Artificial Intelligence, Machine Learning and Modeling for Understanding the Oceans and Climate Change

Nayat Sánchez-Pi **Luis Martí**
Inria Chile Research Center,
Av. Apoquindo 2827 p.12, Las Condes, Chile.
{nayat.sanchez-pi,luis.marti}@inria.cl

André Abreu
Fondation Tara Océan
8 rue de Prague 75012 Paris
andre@fondationtaraocean.org

Olivier Bernard
Université Côte d'Azur, BIOCORE Team,
INRIA, INRAE, CNRS, Sorbonne Université,
2004 Route des Lucioles,
06902 Sophia-Antipolis, France.
olivier.bernard@inria.fr

Colomban de Vargas
GOSEE CNRS Federation,
Station Biologique de Roscoff.
Place Georges Teissier, 29680 Roscoff, France.
vargas@sb-roscoff.fr

Damien Eveillard
Université de Nantes, CNRS, LS2N
2 rue de la Houssinière F-44322 Nantes, France.
damien.eveillard@univ-nantes.fr

Alejandro Maass
Dept. de Ingeniería Matemática &
Centro de Modelamiento Matemático,
Universidad de Chile - IRL 2807 CNRS
Beauchef 851, Santiago, Chile.
amaass@dim.uchile.cl

Pablo A. Marquet
Dept. de Ecología & C. de Cambio Global,
FCB, Pontificia Univ. Católica de Chile
Bernardo O'Higgins 340, Santiago, Chile.
Inst. de Ecología y Biodiversidad,
The Santa Fe Institute, NM 87131, USA.
Inst. de Sist. Complejos de Valparaíso,
pmarquet@bio.puc.cl

Jacques Sainte-Marie **Julien Salomon**
ANGE Team, Inria Paris
2 Rue Simone Iff, 75012 Paris, France
Sorbonne Université, CNRS,
Lab. Jacques-Louis Lions, 75005 Paris, France
{jacques.sainte-marie,
julien.salomon}@inria.fr

Marc Schoenauer **Michèle Sebag**
TAU Team & LISN, Inria & CNRS, Univ. Paris Saclay
1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
marc.schoenauer@inria.fr, michele.sebag@lri.fr

Abstract

The ongoing transformation of climate and biodiversity will have a drastic impact on almost all forms of life in the ocean with further consequences on food security, ecosystem services in coastal and inland communities. Despite these impacts, scientific data and infrastructures are still lacking to understand and quantify the consequences of these perturbations on the marine ecosystem.

Understanding this phenomenon is not only an urgent but also a scientifically demanding task. Consequently, it is a problem that must be addressed with a scien-

Tackling Climate Change with Machine Learning workshop at NeurIPS 2020.

tific cohort approach, where multi-disciplinary teams collaborate to bring the best of different scientific areas.

In this proposal paper, we describe our newly launched four-years project focused on developing new artificial intelligence, machine learning, and mathematical modeling tools to contribute to the understanding of the structure, functioning, and underlying mechanisms and dynamics of the global ocean symbiome and its relation with climate change. These actions should enable the understanding of our oceans and predict and mitigate the consequences of climate and biodiversity changes.

1 Introduction

Considering the importance and amount of oceans in this speck of dust in the middle of nowhere that we inhabit, we should have called it Planet Ocean. Oceans are not only important because of their volume but are also about the functions and contributions they provide to biodiversity, we included [1]. Oceans play a key role in the biosphere, regulating the carbon cycle; absorbing emitted CO_2 through the biological pump, and a large part of the heat that the remaining CO_2 and other greenhouse gases retained in the atmosphere.

The biological pump is driven by photosynthetic microalgae, herbivores, and decomposing bacteria. Whales also play a prominent role by moving nutrients and providing mixing in the ocean [2–4]. Understanding the drivers of micro and macroorganisms in the ocean is of paramount importance to understand the functioning of ecosystems and the efficiency of the biological pump in sequestering carbon and thus abating climate change.

This situation poses a substantial challenge to humanity as a whole. It is not only an urgent but also a scientifically demanding task. Consequently, it is a problem that must be addressed with a scientific cohort approach, where multi-disciplinary teams must collaborate to bring the best of different scientific areas: state-of-the-art artificial intelligence, machine learning, applied math, modeling, and simulation, and, of course, marine biology and oceanography. They will enable us to understand our oceans and to predict and —hopefully— mitigate the consequences of climate change.

Data is essential in this pursuit. Tara Océans[1] has spearheaded the methodological sampling of the different phenomena that are taking place in our oceans. Despite these efforts, scientific data -even with the import contribution from Tara and infrastructures is not sufficient to adequately understand and quantify the consequence of these perturbations on the marine ecosystem. In particular, critical ecosystems need extensive surveys to characterize the biological acclimation to climate perturbations better.

Consequently, it is necessary to not only gather more data but also to develop and apply state-of-the-art mechanisms capable of turning this data into effective knowledge, policies, and action. This is where artificial intelligence (AI), machine learning (ML), and modeling tools are called for. The application of these methods in the context of ecology and climate change is not new [5]. However, the inherent complexity of this problem poses important challenges to modern computer science and applied mathematics.

2 Context and domain challenges

Despite the growth of research applying AI and ML to problems of societal and global good, there remains the need for a concerted effort to identify how these tools may best be applied to tackle climate change. On the other hand, many computer scientists and practitioners wish to act but are uncertain how. Similarly, many field experts have begun actively seeking input from the AI, ML, and modeling communities. To structure the goals of the project around the following domain challenges:

- **Biodiversity and ecosystem functioning [6]**. Biodiversity supports important functions, such as primary productivity and carbon fixation and sequestration, that are directly or indirectly used and affected by humans.

[1] `https://oceans.taraexpeditions.org/en/`

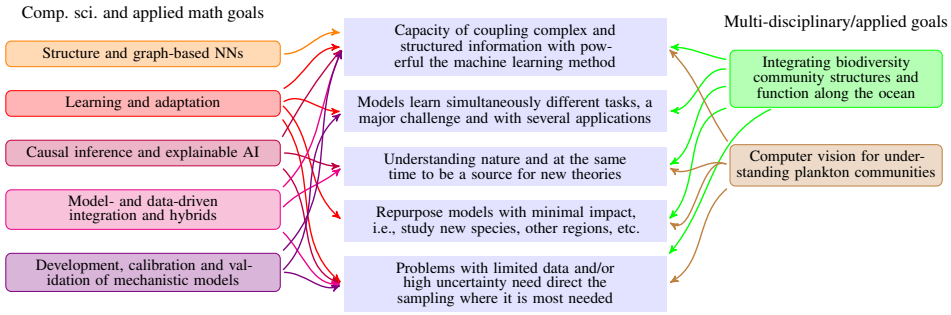

Figure 1: Relation between the AI/ML and modeling approaches with respect to the domain questions.

- **Meta-metabolic modeling**. The objective is to develop a metabolic model including the main microbial oceanic compartments, and couple it with physics. A meta-metabolic model is challenging due to the variety in the pathways and time scales.
- **Phytoplankton biodiversity with regard to temperature [7]**. The main purpose is to create models to properly incorporate plankton complexity into ocean-climate models, assuming the stochastic nature of this system.
- **Data assimilation in biogeochemical models: Predicting the future**. Data assimilation strategies should be developed to calibrate biogeochemical models using the available database. AI tools combined with applied mathematics can allow reaching prediction capability.
- **Computer vision for understanding plankton communities**. Tara Océans has obtained from the samples being extracted as a camera is submerged to grab images of the microscopic organisms found. Some tasks to be address include:
  - *Plankton identification from satellite images*: integrate omics information with high-throughput/high-resolution plankton imaging and environmental data crossed with satellite images.
  - *Connecting images and genomic features*: establish the connection between plankton images and genomic data could state biogeography of the morphological diversity, and identify genes responsible for plankton shapes and morphologies.
  - *Explainable anomaly detection for automatic plankton discovery*: will require extended use of causal inference and image-based explainable AI methods should hint what parts of the observed organism that determining its identification.

3 Addressing the Goals with a Multi-Disciplinary Approach

AI, ML, and modeling tools are key to understanding oceans and climate change. However, their current limitations pose important hurdles in their application. In the case of ML, only recently it has started to be able to handle structured information, like the one required to understand the networks created by interacting populations of different species. Despite the important efforts on data gathering, the current amount of data available conform to a scenario that can be denominated as small data, that heavily contrasts with the data-hungry methods that conform most of the current state of the art in ML. This situation could be overcome either by improving the modeling methods themselves or by taking a stab at developing mechanistic approaches that also seem to be capable of complementing AI and ML in the application domain [8].

The above domain challenges are to be addressed in a multidisciplinary fashion that integrates computer science and applied mathematics. We have identified a group of computer science topics that should be addressed, in particular:

- **Structured and graph-based neural networks [9–12].** The most frequent way to represent biodiversity today is through co-occurrence graphs. These graphs have particular

structures that deserve to be analyzed using the presented techniques and their improvements. A comparison of such graphs is a way to observe the evolution of communities. So having ML methods capable to function on top of this information is essential to understand such dynamics.

- **Learning and adaptation**. This topic comprehends active/few-shot/multi-task learning, transfer learning (TL), and domain adaptation. In problems with limited data and high uncertainty, like the ones to be dealt with here, it is necessary to apply methods that direct the measurements to the areas of the domain where they are most necessary using active learning or Bayesian principles. Here, few-shot learning methods (relying on TL) must take care of producing actionable products with minimal data.
- **Causality [13] and explainable AI [14].** This is a core concern in computer science at the moment. It is also an essential component of the challenge as we intend to use the models created to serve as a means for understanding nature and as sources for new theories.
- **Model-driven and data-driven integration and hybrids.** Biogeophysical models [15] can be very time and CPU time consuming. The idea here is to use deep learning approaches to reproduce the predictions of these resource-demanding models. More precisely, to reduce complex models using deep neural networks. We plan to investigate schemes for decomposing a process model into PDE and statistical components.
- **Development, calibration, and validation of mechanistic models**. The high dimension of the biogeochemical models makes challenging their calibration and validation from a reduced number of measurements.

In addition to the above directions, we plan to deploy an open-access data lake that rely on the M2B3 standard [16] containing or providing transparent access to a diverse set of data sources like Tara Océans data, Copernicus, SeaDataNet, PANGAEA, etc. It will allow to cross-reference and georeference data by providing homogeneous access to all sources and the capacity of merging with other data sources.

We are confident that addressing the domain challenges in the next four years with the focus of researchers of different institutions we will be able to make progress on our understanding of marine biology. We expect to provide actionable decision-making tools that would enable to derive data-informed conclusions and focus resources to deal with the ecological challenges ahead of us.

Acknowledgments and Disclosure of Funding

This work was funded by project CORFO 10CEII-9157 Inria Chile and Inria Challenge project OcéanIA (desc. num 14500). Alejandro Maass is funded by ANID Basal Grant 170001. Pablo Marquet acknowleges funding from ANID Basal Grant 170008.

References

[1] Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stephane Audic, Léo Berline, Jennifer R. Brum, Luis Pedro Coelho, Julio Cesar Ignacio Espinoza, Shruti Malviya, Shinichi Sunagawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Julie Poulain, Sarah Searson, Lars Stemmann, Fabrice Not, Pascal Hingamp, Sabrina Speich, Mick Follows, Lee Karp-Boss, Emmanuel Boss, Hiroyuki Ogata, Stephane Pesant, Jean Weissenbach, Patrick Wincker, Silvia G. Acinas, Peer Bork, Colomban De Vargas, Daniele Iudicone, Matthew B. Sullivan, Jeroen Raes, Eric Karsenti, Chris Bowler, and Gabriel Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470, 2016. ISSN 14764687. doi: 10.1038/nature16942.

[2] J Roman and J J Mccarthy. The Whale Pump: Marine mammals enhance primary productivity in a coastal basin. *PLoS ONE*, 5(10):13255, 2010. doi: 10.1371/journal.pone.0013255.

[3] Jorge León-Munõz, Mauricio A. Urbina, René Garreaud, and José Luis Iriarte. Hydroclimatic conditions trigger record harmful algal bloom in western Patagonia (summer 2016). *Scientific Reports*, 8(1):1–10, 2018. ISSN 20452322. doi: 10.1038/s41598-018-19461-4.

[4] Verena Häussermann, Carolina S. Gutstein, Michael Beddington, David Cassis, Carlos Olavarria, Andrew C. Dale, Ana M. Valenzuela-Toro, Maria Jose Perez-Alvarez, Hector H. Sepúlveda, Kaitlin M. McConnell, Fanny E. Horwitz, and Günter Försterra. Largest baleen whale mass mortality during strong El

Niño event is likely related to harmful toxic algal bloom. *PeerJ*, 2017(5):1–51, 2017. ISSN 21678359. doi: 10.7717/peerj.3123.

[5] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling climate change with machine learning. Technical report, climatechange.ai, 2019. URL http://arxiv.org/abs/1906.05433.

[6] Fons van der Plas. Biodiversity and ecosystem functioning in naturally assembled communities. *Biological Reviews*, 94(4):1220–1245, 2019. doi: 10.1111/brv.12499. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12499.

[7] David Demory, Anne-Claire Baudoux, Adam Monier, Nathalie Simon, Christophe Six, Pei Ge, Fabienne Rigaut-Jalabert, Dominique Marie, Antoine Sciandra, Olivier Bernard, and Sophie Rabouille. Picoeukaryotes of the Micromonas genus: Sentinels of a warming ocean. *The ISME Journal*, 13(1):132–146, jan 2019. ISSN 1751-7362. doi: 10.1038/s41396-018-0248-0. URL https://doi.org/10.1038/s41396-018-0248-0http://www.nature.com/articles/s41396-018-0248-0.

[8] Ruth E. Baker, José-María Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters*, 14(5):20170660, 5 2018. ISSN 1744-9561. doi: 10.1098/rsbl.2017.0660. URL https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0660.

[9] Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. jul 2017. URL http://arxiv.org/abs/1707.05005.

[10] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 12251234, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939753. URL https://doi.org/10.1145/2939672.2939753.

[11] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 855–864, 2016. ISBN 9781450342322. doi: 10.1145/2939672.2939754. URL http://dx.doi.org/10.1145/2939672.2939754.

[12] Balthazar Donon, Benjamin Donnot, Isabelle Guyon, and Antoine Marot. Graph neural solver for power systems. In *IJCNN 2019 - International Joint Conference on Neural Networks*, Budapest, Hungary, July 2019. URL https://hal.archives-ouvertes.fr/hal-02175989.

[13] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82 – 115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012. URL http://www.sciencedirect.com/science/article/pii/S1566253519308103.

[15] Léa Boittin, François Bouchut, Marie-Odile Bristeau, Anne Mangeney, Jacques Sainte Marie, and Fabien Souillé. The Navier-Stokes system with temperature and salinity for free surface flows Part II: Numerical scheme and validation. working paper or preprint, March 2020. URL https://hal.inria.fr/hal-02510722.

[16] Petra ten Hoopen, Stéphane Pesant, Renzo Kottmann, Anna Kopf, Mesude Bicak, Simon Claus, Klaas Deneudt, Catherine Borremans, Peter Thijsse, Stefanie Dekeyzer, Dick MA Schaap, Chris Bowler, Frank Oliver Glöckner, and Guy Cochrane. Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. *Standards in Genomic Sciences*, 10(MAY2015), 2015. ISSN 19443277. doi: 10.1186/s40793-015-0001-5.

[17] Stephane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Sabrina Speich, Romain Trouble, Celine Dimier, and Sarah Searson. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2(Lmd):1–16, 2015. ISSN 20524463. doi: 10.1038/sdata.2015.23.

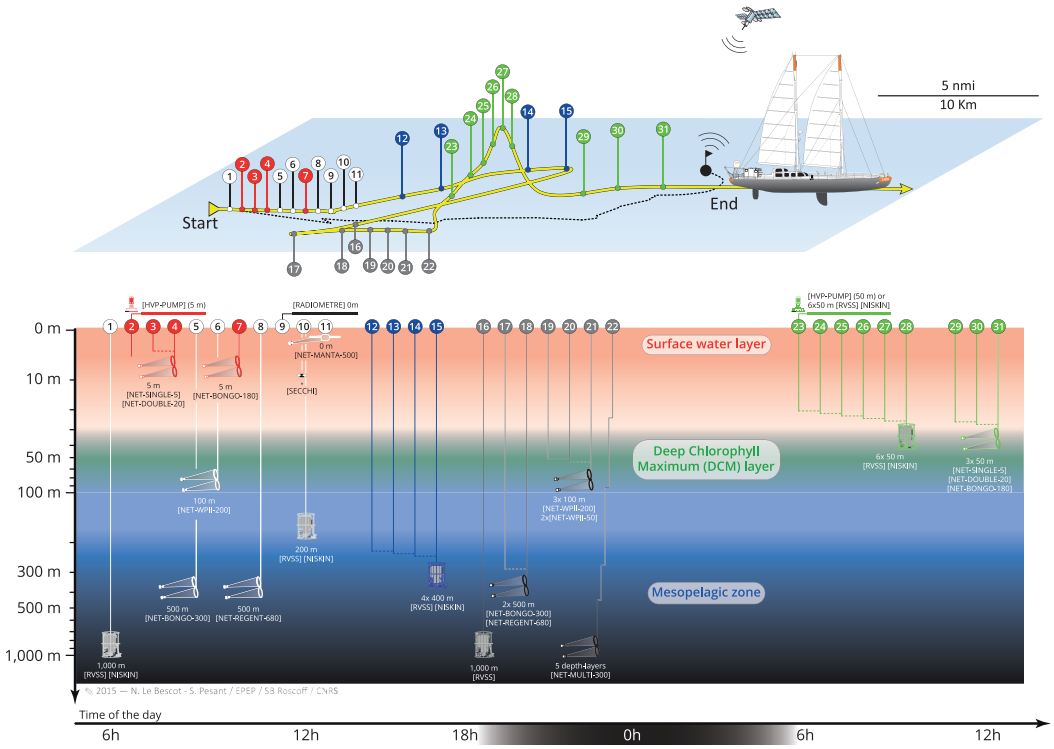

Figure 2: Spatial representation and chronology of Tara sampling methodology events during a 24–48 h station. Colored markers along the route of SV Tara (yellow surface track) correspond to sampling events targeting the surface water layer (in red), deep chlorophyll maximum layer (green, here at 50 m), and the mesopelagic zone (blue, here at 400 m). At some stations, an Argo drifter (10 m floating anchor and satellite positioning) was used to follow the water mass during sampling (black surface track). Taken from [17], shared under a Creative Commons Attribution 4.0 International License.

A Context on Tara Expeditions Data

There is a clear scientific consensus about the effects of climate change on the global ocean: among others a shift of temperatures, an increase of acidification, deoxygenation of water masses, and perturbations in nutrient availability and biomass productivity. Altogether, these abiotic changes will have a drastic impact on almost all forms of life in the ocean with further consequences on food security, ecosystem services and the well-being of coastal communities. In this regard, Tara Océans has spearheaded the actions directed towards sampling and understanding the different phenomena that are taking place. Figure 2 illustrate the complex sampling methodology applied. Despite these numerous impacts, scientific data -even with the import contribution from Tara Océans and infrastructures are not sufficient to adequately understand and quantify the consequence of these perturbations on the marine ecosystem. In particular, critical ecosystems need extensive surveys to characterize the biological acclimation to climate perturbations better. Consequently, it is necessary to not only gather more data but also to develop and apply state-of-the-art mechanisms capable of turning this data into effective knowledge, policies and action. This is where artificial intelligence, machine learning and modeling tools are called for.

The upcomming Tara expetion with cover the Patagonian region. This is a unique ecosystem that represents an open sky laboratory for ecological studies. This pristine region is indeed changing more rapidly under the effects of climate change and describes an oracle of changes to come in the next decades for other parts of the ocean. Patagonia is fundamental to understand the responses of the microbial marine life at the interface between antarctic waters, the coastal ecosystems, and the melting glaciers. This region is also one of the most productive regions in the ocean, accounting for more than 30% of sardines stocks, among other species and one of the most important region in

sequestering carbon. Patagonia is also a hot spot of aquaculture, with an intensive salmon production, an ecosystem that is both impacting, and being impacted by, climate changes. In order to understand the functioning of this large scale ecosystem, the Tara Océans initiative has decided to carry out and intense sampling campaign.

The consortium will build a modeling framework dedicated to ocean modeling, contributing to learn causal and explanatory models; fair data models; and robust models. This project is an opportunity to contribute key scientific knowledge on a global pressing problem as climate change is, capitalizing on the experience and articulation of the teams involved and the availability of data on a key area, as is the Patagonia, that can provide answers that can be transferred to others parts of the oceans.

The motivation of this interdisciplinary project is to develop new AI and mathematical modeling tools to contribute to the understanding of the structure, functioning, and underlying eco-evolutionary mechanisms and dynamics of plankton in the global ocean. Methods like deep learning, causal and inference learning, sequential decision making, transfer learning, multi-criteria optimization are just a few that can be applied to these kinds of complex problems, allowing us to get reliable knowledge from the ocean and its interactions. To do this, we will use the corpus of Tara Océans Expeditions datasets, which is, as far as we know, the most comprehensive case study to develop AI and mathematical modeling methods for studying global ecology along with other related datasets. This fundamental baseline currently makes marine plankton the best-described planetary ecosystem in terms of taxonomic composition, abundance, and genetic diversity, making this project realistic.