# QIEP-B: Quantifying the Impact of Environmental Parameters on Biodiversity

**Projet porté par Clovis Galiez (LJK) et Olivier François (TIMC) en partenariat avec Wilfried Thuiller (LECA) et Samuel Chaffron (LS2N).**

## Context and objectives

Climate change, intensive land use, pollution and biological invasions represent a substantial threat to the diversity of living organisms on Earth. Monitoring and predicting the responses of these organisms to these global changes is thus essential to anticipate and protect biodiversity and the associated ecosystem services. More specifically, we know that each species occupies a particular ecological niche that defines the multi-dimensional space of resources (e.g., light, temperature, nutrients, structure, etc.) available to (and specifically used by) organisms, where the population can persist. Under environmental change, we want to understand how species will respond to given their ecological niche.

Environmental changes could lead to range expansion of some species, range extirpation of others, local extinction, with important consequences on the composition of ecological communities. Interactions among organisms also come into play, for example due to symbiotic or trophic relationships, complexifying the response of communities to environmental changes. Moreover, the relationship between an environmental variable and the shift in community composition is likely to be highly non-linear. This means that, for some range of values, the composition of an ecosystem may be unaffected. For some other range of values, a small shift may lead to tremendous changes in the composition of a community, and therefore impacts the potential resilience and stability of the ecosystem. The critical values of environmental parameters leading to dramatic changes in ecosystem composition are coined the "tipping points" of an ecosystem (Lenton et al. 2008). The study of tipping points is of hot topic in the context of climate change (Lenton et al. 2019, Dakos et al. 2019).

For decades, theoretical ecologists have investigated how environmental fluctuations affect species composition of ecosystems, modelling the response of each organism to environmental parameter changes, possibly considering or inferring interactions between species (or higher taxa). Most of the previous works assume an implicit regularity, such as the underlying Gaussian distribution of species abundances (Faisal et al. 2010), and fail to handle the actual variability of communities at fixed environmental conditions.

In this project, we want to develop *data-driven* methods to simulate ecological communities under observed environmental conditions (WP1) and quantify the shift in composition of ecosystems caused by a change of an environmental variable such as temperature (WP2). To carry out this work, members of LJK, TIMC, LECA and LS2N will collaborate to leverage high-throughput DNA sequencing data. Our project will use environmental DNA and metagenomic information, allowing the observation of relative abundances of a large number of species without heavy wet lab protocols. This project will develop new computational methods for analyzing global biodiversity and simulating its response to environmental changes by combining machine learning and bioinformatic techniques with genomic data.

We propose to study three case studies in which data are readily available: alpine ecosystems (WP3.T1), ocean microbiome (WP3.T2), and human gut microbiome (WP3.T3). We therefore think that our project contributes to the Data Science strategic axis of the LabEx, opening up a new field of application for Persyval-Lab toward modern quantitative ecology with concrete societal and clinical impacts.

# Proposed collaboration

The consortium is highly complementary. It involves specialists in machine learning and computational biology (Clovis Galiez, LJK, Olivier François, TIMC), in statistics (Olivier François), in ecology and modelling (Wilfried Thuiller, LECA) and system ecology (Samuel Chaffron). The collaboration between partners in LJK (team SVH) and TIMC (team BCM) is particularly suited since the present work relies on computational methods for DNA data. Clovis Galiez has experience into applying machine learning techniques for metagenomic data classification, and will benefit from interactions with LJK members for statistical learning and optimization. The BCM team in TIMC has a long experience of genomic data analyses. The project on metagenomics of the human gut is specifically relevant to clinical applications given the profound implications of the microbiome in health and diseases (Durack & Lynch 2019). The PhD student hired for this project will be co-supervised by Clovis Galiez and Olivier François (HDR). Moreover, the work developed in WP1 will benefit from the experience gained in methods developed by Clément Gain during his thesis work (Dir: Olivier François, TIMC).

Wilfried Thuiller (LECA, Grenoble) and Samuel Chaffron (LS2N, Nantes) are experts in the Orchamp (WP3.T1) and Tara (WP3.T2) data respectively, a great asset for the project to ensure consistency of data processing and to avoid biases in its treatment. They will contribute to draw the ecological conclusions provided by the methods developed in this project. They will also actively participate in the PhD supervision and be part of the steering committee of the thesis.

# Consortium members

**Clovis Galiez** is an assistant professor in computer science at Ensimag recruited in September 2018. His previous works mainly dealt with machine learning for classification problems in bioinformatics and metagenomics. He recently joined the LJK where his current research focuses on leveraging machine learning techniques for modeling microbial populations through metagenomics data. He regularly participates in workshop sessions in a project with Olivier François and Wilfried Thuiller about ecological network inference for global ecosystem monitoring. CG worked in H2020 European Project involving biologists, physicists and computer scientists as well as industrial partners, and he has experience with working in multidisciplinary and interdisciplinary partnerships.

**Olivier François** is a professor in applied mathematics at Ensimag, Univ. Grenoble-Alpes. His research is to develop statistical methods for ecological and evolutionary genetics. He studies evolutionary questions including the demographic history of model species and their adaptation to the environment. OF has created a research group in mathematical and computational biology at TIMC, and had been the leader of the group for thirteen years. He has authored 110 peer-reviewed publications and supervised sixteen PhD theses.

**Wilfried Thuiller** is a senior research director at CNRS. He is specialized in biodiversity modelling using both mechanistic and statistical models, and he is leading the Orchamp monitoring program. Most of his current work is focused on analysing interaction networks through space and time, and modelling them with the development of new statistical models. He is one of the 8 Highly Cited Scientist at Univ. Grenoble Alpes, and the most cited scientist in the world over the last ten years in Environmental Science. His work has been heavily used on the IPCC and the IPBES assessments in which he has been coordinating leading author. He is the co-PI of the chair MIAI@Univ.Grenoble.Alpes on artificial intelligence and remote sensing.

**Samuel Chaffron** is a computational biologist and microbiologist holding a Ph.D. from the University of Zurich. Since 2017, he is a CNRS researcher with well-established international expertise in the understanding of the structure, function, and diversity of natural microbial communities. By integrating meta-omics information through systems biology approaches in various habitats, including the world's oceans and the human intestinal tract, his research has helped to understand ecological systems at the molecular scale and revealed the diversity and complexity of microbial ecosystems. His research lies in modeling microbial communities and their structures at different levels of organization (from genes to communities and ecosystems). He develops systems ecology and functional (meta-)genomics approaches to reveal universal patterns shaping natural microbial communities. He also builds computational models to gain a predictive understanding of community function and dynamics through metabolic modeling, and acquire a mechanistic understanding of species interactions (e.g., marine symbioses, microbiome-gut-brain axis) and ecosystem functioning.

| Participant | Lab | Employer | Role | Participation |
|---|---|---|---|---|
| Clovis Galiez | LJK* | G-INP | PI, co-PhD advisor | 15 p.month |
| Olivier François | TIMC* | G-INP | co-PhD advisor | 12 p.month |
| PhD Student | LJK | UGA | all WPs | 36 p.month |
| Wilfried Thuiller | LECA | CNRS | all WPs | 6 p.month |
| Samuel Chaffron | LS2N | CNRS | all WPs | 3 p.month |

***: In LabEx Persyval**

## Methodological details and work breakdown organization

The ultimate goal of the project is to quantify ecosystem shifts between environmental conditions. As the number of species (taxa) is higher than the number of samples, and given that species interactions cannot be ignored, we will adopt a dimension reduction approach (WP1). We will make use of *variational auto-encoders* (Doersch 2016, Kingma 2014) tha will allow us not only to model a complex system through a non-linear embedding in a latent space of much smaller dimension, but also to generate simulated ecological communities corresponding to modified environmental conditions and to quantify the ecological offset resulting from the simulation.

WP2 aims at quantifying the ecological cost of environmental changes induced on species communities. A change will be modelled by a shift of some parameter, for instance a temperature variation from $T = T_0$ to $T = T_1$. The quantification of the induced ecological cost will be implemented using the optimal transport theory (Peyré & Cuturi 2018), which computes a minimal cost for the transportation of the distribution of samples at initial condition $S|T = T_0$, to the distribution after the environmental change $S|T = T_1$. The transport cost is defined with respect to a *ground metric* between samples $S$. More precisely, if samples represented in the distribution $S|T = T_0$ are different but biologically similar (as measured by the ground metric) to those in $S|T = T_1$, then the transport cost will be low, and we can assume that ecosystems will adapt easily to such a change and will keep providing the same ecosystem functions. However, if it turns out that the samples in the two distributions are biologically very different, the cost will be high and the shift may correspond to a tipping point of the ecosystems.

WP3 will consider applications of VAE and optimal transport theory to available data: eDNA of alpine ecosystems from the Orchamp monitoring project (WP3.T1), marine microbiome shotgun sequencing data from the TaraOcean expedition (WP3.T2) and finally human gut microbiome using the 16S amplicon data from the American Microbiome Project (WP3.T3).

### WP 1 - Develop a method to simulate the distribution of taxa conditionally on environmental variables

**Participants**: *PhD student co-supervised by Olivier François (TIMC) and Clovis Galiez (LJK). In collaboration with Wilfried Thuiller (LECA) in particular for the interpretation of the ecological assessment.*

In WP1, we will simulate ecological communities conditionally on environment variables by training a conditional variational autoencoder model (cVAE). Variational auto-encoders (VAE) are unsupervised regression models that can be used in a generative way. VAEs learn an embedding of a vector representing the ecological community in a latent space. Each component of the input vector represents the abundance of a species (taxon), and the latent variables are optimized so that a probabilistic generative model reproduces the input distribution.

Recent advances (cVAE) allow us to condition the latent variables on the environmental variables associated to the training samples (Sohn et al. 2015). Conditionally on environmental variables, cVAEs can model the joint distribution of taxa in samples. Since a generative model is defined by the VAE, the algorithm can simulate new samples while taking into account abiotic (i.e. environmental factors) and biotic (i.e. interspecies) interactions.

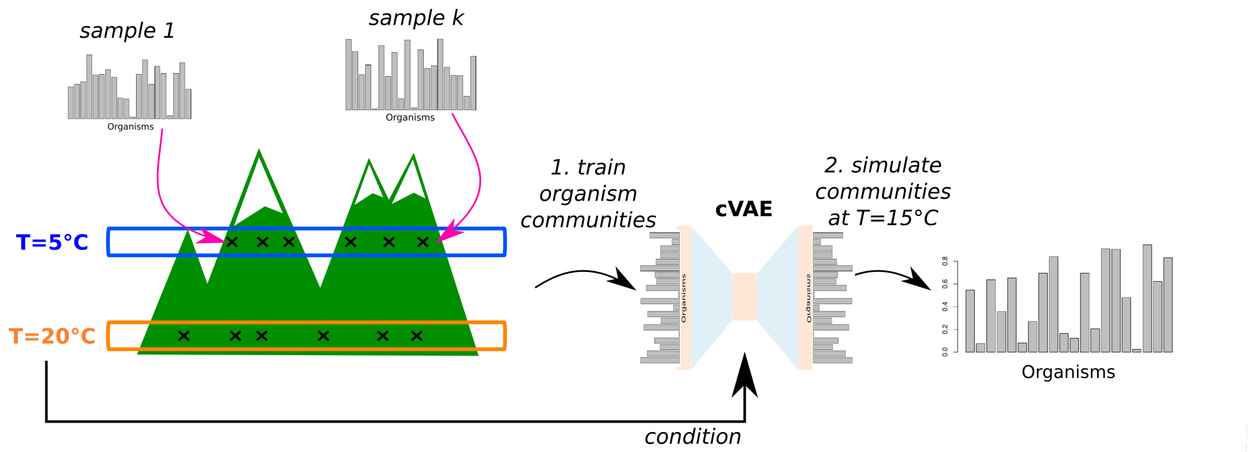

Figure 1: WP1 principle: train a cVAE on organisms communities conditionnally on an environmental variable. Use the model to simulate communities at different temperatures.

For the majority of ecosystems, cVAEs will interpolate between data points from lower latitude and data from same latitude, and we are aware of uncertainty in prediction for taxa at the lower edge of the range (ie, equator). We nevertheless expect that even at the edge of the observed range of parameters, for very small shifts, the first order approximation - called tendency of ecosystem compositional shift - will be correctly captured by the model.

### WP 2 - Develop a method to quantify the shift of biodiversity due to a shift of an environmental variable

**Participants**: *PhD student co-supervised by Olivier François (TIMC) and Clovis Galiez (LJK). In collaboration with Wilfried Thuiller (LECA) in particular for the interpretation of the ecological assessment.*

The goal of WP2 is to quantify the "evolutionary effort" that would be needed by an ecosystem to adapt to a shift of environmental variables. It is theoretically similar with the challenge of identifying tipping points of ecosystems to provide a safe management of land use and to regulate critical anthropogenic activities.

Common approaches for quantifying the "evolutionary effort" use latent structures modeling, such as partial network inference, to be able to measure the impact of an environmental factor on the ecosystem. Although these approaches provide novel knowledge on ecosystemic mechanisms, they also carry many hypotheses on the biotic and abiotic interactions - such as sparsity - which are hard to validate. Moreover, every tractable model supposes that the environment deterministically biases the ecosystem toward a given ecosystem composition, and that the observed realization is a variation around this optimum. Here we take a data-driven approach, which does not make any hypothesis on the biotic and abiotic interactions in ecosystems. We only hypothesize that ecosystems with similar environmental parameters are mirroring all the possible species compositions.

**A simple example.** Let us consider alpine ecosystems at two different temperatures 5°C to 7°C, with 2 samples for each condition:

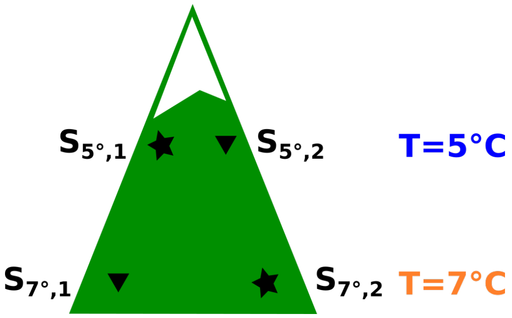

Figure 2: Let us suppose that these samples are representative of the possible species composition found at these temperatures.

If $S_{5^\circ C,1}$ is very similar to $S_{7^\circ C,2}$, and if $S_{5^\circ C,2}$ is very similar to $S_{7^\circ C,1}$, then it is likely that $S_{5^\circ C,1}$ will adapt to $S_{7^\circ C,2}$ and $S_{5^\circ C,2}$ to $S_{7^\circ C,1}$, and that the overall change will be neglectable.

However, if still $S_{5^\circ C,1} \approx S_{7^\circ C,2}$ but $S_{5^\circ C,2}$ is very different from both $S_{7^\circ C,1}$ and $S_{7^\circ C,2}$, then it is likely that $S_{5^\circ C,1}$ will still adapt to $S_{7^\circ C,2}$ at low "evolutionary" or "ecosystemic function" cost, but that $S_{5^\circ C,2}$ will have to adapt to $S_{7^\circ C,1}$ or $S_{7^\circ C,2}$ with much more effort (i.e. at a much higher cost).

Given a measure of dissimilarity between samples (called the *ground metric*), the minimal cost that Nature has to "pay" to change from $T = 5°C$ to $T = 7°C$ can be formally computed using *optimal transport* (OT) theory.

The ground metric will be defined by WP2.T1, and the optimal transport quantification by WP2.T2.

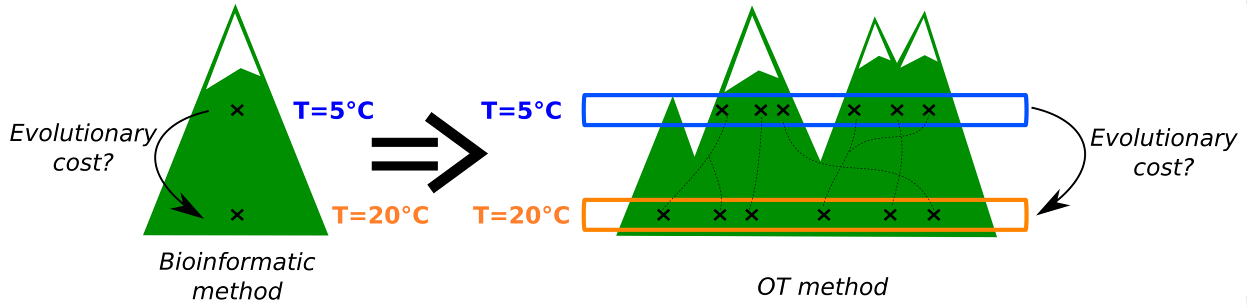

Figure 3: WP2 principle: quantify the evolutionary cost between two samples (WP2.T1), and consider a pool of samples as a representation of the possible dynamic equilibria and quantify (WP2.T2) - using optimal transport (OT) theory - the cost of shift of biodiversity due to a shift of an environmental variable. The dotted lines are an idealized representation of a transport plan of OT.

### T1 - Define a "biological" distance between samples

First, we will investigate various possibilities for measuring dissimilarity between species distributions (i.e. samples) identified from genetic information using metagenomics of environmental DNA tools.

The choice of the dissimilarity measure $D_{i,j}$ between samples $i$ and $j$ will depend on the downstream ecological problem we want to look at.

Using the Unifrac distance (Lozupone & Knight 2005), $D_{i,j}$ will measure the phylogenetic turnover between samples. This can be interpreted from an evolutionary perspective: the more dissimilar are $i$ and $j$, the more time it would take for taxa to evolve from $i$ to $j$. Variants of this measure will be considered by transforming the phylogenetic tree with the delta transformation (Pagel, 1997) to emphasize either on recent or ancient splits.

When considering faster environmental changes, species do not have time to evolve. In that case, one can use a $D_{i,j}$ accounting for taxa tunover between samples $i$ and $j$ such as Bray-Curtis dissimilarity (Bray & Curtis 1957). $D_{i,j}$ is thus interpreted as the quantity of species to be carried by opportunistic migration. We will evaluate possible variants of this measure to emphasize importance of abundant or rare taxa, in the same spirit as Hill numbers for $\beta$-diversity measurements (Hill, 1973).

To measure the shift in terms of ecosystemic functions or services, $D_{i,j}$ based on taxa or phylogeny are not suited since similar ecosystem services can be carried by different species. In that case, a $D_{i,j}$ will be chosen to represent for instance the shared metabolic pathways between the two samples.

Finally, other dissimilarity measures will be implemented and compared, such as the *Wasserstein distance* computed from species embeddings that has been already explored in a preliminary study supervised by Clovis Galiez through a master project (B. Fayolle).

**T2 - Evaluate the impact of environmental variable shifts**

WP1 provide a way of simulating samples for given environmental conditions. By identifying similar samples together according to the ground metric (clustering), for a given temperature $T$, we will get a probability distribution $S|T$ over samples.

The global impact on samples of a shift from a temperature $T_0$ to $T_1$ is computed as the minimal cost in terms of optimal transport (Peyré & Cuturi 2018) of the distribution $p(S|T = T_0)$ of samples at temperature $T = T_0$ to the distribution of samples $p(S|T = T_1)$ at temperature $T = T_1$. According to optimal transport theory, the optimal cost is described by the *Wasserstein distance* $\mathcal{W}(S|T_0; S|T_1)$.

In particular, when samples at $T_0$ are similar to samples at a temperature $T_1$, then the *Wasserstein distance* will be close to zero, meaning that little cost will occur. On the contrary, if the samples present in the distribution $p(S|T = T_0)$ are very dissimilar to those in the distribution $p(S|T = T_1)$, then the *Wasserstein distance* will be large, indicating a stronger cost on ecosystems. We therefore consider the cost of a temperature shift on samples as being defined by the *Wasserstein distance* between the conditional distributions

$$\iota(T_0, T_1) = \mathcal{W}(S|T_0; S|T_1)$$

Several interesting quantifications follow from this definition. For instance we can define the sensitivity of an ecosystem to a change of temperature as

$$s(T) = \lim_{\delta T \to 0} \tfrac{\iota(T, T+\delta T)}{\delta T}$$

At a given temperature $T = T_0$, if a small change of $T$ has no impact on ecosystems, then $\iota(T, T + \delta T) \approx 0$, and the sensitivity $S(T_0)$ will be also small. Conversely, high values of $s(T)$ correspond to temperatures with high impact on ecosystems.

Ecosystem tipping points will be defined as the environmental parameter values having the highest sensitivity. We will compare to known reported critical values for different ecosystems (in particular marine and alpine).

To limit the risks of the project, in case the simulations of WP-1 turn out to be unreliable, we will associate the modeled abiotic niche (i.e. a distribution $T|S$) to every sample, and compute the distributions $S|T$ using simple Bayes' rule:

$$p(S|T) = \tfrac{p(T|S)p(S)}{\sum\limits_i p(T|S_i)p(S_i)}$$

Then, we can proceed to compare the distribution $S|T_0$ and $S|T_1$ using optimal transport theory as previously described.

### WP 3 - Apply the methods on real data

We will apply the methods developed in WP1 & WP2 to different types of ecosystems with slightly different but convergent scientific goals.

Mainly three main sources of environmental genetic sequencing data are currently available for our approach: environmental DNA (eDNA), metagenomic shotgun, and amplicon data. Environmental DNA focuses on extra-cellular DNA resulting from the degradation of the cells of organisms, giving a snapshot of the full biodiversity composition, from micro-organisms to large mammals. The other techniques mainly focus only on micro-organisms, that are acknowledged as major players, proxies and drivers for higher level biodiversity. Shotgun metagenomics gives access to the full genomic information, resulting in very high amount of data, providing access to gene content, metabolic pathways, but also taxonomic identification. Amplicon DNA focus on the identification of particular taxa in samples, resulting in much lighter data although not providing functional interpretations.

#### T1 - Alpine ecosystems: Orchamp

**Participants**: *PhD student, Wilfried Thuiller (LECA).*

We will consider extra-cellular DNA from soil samples that give a snapshot of the whole biodiversity present in an ecosystem from micro-organisms to mammals. The Orchamp data is made of 25 elevation gradients (e.g. temperature) across the entire French Alps (UGA IRS Montane and CDP Trajectories). Each gradient consists of 5 to 8 permanent plots spaced every 200m of altitude, from down the valley to the top of mountain, with three subplots each. For each subplot, soil samples were collected and the DNA extracted, amplified and sequenced. The data has been processed, curated and cleaned already by the LECA team. Applying our approach on these data will allow us to estimate the impact of global warming on the whole biodiversity of alpine ecosystems.

Moreover, on one of the 25 gradients (close to the Jardin du Lautaret), 40 meters (40 times 1x1m) land plots have been swapped between two different altitudes (2450m and 1950m) in 2016, and yearly eDNA sampling has been carried out to monitor biodiversity evolution across time. We will use this experiment to quantify how fast alpine ecosystems are stabilizing after a drastic change of temperature and measure the recovery status with respect to their the original ecosystem composition.

The eDNA data comes with a cost of having more noise in the data and less precise species identification than more traditional sequencing. This noise should be taken into account when measuring distances between samples. A potential way to correct the noise is to use a particular form of denoising autoencoders, in phase with the developments of WP1.

#### T2 - Ocean microbiome: Tara Oceans

**Participants**: *PhD student, Samuel Chaffron (LS2N).*

Sunagawa et al. (2015) showed that temperature is the main abiotic factor shaping microbial community composition in the euphotic zone of oceans. Moreover, climate warming is expected to worsen in the next decades, and it is therefore of a major challenge to evaluate to what extent a rise of temperature will impact ocean ecosystems (Baltar et al., 2019). Here, we propose to use our method for a data-driven estimation of this impact. The data consist of $\sim$ 200 shotgun metagenomic samples from the global ocean, for which the major part of the biodiversity is already known (Sunagawa et al. 2015). Thanks to the full sequencing of the DNA information, by mapping DNA information on reference databases, we can detect known metabolic function encoded in organisms' genomes. This enables us to consider distances between samples in terms of metabolic functions, thereby measuring the shifts in terms of impact on ocean ecosystem functions and services.

### T3 - Gut amplicon metagenomics: The American Gut Project

**Participants**: *PhD student, Clovis Galiez (LJK), Olivier François (TIMC).*

The American Gut Project (AGP, McDonald et al. 2018) is a collaborative project which currently consists of almost 20,000 samples of American gut amplicons, together with individual metadata such as diet, drug intake, etc. (Tataru et al. 2019). This amount of data is still expected to increase during the QIEP-B project and it has been under-exploited. The meta data of WGP is mostly discrete (e.g. low, medium, high intake of sugar) and we will use WP-1 to simulate gut communities on a continuous scale.

Applying the methods of WP2, we will be able to measure if changes of gut conditions (in sugar intake for instance) have a lot of impact in terms of community and functional turnorver.

Moreover, some samples are labeled with disease informations (such as inflammatory Bowel disease) and we will quantify if samples differ significantly between healthy and diseased individuals in terms of species composition and metabolic function. To assess the significance, we will compare to a null hypothesis by computing optimal transport distances between random groups of individuals.

Samuel Chaffron has a strong experience working on gut microbiome communities, and Clovis Galiez already supervised a master project (student: B. Fayolle) dealing species embeddings on AGP data.

## Planning

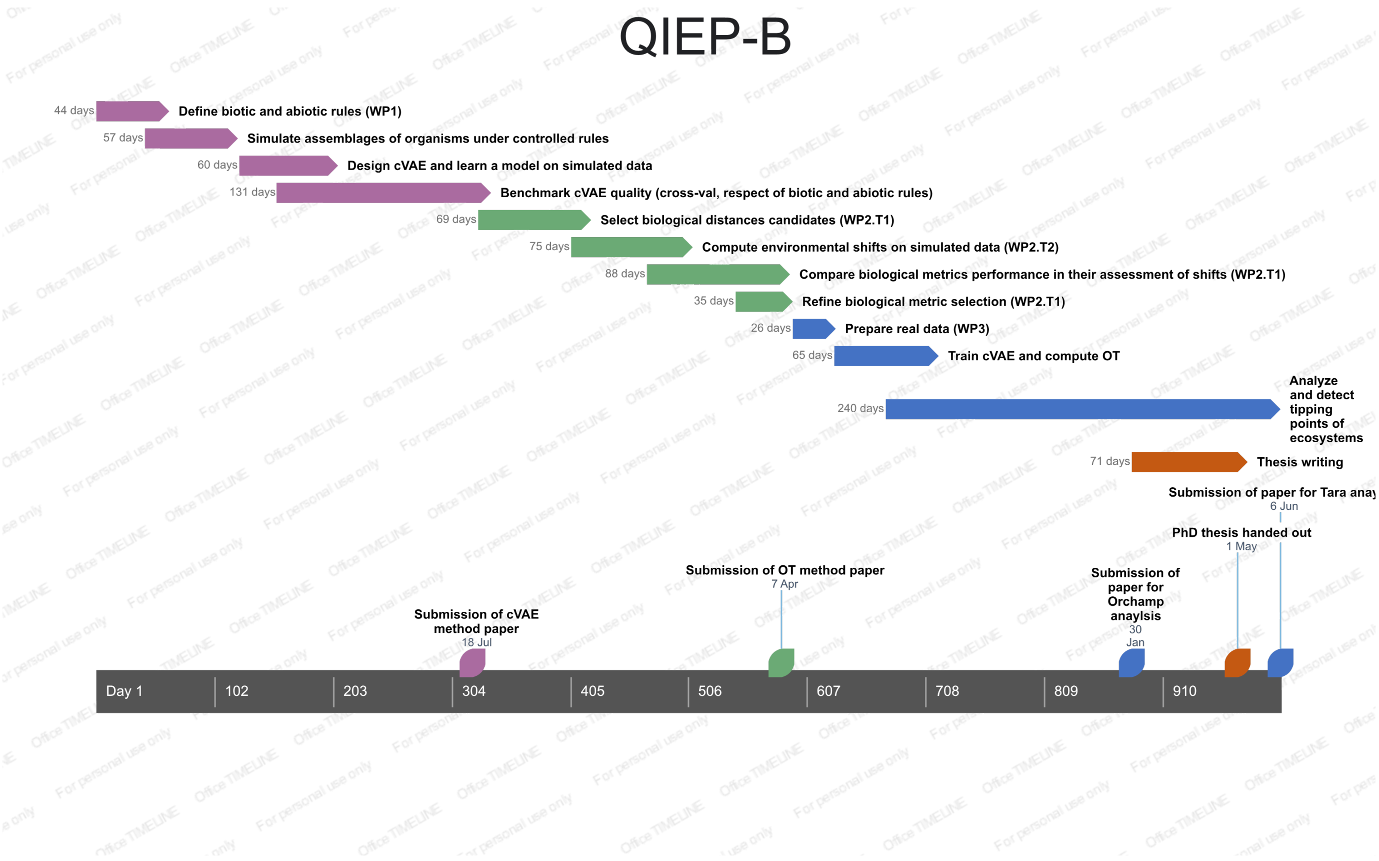

Figure 4: Overview of the QIEP-B project timeline. Duration are in working days.

# Expected outcomes

## Scientific publications

We expect the methods developed in WP1 and WP2 to be published in applied computer science journals (*Methods in Ecology and Evolution, PLoS Computational Biology, Oxford Bioinformatics*, etc.). The usage of the methods on real data and the analysis of results are expected to be published in specialized high impact journals (*Nature Communications, Cell Systems, Microbiome, PLoS Computational Biology*). All publications will be open access.

## Scientific animation

Among participants in the consortium, we plan to meet regularly to coordinate the project. Moreover, we plan to organize once a year a workshop including collaborators and experts for specific scientific questions that may arise during the project.

Some members of the consortium (Clovis Galiez, Olivier François, Wilfried Thuiller) are already interacting through a regular journal club, in which we will discuss the bibliography of the present project.

## Software

The methods will be implemented as R packages or Python libraries allowing users to retrain the models with their own data. Moreover, the models trained on the datasets of WP3 will be available to the community, allowing to simulate organisms communities under environmental parameters. Every software and model will be released under GPL licence.

## Outreach

Our contribution to the tipping point analysis of Earth ecosystems is of interest for the scientific and general public audience. We plan to communicate our methods and results in scientific conferences, but we are also convinced that it is our role to diffuse the scientific investigations and conclusions to the general audience as well. Therefore, we will strive to make our results available in local journals (specifically concerning the Orchamp data) or in national general audience journals (e.g. The conversation).

# Requested resources

This project is planned to be a collaborative work around a PhD thesis (36 p.months). It also require additional administrative and travel costs, and some publication costs.

| Expenses | Amounts |
|---|---|
| Salary (PhD) | 103k |
| Travel | 3k |
| Publication fees | 4k |
| Scientific animation | 7k |
| Total | 117k |