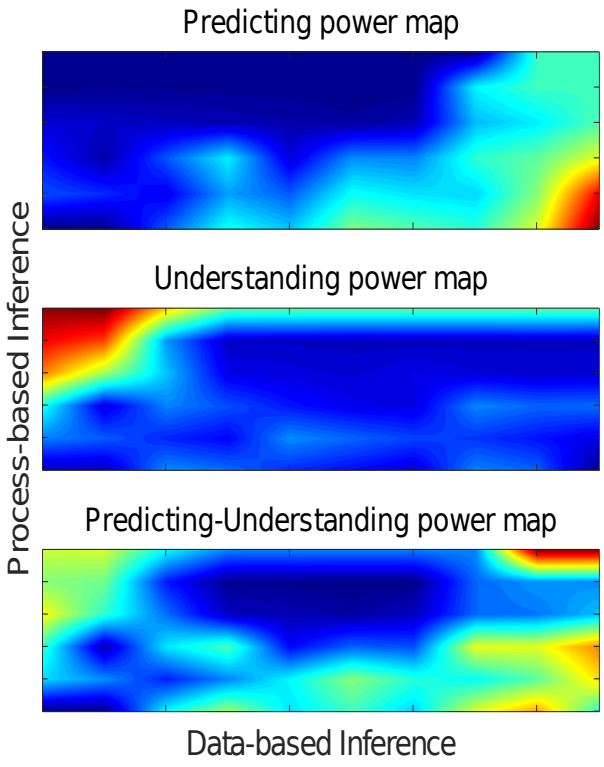

Figure 3: **Prediction and understanding power map**. This figures shows a cartoon of a prediction power map (top), an understanding power map (middle), and a predicting-understanding power map (bottom). x- and y-axis represent data-based inference (i.e., gradient of AI methods from low (left) to high (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 research joining predicting and understanding power of the empirical data might occur.

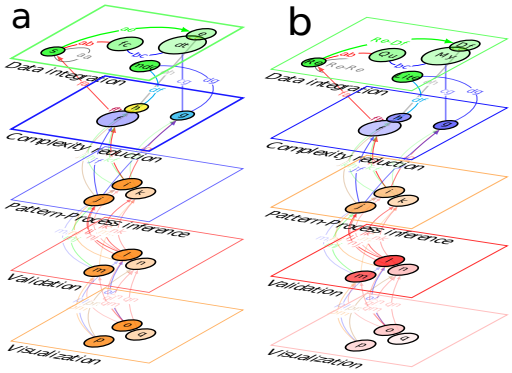

Figure 4: **Figure 4: Prototyping a distributed and open automated research platform: a)** Our initial prototype will contain five layers (this is not an exhaustive number. Some might be merged and others, like reporting generation, can be introduced): Data Integration, Complexity reduction, Pattern-process inference, Validation, and Visualization. Nodes and links represent algorithms and interactions between two algorithms, respectively. The inter-layer interactions will be implemented using the Renku-SDSC platform [8]. The intra-layer interactions will be developed initially in julia language (other languages will come into play during the development of each layer). **b)** A julia-computing-language prototype of an automated research platform. Nodes and links in each layer represent julia packages and interactions between two packages, respectively. The figure shows the julia packages to be used for the Data integration layer containing the packages "Retriever.jl" (**Re**), "Query.jl" (**Qu**), "MySQL.jl" (**My**), "SQlite.jl" (**lite**), and "DataFrames.jl" (**df**). This cartoon representing many intra- and inter-layer connections might be helpful to show the vision of the platform. For example, the path taken to solve a specific intra- or inter-domain (fundamental or applied) question can be quantified by many metrics each producing a distribution of automated solutions across many nodes in a distributed and open network, the Robhoot Open Network (RON). This distribution can be analyzed to quantify properties as robustness, reproducibility and bias of a fundamental or applied solution.

References

[1] Automated statistician. https://www.automaticstatistician.com/index/.

[2] Bigquery. https://cloud.google.com/bigquery/.

[3] Ease.ml. https://github.com/DS3Lab/easeml.

[4] Google ai. https://ai.google/.

[5] Iris ai. https://iris.ai.

[6] Modulus. http://www.modulos.ai/.

[7] Nakamoto terminal. https://www.nterminal.com.

[8] Renku. https://renku.readthedocs.io/en/latest/.

[9] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521:452–459, 2015.

[10] C. J. Melián, B. Matthews, C. S. de Andreazzi, J. P. Rodríguez, L. J. Harmon, and M. A. Fortuna. Deciphering the interdependence between ecological and evolutionary networks. *Trends in Ecology and Evolution*, xx:1–9, 2018.

[11] Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–2024, 2019.

[12] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, jan 2015.

[13] S. Sheehan and Y. S. Song. Deep learning for population genetic inference. *PLoS computational biology*, 12(3):e1004845, mar 2016.