Interaction Graphs of Phytoplankton Species Interactions using Logical Learning

Madeleine Eyraud[1], Maxime Folschette[1], Katsumi Inoue[2], Sébastien Lefebvre[3], and Cédric Lhoussaine[1]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
maxime.folschette@centralelille.fr
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[3] Univ. Lille, CNRS, Univ. Littoral Côte d'Opale, UMR 8187 - LOG - Laboratoire d'Océanologie et de Géosciences, Station marine de Wimereux, F-59000 Lille, France

Abstract. The functioning of marine ecosystems depends on key processes such as climate regulation and water quality. Phytoplankton, unicellular microalgae at the base of marine food webs, play a central role in these dynamics that are threatened by global change. Although abiotic factors (e.g., temperature) are well studied, biotic interactions (i.e., between species) are poorly understood. Classical machine learning models, while effective at prediction, often operate as black boxes and provide limited biological interpretability. Meanwhile, symbolic and qualitative modeling approaches, which could offer greater explanatory power, remain largely underused in marine ecology. To address this gap, we investigate phytoplankton interactions using an explainable machine learning method, LFIT (Learning From Interpretation Transitions), which infers logical rules from observational time series data. Adapting this framework required methodological contributions, notably a species-specific discretization strategy informed by ecological theory. The extracted rules enable the construction of an interaction graph, where edges represent probable interspecies interactions. This graph offers an interpretable representation of the dynamics of the community and helps identify key drivers of the development of phytoplankton.

Keywords: Marine Ecology · Explainable Machine Learning · Symbolic Learning · interaction Graphs · Network Inference · Data Discretization.

1 Introduction

Marine ecosystems are complex adaptive systems shaped by physical, chemical, and biological interactions. At the base of these ecosystems lie phytoplankton communities, whose dynamics drive primary production, biogeochemical cycling, and trophic networks [7,33]. Understanding and predicting phytoplankton dynamics is therefore critical for both ecological science and practical challenges such as monitoring harmful algal blooms and coastal water quality, anticipating climate change and tracking climate-driven biodiversity shifts [24,22]. While

the role of *abiotic* factors (such as temperature, light, and nutrients) is well studied [8,6], recent studies suggest that *biotic* interactions (i.e., interactions between species) also play a major role in structuring phytoplankton communities [21,26,16]. These interactions, however, remain poorly characterized, in part due to the difficulty of disentangling them from environmental effects in observational data [26].

Studying phytoplankton communities has been already performed using ODE systems, but they usually require the identification of a lot of parameters and fine-tuning. Statistical machine learning methods, often described as "black boxes", offer predictive models but at the cost of explainability. When understanding the drivers of interspecies interactions is crucial for interpretation and management, this lack of transparency is a significant limitation. To address it, explainable artificial intelligence (XAI) has emerged as a bridge between interpretability and predictive modeling. Here, we propose a symbolic and explainable modeling approach based on *Learning From Interpretation Transitions (LFIT)* [18], which infers human-readable logical rules from time-series data.

As phytoplankton community structure (abundance per species) indicates rapid change in the functioning of ecosystems, long-term monitoring networks (>30 years) were implemented in some coastal environments at bi-weekly frequencies in the framework of European marine water directives and French monitoring networks [32,24]. We apply the LFIT framework to this multi-decadal, species-level records across several coastal locations in the north of France. This dataset offer a realistic and ecologically meaningful testbed for interaction modeling. Our goal is to extract logical rules that describe interspecies interactions (while accounting for abiotic effects) and use them to construct a directed, weighted interaction graph. Such a graph provides a compact and interpretable summary of the full logic program, which may be long and difficult for domain experts to interpret directly.

Our method builds on recent advances in explainable AI and Boolean network inference in ecology [28,3]. We incorporate species-specific discretization strategies informed by ecological knowledge to ensure biologically meaningful input representations. Unlike black-box models, our approach produces symbolic rules that are both data-driven and interpretable.

In addition to modeling phytoplankton dynamics, a key goal of this study is to provide a general framework that is modular and applicable to a broad range of ecological time series involving species abundance (number of individuals per water volume) or presence–absence data.

Our main contributions are the following:

- We develop a species-specific discretization strategy tailored for symbolic learning from ecological data.
- We apply LFIT to long-term phytoplankton time series to infer interspecies interaction rules.
- We construct a signed and weighted interaction graph from the extracted rules, using support and confidence to quantify interaction strength.

We provide source code for all the steps of our approach under the form of Python notebooks available online at: `https://zenodo.org/records/15389109`.

The rest of this paper is organized as follows: Section 2 reviews related work on ecological modeling and symbolic learning. Section 3 describes the data sources, preprocessing steps and the species-specific discretization. Section 4 details the symbolic learning framework and interaction graph construction. Section 5 presents our results and interpretations. We discuss limitations and conclude in Section 6 with prospects for future work.

2 Background and Related Work

2.1 Modeling Approaches in Marine Ecosystems

Various modeling approaches have been used to study phytoplankton dynamics. Ecosystem models (systems of ordinary differential equations based on reaction networks) are constructed to predict the responses of marine ecosystems to global changes. They typically account for the interplay between nutrients, phytoplankton, zooplankton and detritus (also called NPZD models) depending on temperature, light, hydrodynamics and some other abiotic components, but largely simplify planktonic biodiversity to a few functional groups (a set of species with similar characteristics) and translate biotic interactions only through the lens of predation or competitive exclusion [14,19,5]. Facilitation [9] or allelopathy [35] are other known mechanisms of biotic interaction that are difficult to evidence from field approaches or experiments.

These models are theory-driven and interpretable but require strong assumptions and detailed parameterization, which can limit their applicability to real-world ecosystems showing high biodiversity. Although intensively used to predict the effects of climate change for example (they are included in earth system and climate models used by the GIEC for predictions of global temperature increase, see for example [23]), these models are in reality of poor predictive power of changes in the phytoplankton community structure and therefore of the future functioning of marine ecosystems. Models of "infinite biodiversity" [1,11,5], i.e. NPZD models with more than 30 phytoplankton species, were developed so far to understand how phytoplankton community will change under the climate induced alterations of oceanic conditions. Parametrization of such models is idealized by using size of the phytoplanktonic cells as a driver of all parameters. They are used for theoretical questions in ecology, not for practical ones.

Statistical models (e.g., GLMs, GAMs, multivariate analysis and niche models) offer flexibility for analyzing species–environment relationships but often struggle with nonlinearities and high-dimensional data [10,26]. More recently, trait-based ecological modeling has offered an alternative perspective: instead of focusing on species identity alone, models focus on functional traits that interaction species' responses to their environment (e.g., thermal tolerance range, cell size [7]). This framework has been used to delineate the realized niches of phytoplankton taxa, that is, the set of environmental conditions under which

these taxa are actually observed to persist, taking into account both abiotic (non-living) and biotic (living) factors [20,22,10]. Trait-based modeling has supported recent large-scale analyses of community composition and shifts in French coastal waters [22,24]. Although these methods have highlighted the global importance of biotic interactions in population dynamics, it is unable to precisely characterize them individually.

More recently, machine learning methods have increasingly been used for ecological forecasting and pattern discovery from high-dimensional time series [26,4]. These models can capture nonlinear relationships and scale well with data, but they generally lack interpretability, making it difficult to extract mechanistic insights or generate ecological hypotheses. This is particularly limiting in marine ecology, where the identification of biotic interactions and feedbacks is essential for understanding community dynamics.

2.2 Symbolic and Constraint-Based Networks Inference

To overcome the limitations of black-box models, recent research has explored symbolic and logic-based methods for inferring interaction networks directly from observational data. These approaches aim to extract interpretable models encompassing the structure of the system, often under formal guarantees of consistency with observed dynamics.

CASPO (Cell ASP Optimizer) employs ASP to explore logical models consistent with perturbation-based experimental data, particularly in signaling networks [15]. It uses formal reasoning to identify minimal models that reproduce observed behavior, but its reliance on perturbation data prevents its application to natural ecological systems. MIIC (Multivariate Information-based Inductive Causation) is an information-theoretic method that reconstructs causal networks from multivariate observational data by combining conditional mutual information with constraint-based structure learning [34]. It introduced a feature to learn from time-series data only very recently, and this was applied to live-cell imaging data [31]. To our knowledge, these methods have not been applied to ecological data.

BoNesis [3] infers Boolean network models using Answer-Set Programming (ASP). It integrates prior knowledge, under the form of a prior knowledge network, with dynamical constraints, which can be derived from observed transitions. BoNesis explores the full set of Boolean networks that are consistent with these structural and dynamical constraints. It supports reasoning under uncertainty by producing ensembles of models that might be large and often requires additional assumptions to reduce ambiguity. The viability of this approach was demonstrated for biological networks, showing that it can generate minimal models compatible with observational constraints [2], and even more specifically for ecological networks inference from observational data [28]. Nevertheless, since this method requires a prior knowledge network, which is not available for biotic phytoplankton interactions, this method is not applicable to our work.

Qualitative symbolic frameworks, such as ASP-based modeling, are well suited to ecological applications, as they integrate domain knowledge, handle discrete

and noisy time series, and support reasoning under uncertainty [13,12]. These features are particularly relevant in planktonic systems, where both data sparsity and process complexity are prevalent. Our approach builds on this line of work by combining symbolic learning with a species-specific discretization strategy to generate interpretable interaction rules directly from real-world ecological data.

2.3 Qualitative and Explainable Approach: Learning from Interpretation Transitions (LFIT)

Among symbolic learning techniques, the Learning From Interpretation Transitions (LFIT) framework, originally introduced by Inoue et al. [18], has shown particular promise in inferring discrete dynamical systems from observational data. LFIT has been especially applied to Boolean networks and cellular automata [18], robotics [25], customer journey mapping [27]. It is particularly suited for biological systems, as they naturally accommodate asynchronous and non-deterministic dynamics—features common in gene regulation, signaling pathways, and ecological communities [29,18].

The LFIT framework takes as input a set of dynamical transitions, that is, pairs of states so that the system has been observed evolving from the first to the second. Such transitions can be simply extracted from a state graph or from time series observations, as in our case (see Figure 2 in Appendix). The framework performs logic rule refinement in order to output a set of logic rules that describe the local conditions on a variable to change its state.

To perform rule learning with LFIT, the main algorithm is GULA (General Usage LFIT Algorithm) by Ribeiro et al. [29]. GULA is a complete and sound learning algorithm that supports memory-less semantics and produces all minimal rules that explain the transitions. However, it suffers from exponential complexity, which limits its scalability to systems with a very low number of variables, typically less than 15. Since our model contains 66 variables variables, this algorithm cannot be used.

The scalability of LFIT was significantly improved by the PRIDE algorithm [30], which offers a polynomial-time learning approach by trading completeness for tractability. PRIDE is particularly suitable for applications in large or complex ecological systems, at the cost of potentially missing some explanations. What explanations are prioritized is given by an ordered list of variables given to PRIDE: the variables appearing first are considered first to construct the logical rules and might be strongly overrepresented compared to the last variables. This requires comparing several runs with different orderings for a complete view of the results.

A prior attempt to apply symbolic learning to ecological systems was proposed by Iken et al. [17], who explored the use of LFIT for rule extraction in a phytoplankton context. Their work introduced heuristics for rule simplification and interaction graph construction but remained preliminary and limited in scope. In this paper, we extend that direction by applying a refined LFIT-based framework, incorporating species-specific discretization and validating the output through ecological interpretation and interaction graph analysis.

In this context, our work leverages symbolic learning to generate interpretable interaction graphs from phytoplankton time-series data. By incorporating expert-informed discretization and formal rule extraction, we aim to provide biologically grounded representations of interspecies dynamics in coastal ecosystems.

3 Data and Discretization

3.1 Ecological Dataset

This study is based on long-term phytoplankton monitoring data from the SRN program (Regional Observation and Monitoring Program for Phytoplankton and Hydrology in the eastern English Channel [32,24]). The dataset comprises time series of environmental measurements and species-level phytoplankton counts sampled approximately every 15–30 days from 1992 to 2020 along the eastern English Channel coast, with a particular focus on the Boulogne-sur-Mer coastal station. We consider only surface-level observations and restrict our analysis to 12 phytoplankton species and 11 abiotic factors that are both commonly available and ecologically meaningful, following the selection in [20]. These 12 species were thus chosen due to their ecological relevance, and consistent measures.

The biotic variables consist of species-level abundance, expressed as cell counts, and represent macro-level community structure rather than biomass. The selected species include key diatoms such as *Chaetoceros danicus*, *Guinardia delicatula*, *Skeletonema*, and *Pseudo-nitzschia seriata*, among others, as well as representatives of prymnesiophytes, notably *Phaeocystis*.

The abiotic variables considered are: temperature (TEMP), salinity (SALI), nitrite (NO_2), nitrate (NO_3), ammonia (NH_4), phosphate (PO_4), silanol ($Si(OH)_4$), turbidity (TURB), and other light-related proxies such as organic and inorganic suspended matter (MESORG, MESINORG). These variables were selected for their known interaction on phytoplankton dynamics, supported by ecological literature and theoretical growth models [8,19].

To accommodate irregular sampling and missing values, we resampled the data to a uniform monthly time step and used linear interpolation to impute missing measurements. While imputation inevitably introduces some uncertainty, this approach preserves the temporal structure needed for state transition learning.

Before applying symbolic learning, we performed a preliminary analysis to assess whether including other species as predictors improves the ability to model the dynamics of a target species. This approach, inspired by previous work on data-driven interaction inference [26], quantifies how much additional variance can be explained when biotic variables are included alongside abiotic factors. For each species in our dataset, we trained two random forest models: one using only abiotic variables, and one including both abiotic and species abundance variables. We report the resulting R^2 scores in Table 1 in Appendix. The coefficient of determination R^2 measures the proportion of variance in the target variable that is explained by the model. An R^2 of 1 indicates perfect prediction, while an R^2 of 0 means the model performs no better than the mean.

In nearly all cases, including species variables led to improved predictive performance, suggesting that interspecies influences play a significant role in shaping community dynamics. Note that in complex ecological systems, low R^2 values are common due to noise and unmeasured variables. Here, even modest improvements support the presence of informative structure in the data, especially for large datasets. These results support the hypothesis that species interactions are embedded in the observational data and justify the use of an explainable learning framework to characterize them.

3.2 Discretization Strategy

LFIT requires all variables to be expressed in discrete states. For species abundance, we convert the time series into Boolean values (presence/absence) using a quantile-based discretization: values above a defined threshold are assigned 1 (presence), the rest 0.

For abiotic variables related to temperature and nutrients, we apply a species-specific discretization strategy informed by theoretical growth models. Each abiotic variable is duplicated per species (e.g., T_a, T_b, etc.) and discretized according to that species' theoretical physiological response. This enables the learning algorithm to consider the same environmental condition differently for different species.

We rely on known functional forms—such as Arrhenius-type responses for temperature and Michaelis-Menten kinetics for nutrients [7,1]—and align them with empirical species presence distributions to validate the relevance of the thresholds.

Temperature Effect. Temperature response is modeled using a Gaussian growth curve specific to each species [8]:

$$f_{\text{temp}}(T) = \exp\left(-\frac{(T - T_{\text{opt}})^2}{2\sigma^2}\right) \qquad (\text{in days}^{-1})$$

This captures unimodal thermal responses around a species-specific optimal temperature T_{opt} with tolerance σ. Thresholds for Boolean discretization are chosen as ± 1 standard deviation around T_{opt}, defining the range of favorable conditions with assigned value 1. Outside this range, values are assigned 0. Figure 1(a) shows this alongside observed presence rates.

The species occurrences distribution (presence rate per unit of temperature) is used to verify that the theoretical bins align with observed patterns. Presence rates in the dataset are not expected to fit the theoretical growth potential. The theoretical growth represents the species' *growth* potential under otherwise optimal conditions (normalized number of individuals per time unit) while the *presence rate* of the species gives the number of observations featuring the presence of the species above a certain threshold for a given temperature. Although it is expected that presence rates are higher when the theoretical growth rate is high, in practice, presence rates are lower than what one could expect due to

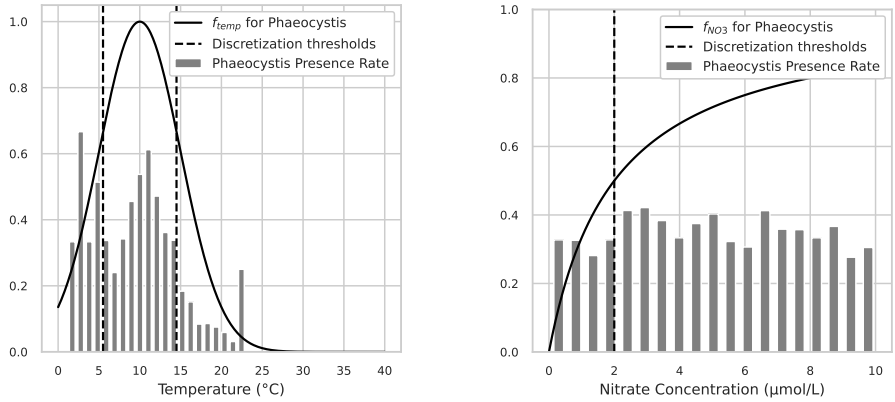
additional limiting factors—such as other environmental constraints and interspecies interactions—as well as natural fluctuations of temperature over time.

Nutrient Limitation. Nutrients (i.e., NO_2, NO_3, NH_4, PO_4 and $Si(OH)_4$) are modeled using a Michaelis-Menten function:

$$f_X([X]) = \frac{[X]}{[X] + K_X} \qquad \text{(dimensionless)} \tag{1}$$

where $[X]$ is the concentration of the nutrient and K_X its half-saturation constant.

For three nutrient variables deemed the most important, namely NO_3, NH_4 and PO_4, discretization thresholds were defined based on their respective half-saturation constants (K_X). Figure 1(b) shows this applied to nitrate (NO_3) for *Phaeocystis*. These constants represent the nutrient concentration at which the growth rate reaches half its maximum value and are widely used in ecological models to characterize nutrient limitation [1].



(a) Theoretical temperature growth function and observed presence rate of *Phaeocystis*.

(b) Theoretical nitrate limiting function and observed presence rate of *Phaeocystis*.

Fig. 1. Theoretical responses to abiotic factors (black lines) vs. observed presence rate of species (gray bars) for *Phaeocystis*. Dashed lines indicate discretization thresholds.

The distribution of species occurrences for temperature shows a closer alignment with the theoretical growth function than for nutrient variables. This is consistent with the strong thermal sensitivity of phytoplankton species, which exhibit narrow optimal temperature ranges and sharp declines outside of those conditions [8,33]. In contrast, nutrient uptake responses are generally more flexible.

To sum up, our model encompasses 66 variables: 12 phytoplankton species, 4 abiotic factors (temperature and 3 nutrients) duplicated into as many variables as there are species, and the remaining 6 abiotic factors.

4 Interaction Graph Extraction from LFIT program

4.1 Rule inference with LFIT

The rules inferred by LFIT form a logic program (called *Dynamical Multi-Valued Logic Program*, or DMVLP in [29]) where each rule r has the general form:

$$(v^{t+1} = val) \leftarrow (v_1^t = val_1) \wedge (v_2^t = val_2) \wedge \cdots \wedge (v_m^t = val_m) \tag{2}$$

Such a rule means that variable v will take value val at time $t + 1$ (left hand side, denoted head(r)) if all the conditions in the body (right-hand side, denoted body(r)) are satisfied at time t. The framework supports multi-valued variables and non-deterministic updates, making it particularly well-suited for modeling complex biological systems.

In addition to inferring a logic program, we use the extended framework of *weighted logic programs* (WDMVLP) [29] where each rule is assigned a weight that reflects how often the rule's body (conditions) is satisfied in the observed data — a measure of its empirical coverage. A WDMVLP is composed of:

- Likeliness rules — from transitions observed in the data;
- Unlikeliness rules — from transitions never observed.

The weights serve as a proxy for rule reliability and enable probabilistic prediction by contrasting the total support of contradictory rules. In the WDMVLP framework, unlikeliness rules—representing transitions not observed in the dataset—are used to improve predictive accuracy. However, only likeliness rules are considered for the construction of interaction graphs below, as they correspond to transitions actually observed in the data and therefore represent plausible ecological interactions. These rules form the starting point of graph construction in our approach.

In this work, we exclusively use the PRIDE algorithm, given the performance issues of GULA when applied to systems of more than 15 variables. As a result, the logic programs of the WDMVLP are not complete and might thus miss some explanations. To compensate for this limitation, we perform multiple runs of PRIDE with different variable orderings and take the union of the resulting rule sets, thereby approximating broader coverage while retaining polynomial-time complexity.

4.2 Interaction Graphs Extraction

Although singular rules produced by LFIT are human-readable, a WDMVLP can be very large: 20 000 rules in total in the case of this work. Summing all the information of the interactions into a graph is therefore beneficial, but such

a translation is not trivial. In the following, we propose a method to produce a signed and weighted interaction graph where each node represents a variable (e.g., a phytoplankton species) and directed and weighted edges represent interactions inferred from the learned rules.

Edge weight computation. To construct the interaction graph, we assign a weight to each directed edge based on the rules in the WDMVLP. For each pair of variables (i, j), we define the edge $i \to j$ and compute a total weight $w_{i\to j}$, which aggregates the contributions of all rules where variable i appears in the body and variable j appears in the head. We denote the set of such rules as $\mathcal{R}_{i\to j}$. For each rule $r \in \mathcal{R}_{i\to j}$, we compute a rule weight $w(r)$, which represents the contribution of rule r on the total edge weight $w_{i\to j}$. To calculate $w(r)$, we need to define the following metrics:

- The **coverage** of a rule r, noted coverage(r), is the number of transitions in the dataset for which the rule's body is satisfied. Coverage reflects how many times a particular condition appears in the observed data.
- The **support** of a rule r, noted support(r), is the number of transitions in the dataset for which the whole rule is satisfied: its body is satisfied at time t and its head at time $t+1$. This reflects how many times the rule is realized.

From these metrics, we can derive the **confidence** of a rule r representing the proportion of realization of this rule in the dataset. In other words, it represents how often the rule's head at time $t + 1$ is realized, knowing that the body held at time t. The confidence is defined as the conditional probability:

$$P(\text{head}(r) \mid \text{body}(r)) = \frac{\text{support}(r)}{\text{coverage}(r)} \tag{3}$$

To further emphasize the specificity of the rule, we normalize this by the marginal probability of the head $P(\text{head}(r))$, and multiply by the support, thus:

$$w(r) = \text{support}(r) \cdot \frac{P(\text{head}(r) \mid \text{body}(r))}{P(\text{head}(r))} \tag{4}$$

The normalization helps to down-weight rules that match only a few transitions in the dataset, which may be due to noise or rare events. Rules with higher support are generally more robust, as they apply more often, and thus reflect recurring dynamics. This usually makes a shorter body containing less conditions to match, and these conditions imply a stronger interaction on the head variable. Also, the multiplication by the support emphasizes how strongly the body predicts the head relative to how often the head occurs overall. This allows us to favor rules whose predictive power exceeds what would be expected from the marginal distribution of the head variable alone. Both components are essential: support ensures statistical relevance and generality, while normalized confidence ensures predictive reliability. The combination yields an interpretable measure of interaction strength in the resulting graph.

Finally, the general weight $w_{i\to j}$ is given by:

$$w_{i\to j} = \sum_{r\in\mathcal{R}_{i\to j}} w(r) \tag{5}$$

Edge sign computation. In addition to weighting edges, we assign each one a signed weight, which is an unbounded positive or negative value representing the nature of the interaction. To determine the signed weight of an edge $i \to j$, we compute a signed weight from each rule $r \in \mathcal{R}_{i\to j}$. The signed weight of a rule r on $i \to j$ is computed as:

$$\text{sign}_{i\to j}(r) = \delta_{i\to j}(r) \cdot w(r) \tag{6}$$

where:

- $\delta_{i\to j}(r)$ equals -1 if the head j and body variable i differ in value (e.g., one is high while the other is low) and 1 if they are aligned.
- $w(r)$ is the (positive) weight of the rule, calculated as described above.

The idea behind $\delta_{i\to j}(r)$ is that a positive influence from i to j is characterized by variable i pushing j towards the same value. On the contrary, if an increase of i makes j decrease, or conversely if a decrease of i makes j increase then this rather characterizes a negative influence.

Example 1. For instance, let us consider a likeliness rule r of this form: $(v^{t+1} = 1) \leftarrow (v_1^t = 1) \wedge \ldots$ The signed weight of this rule r on edge $v_1 \to v$ is: $\text{sign}_{v_1\to v}(r) = 1 \cdot w(r)$ since the values of v^{t+1} and v_1^t are aligned.

The total signed interaction from variable i to variable j is then the sum of all signed rule contributions:

$$\text{sign}_{i\to j} = \sum_{r\in\mathcal{R}_{i\to j}} \text{sign}_{i\to j}(r)$$

Representation in the Graph To sum up, this section, for each directed edge $i \to j$ in the influence graph, two complementary measures are derived from the set of rules $\mathcal{R}_{i\to j}$ (i.e., where variable i appears in the body and j in the head).

The first is the edge **weight**, defined as the sum of the weights of all such rules. This value reflects the overall strength of the inferred interaction, and is used to determine the **thickness** of the edge in the graph.

The second is the edge **signed weight**, which incorporates the direction of influence. For each rule, we assign a sign (+1 or –1) based on the type of interaction (i.e., alignment or opposition of variable values), and multiply it by the rule's weight. The signed weight is then the sum of these signed contributions across all relevant rules. This value is mapped to the edge **color**: green for positive interactions, red for negative ones, and gray for ambiguous or mixed interactions (i.e., when the signed weight is close to zero).

This dual encoding allows us to distinguish between interactions that are consistently positive or negative (strongly colored, but not necessarily thick), and those that are strongly supported but directionally ambiguous (thick but near-gray).

This graph provides an interpretable, rule-based summary of system dynamics, enabling further analysis of community interactions and key drivers in ecological networks. The graphs obtained in this work are presented in the next section.

5 Phytoplankton interaction Graphs

5.1 Application of PRIDE

The PRIDE algorithm, while efficient and scalable, is sensitive to the order in which input variables are provided. It prioritizes explanations based on the early variables in the input list, which can lead to incomplete coverage of potential rule sources. To mitigate this bias and better approximate the exhaustive rule set produced by GULA, we performed 5 independent runs of PRIDE, each using a different random permutation of the variable ordering. The variable orderings and the corresponding accuracy of each model are reported in Table 2 in the Appendix. In all runs, abiotic variables were placed first to ensure that the algorithm first attempts to explain species dynamics through environmental factors. This prioritization helps reduce the risk of inferring spurious biotic interactions that may in fact result from shared responses to abiotic drivers. We then took the union of all resulting rules across these runs to construct a more comprehensive logic program. Since PRIDE outputs minimal rules, this aggregation does not introduce overlap. Importantly, this approach preserves PRIDE's polynomial-time complexity, unlike GULA's exponential cost, and remains tractable for our dataset. The aggregated model achieved an accuracy of 0.86, outperforming each individual run. This suggests that combining the rules improves overall predictive performance.

5.2 Interaction Graph over all Phytoplankton Species

We first present the global interaction graph generated from the symbolic learning process, which captures both biotic and abiotic interactions across all species in the dataset (Figure 3 in Appendix). In this graph, each node corresponds to a phytoplankton taxon, and directed edges represent inferred interactions derived from logical rules. For readability reasons, abiotic factors have not been represented here.

Edge weight and signed weight were calculated using the weighted rule aggregation method described in Section 4.

It is important to note that while these graphs depict directed influences between species, they do not directly indicate specific ecological interaction types such as competition, allelopathy, predation, or facilitation. The symbolic learning

framework identifies patterns consistent with influence—i.e., how the presence of one species may relate to changes in another—but cannot distinguish whether this influence is direct or indirect. For example, an inferred interaction may result from a shared dependency on a resource, rather than a direct species-to-species influence.

Moreover, while our ultimate goal is to uncover causal ecological relationships, the LFIT-based framework and PRIDE algorithm used here cannot guarantee causality. The rules capture patterns in observed transitions, but without further constraints or integration of ecological prior knowledge, causal inference remains out of reach.

In this context, expert knowledge is essential to interpret the inferred graphs meaningfully. Ecologists familiar with the system must assess whether observed patterns align with known mechanisms, suggest new hypotheses, or may reflect confounding environmental structures. Therefore, we view these graphs as a hypothesis-generating tool, rather than a definitive mapping of ecological interaction types.

5.3 Species-Centered Graphs: Focus on Target Species

To improve interpretability, we extract subgraphs centered on individual target species. Each of these visualizations shows all inferred interactions from other species to a single focal species. For each graph, edge thickness and color intensity are rescaled independently based on the minimum and maximum values within that graph. This local normalization facilitates comparison of relative interaction strengths and signs within each subgraph, while preventing domination by extreme values present in other graphs.

Figure 4 in Appendix shows the inferred interaction graph for *Phaeocystis*. Some interactions appear consistent with previous ecological interpretations and results reported in [20]. That study identified two different environmental trajectories: one associated with years of high *Phaeocystis* abundance and another with low abundance, based on combinations of abiotic conditions such as irradiance, turbidity, and nutrient concentrations. Species were grouped according to their presence along these trajectories, reflecting their realized ecological niches.

According to that study, certain species such as *Skeletonema*, *Thalassionema nitzschioides*, *Paralia sulcata*, *Guinardia striata*, and *Guinardia delicatula* appear in both high and low *Phaeocystis* abundance trajectories. Their ecological niches are broad enough to accommodate a range of conditions, including those that precede *Phaeocystis* blooms. As such, their impact on *Phaeocystis* dynamics might be relatively neutral in the inferred graph, although there are some exceptions, such as a weak positive interaction from *Guinardia delicatula*.

In contrast, species that appear only in the high-abundance trajectory, such as *Ditylum*, *Chaetoceros danicus*, *Nitzschia longissima*, and *Leptocylindrus danicus*, are expected to have a more clearly positive influence. They occupy realized niches that align with the environmental trajectory preceding a *Phaeocystis* bloom. These species are associated with increasing irradiance, reduced turbidity, and moderate nutrient levels, conditions favorable to bloom development. This

is reflected for *Ditylum* and *Chaetoceros danicus* in the interaction graph, but not as clearly for *Nitzschia longissima* and *Leptocylindrus danicus*. In the case of *Leptocylindrus danicus*, its appearance later in the season, after *Phaeocystis* blooms, could explain the observed negative interaction.

Finally, species restricted to low-abundance years, such as *Thalassionema gracilis*, tend to show negative interactions with *Phaeocystis*. An exception is *Pseudo-nitzschia seriata*, which is sometimes observed during bloom events and is known to physically insert itself into *Phaeocystis* colonies, potentially explaining its weak positive effect.

It is important to note that our inference method does not model delayed effects. Therefore, some inferred interactions may not align directly with the mechanistic hypotheses proposed in [20], particularly those involving time-lagged processes.

5.4 Abiotic Factors Influence Graphs

In addition to the species interaction graph, we can also look at the abiotic factors graph, representif the potential effects of abiotic variables on *Phaeocystis*, shown in Figure 5 in Appendix. This graph serves as a form of validation for the method, since the relationships between *Phaeocystis* and abiotic factors are relatively well understood. As expected, we observe positive effects from NO_2 and PO_4, which are typically associated with nutrient enrichment preceding blooms. Conversely, *SiOH* (silicate) has a negative effect, likely due to its association with diatom growth: as long as silicate is abundant, diatoms thrive, often outcompeting *Phaeocystis*. Similarly, high salinity and elevated NH_4 concentrations correspond to summer conditions when *Phaeocystis* is typically absent, and are therefore also associated with negative effects in the model.

6 Conclusion and Future Works

This study introduces a symbolic learning framework for inferring interspecies and environmental interactions in phytoplankton communities, using long-term observational data from the SRN [32] monitoring program. Building on the LFIT framework and its most scalable algorithm, PRIDE, we computed interpretable logic rules. Our method differs from black-box machine learning in that it emphasizes explainability and rule-level insight. In particular, we introduced a rule weighting scheme based on both rule support and normalized confidence, from which we extracted biotic and abiotic interactions in graph form. These graphs offer a compact, readable representation of species dynamics that complements classical modeling approaches and provides domain experts with a new lens for ecological interpretation. While our framework ultimately aims to capture causal ecological dynamics, it currently does not enforce constraints that would guarantee causality. As such, expert ecological knowledge remains essential to interpret the interaction graphs and to hypothesize the possible underlying mechanisms behind the observed patterns. For certain focal species, such as *Phaeocystis*, the

resulting graphs highlighted at least one interaction in agreement with previous work [20]. The other interactions might be new information that requires deeper expert analysis. The source code for this work is available online as Python notebooks at: `https://zenodo.org/records/15389109`.

These results open directions for future work. First, the algorithm used is memoryless and cannot account for lagged dependencies or historical patterns. Incorporating memory into the rule-learning process (e.g., using sliding windows or temporal embeddings) could improve accuracy and ecological plausibility. Second, light-related variables were not fully discretized due to data limitations, and further work is needed to apply trait-informed discretization consistently across all abiotic dimensions. Third, the current model learns from all transitions without incorporating known ecological constraints. Future versions could integrate prior knowledge to guide or restrict rule inference—for example, by excluding interactions between species with disjoint seasonal niches or well-documented independence. Finally, our method could also be extended to multi-valued logic variables, requiring to better fine-tune the discretization but also adapt the influence graph inference.

Overall, this work demonstrates the feasibility and value of symbolic learning in marine ecology. It bridges observational data and ecological theory through interpretable modeling and opens the door for more robust, reusable frameworks that combine data-driven learning with expert knowledge. We believe that this approach can be extended to a wide range of ecological systems, offering a new tool for understanding complex species interactions in dynamic environments.

Acknowledgments. This project has received financial support from the French National Research Agency (ANR) through project REBON (ANR-23-CE45-0008) and from the CNRS through the MITI interdisciplinary programs. The authors would like to thank Tony Ribeiro for his help and fruitful discussions.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bruggeman, J., Kooijman, S.: A biodiversity-inspired approach to aquatic ecosystem modeling. Limnology and Oceanography **52** (2007). `https://doi.org/10.4319/lo.2007.52.4.1533`
2. Chevalier, S., Noël, V., Calzone, L., Zinovyev, A., Paulevé, L.: Synthesis and simulation of ensembles of boolean networks for cell fate decision. In: Abate, A., Petrov, T., Wolf, V. (eds.) Computational Methods in Systems Biology. pp. 193–209. Springer International Publishing (2020). `https://doi.org/10.1007/978-3-030-60327-4_11`
3. Chevalier, S., Froidevaux, C., Paulevé, L., Zinovyev, A.: Synthesis of boolean networks from biological dynamical constraints using answer-set programming. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). pp. 34–41 (2019). `https://doi.org/10.1109/ICTAI.2019.00014`
4. Cutler, D., Edwards, T., Beard, K., et al.: Random forests for classification in ecology. Ecology **88**(11), 2783–2792 (2007). `https://doi.org/doi.org/10.1890/07-0539.1`

5. Dutkiewicz, S., Boyd, P., Riebesell, U.: Exploring biogeochemical and ecological redundancy in phytoplankton communities in the global ocean. Global Change Biology **27** (2020). https://doi.org/10.1111/gcb.15493
6. Edwards, K., Thomas, M., Klausmeier, C., Litchman, E.: Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton. Limnology and oceanography **57**, 554–566 (2012). https://doi.org/10.4319/lo.2012.57.2.0554
7. Edwards, K., Thomas, M., Klausmeier, C., Litchman, E.: Light and growth in marine phytoplankton: Allometric, taxonomic, and environmental variation. Limnology and Oceanography **60**, 540–552 (2015). https://doi.org/10.1002/lno.10033
8. Edwards, K., Thomas, M., Klausmeier, C., Litchman, E.: Phytoplankton growth and the interaction of light and temperature: A synthesis at the species and community level. Limnology and Oceanography **61**, n/a–n/a (2016). https://doi.org/10.1002/lno.10282
9. Emna, K., Rapaport, A., Le Floc'h, E., Fouilland, E.: Demonstration of facilitation between microalgae to face environmental stress. Scientific Reports (2019). https://doi.org/10.1038/s41598-019-52450-9
10. Fariñas, T., Bacher, C., Soudant, D., Belin, C., Laurent, B.: Assessing phytoplankton realized niches using a french national phytoplankton monitoring network. Estuarine Coastal and Shelf Science **159**, 15–27 (2015). https://doi.org/10.1016/j.ecss.2015.03.010
11. Follows, M., Dutkiewicz, S., Grant, S., Chisholm, S.: Emergent biogeography of microbial communities in a model ocean. Science (New York, N.Y.) **315**, 1843–6 (2007). https://doi.org/10.1126/science.1138544
12. Gaucherel, C., Fayolle, S., Savelli, R., Philippine, O., Pommereau, F., Dupuy, C.: Diagnosis of planktonic trophic network dynamics with sharp qualitative changes. Peer Community Journal **4**(e58) (2023). https://doi.org/10.24072/pcjournal.417
13. Gaucherel, C., Pommereau, F.: Using discrete systems to exhaustively characterize the dynamics of an integrated ecosystem. Methods in Ecology and Evolution **10** (2019). https://doi.org/10.1111/2041-210X.13242
14. Grangere, K., Lefebvre, S., Bacher, C., Cugier, P., Ménesguen, A.: Modelling the spatial heterogeneity of ecological processes in an intertidal estuarine bay: Dynamic interactions between bivalves and phytoplankton. Marine Ecology Progress Series **415**, 141–158 (2010). https://doi.org/10.3354/meps08659
15. Guziolowski, C., Videla, S., Eduati, F., Thiele, S., Cokelaer, T., Siegel, A., Saez-Rodriguez, J.: Exhaustively characterizing feasible logic models of a signaling network using answer set programming. Bioinformatics **30** (2013). https://doi.org/10.1093/bioinformatics/btt393
16. Houliez, E., Lefebvre, S., Dessier, A., Huret, M., Marquis, E., Bréret, M., Dupuy, C.: Spatio-temporal drivers of microphytoplankton community in the bay of biscay: Do species ecological niches matter? Progress In Oceanography **194**, 102558 (2021). https://doi.org/10.1016/j.pocean.2021.102558
17. Iken, O., Folschette, M., Ribeiro, T.: Automatic Modeling of Dynamical Interactions Within Marine Ecosystems. Master's thesis, University of Lille, France (2021)
18. Inoue, K., Ribeiro, T., Sakama, C.: Learning from interpretation transition. Machine Learning **94** (2014). https://doi.org/10.1007/s10994-013-5353-8
19. Vanhoutte-Brunier de Joux, A., Fernand, L., Ménesguen, A., Lyons, S., Gohin, F., Philippe, C.: Modelling the karenia mikimotoi bloom that occurred in the western english channel during summer 2003. Ecological Modelling (0304-3800) (Elsevier),

2008-02 , Vol. 210 , N. 4 , P. 351-376 **210** (2008). https://doi.org/10.1016/j.ecolmodel.2007.08.025

20. Karasiewicz, S., Breton, E., Lefebvre, A., Fariñas, T., Lefebvre, S.: Realized niche analysis of phytoplankton communities involving hab: Phaeocystis spp. as a case study. Harmful Algae **72**, 1–13 (2018). https://doi.org/10.1016/j.hal.2017.12.005
21. Karasiewicz, S., Dolédec, S., Lefebvre, S.: Within outlying mean indexes: refining the omi analysis for the realized niche decomposition. PeerJ **5**(e3364) (2017). https://doi.org/https://doi.org/10.7717/peerj.3364
22. Karasiewicz, S., Lefebvre, A.: Environmental impact on harmful species pseudonitzschia spp. and phaeocystis globosa phenology and niche. Journal of Marine Science and Engineering **10** (2022). https://doi.org/10.3390/jmse10020174
23. Kawamiya, M., Hajima, T., Tachiiri, K., Watanabe, S., Yokohata, T.: Two decades of earth system modeling with an emphasis on model for interdisciplinary research on climate (miroc). Progress in Earth and Planetary Science **7**(1) (2020). https://doi.org/10.1186/s40645-020-00369-5
24. Lefebvre, A., Devreker, D.: How to learn more about hydrological conditions and phytoplankton dynamics and diversity in the eastern english channel and the southern bight of the north sea: the suivi régional des nutriments data set (1992–2021). Earth System Science Data **15**, 1077–1092 (2023). https://doi.org/10.5194/essd-15-1077-2023
25. Martínez, D., Alenyà, G., Ribeiro, T., Inoue, K., Torras, C.: Relational reinforcement learning for planning with exogenous effects. Journal of Machine Learning Research **18**(78), 1–44 (2017), http://jmlr.org/papers/v18/16-326.html
26. Mutshinda, C., Finkel, Z., Widdicombe, C., Irwin, A.: Phytoplankton traits from long-term oceanographic time-series. Marine Ecology Progress Series **576** (2017). https://doi.org/10.3354/meps12220
27. Okazaki, K., Inoue, K.: Explainable model fusion for customer journey mapping. Frontiers in Artificial Intelligence **5** (2022). https://doi.org/10.3389/frai.2022.824197
28. Paulevé, L., Gaucherel, C.: Inference of ecological networks and possibilistic dynamics based on boolean networks from observations and prior knowledge. Tech. rep., University of Bordeaux (2024). https://doi.org/10.1101/2024.07.01.601264
29. Ribeiro, T., Folschette, M., Magnin, M., Inoue, K.: Learning any memory-less discrete semantics for dynamical systems represented by logic programs. Machine Learning **111**, 1–78 (2021). https://doi.org/10.1007/s10994-021-06105-4
30. Ribeiro, T., Folschette, M., Magnin, M., Inoue, K.: Polynomial algorithm for learning from interpretation transition. 1st International Joint Conference on Learning & Reasoning (2021). https://doi.org/hal-03347026
31. Simon, F., Comes, M.C., Tocci, T., Dupuis, L., Cabeli, V., Lagrange, N., Mencattini, A., Parrini, M.C., Martinelli, E., Isambert, H.: CausalXtract, a flexible pipeline to extract causal effects from live-cell time-lapse imaging data. eLife **13**, RP95485 (2025). https://doi.org/10.7554/eLife.95485
32. SRN-Regional Observation And Monitoring Program For Phytoplankton And Hydrology In The Eastern English Channel: SRN dataset - Regional Observation and Monitoring Program for Phytoplankton and Hydrology in the eastern English Channel (2025). https://doi.org/10.17882/50832, https://www.seanoe.org/data/00397/50832/

33. Thomas, M., Kremer, C., Litchman, E.: Environment and evolutionary history determine the global biogeography of phytoplankton temperature traits. Global Ecology and Biogeography **25**, 75–86 (2016). https://doi.org/10.1111/geb.12387
34. Verny, L., Sella, N., Affeldt, S., Singh, P., Isambert, H.: Learning causal networks with latent variables from multivariate information in genomic data. PLOS Computational Biology **13**, e1005662 (2017). https://doi.org/10.1371/journal.pcbi.1005662
35. Śliwińska Wilczewska, S., Wisniewska, K., Konarzewska, Z., Cieszyńska, A., Barreiro Felpeto, A., Lewandowska, A., Latała, A.: The current state of knowledge on taxonomy, modulating factors, ecological roles, and mode of action of phytoplankton allelochemicals. Science of The Total Environment **773**, 145681 (2021). https://doi.org/10.1016/j.scitotenv.2021.145681

Appendix

Time	a	b	c	d
1	0	0	1	1
2	1	1	0	0
3	0	0	0	1
4	0	1	1	0

$$\begin{pmatrix} a^0 \\ b^0 \\ c^1 \\ d^1 \end{pmatrix} \longrightarrow \begin{pmatrix} c^0 \\ d^0 \end{pmatrix}$$

$$\begin{pmatrix} a^1 \\ b^1 \\ c^0 \\ d^0 \end{pmatrix} \longrightarrow \begin{pmatrix} c^0 \\ d^1 \end{pmatrix}$$

$$\begin{pmatrix} a^0 \\ b^0 \\ c^0 \\ d^1 \end{pmatrix} \longrightarrow \begin{pmatrix} c^1 \\ d^0 \end{pmatrix}$$

Fig. 2. Illustration of a discrete multivariate time series (left) and its corresponding set of state transitions (right). Variables a and b represent abiotic factors, while variables c and d represent phytoplankton species. Each row in the table shows variable values at a time step. Each transition shows how the system state changes from one time to the next, the target variables being species. In our application, all variables take values in $\{0, 1\}$.

Table 1. Comparison of R^2 scores from random forest models predicting each species. "Abiotic" includes only abiotic factors as predictors; "Abiotic + Biotic" includes both abiotic factors and the other 11 species as predictors. Including other species as predictors improves the ability to model the dynamics of a target species.

Target species	**Abiotic**	**Abiotic + Biotic**
Chaetoceros danicus	0.16	0.21
Ditylum	0.17	0.33
Guinardia delicatula	0.18	0.30
Guinardia striata	0.12	0.17
Leptocylindrus danicus	−0.007	0.12
Nitzschia longissima	0.05	0.20
Paralia sulcata	0.08	0.16
Pseudo-nitzschia seriata	0.13	0.13
Skeletonema	0.01	0.11
Thalassionema nitzschioides	0.12	0.22
Thalassiosira gravida	0.20	0.37
Phaeocystis	0.50	0.47

Table 2. Variable orders and model accuracies for different PRIDE runs. The last row reports the accuracy of the aggregated model obtained by taking the union of the rules from all five runs.

Run	**Variable Order**	**Accuracy**
1	SIOH, NO2, NH4, TEMP, PO4, SALI, TURB, NO3, MESINORG, MESORG, Nit. lon., Gui. str., Gui. del., Ske., Lep. dan., Tha. gra., Pseudo-nit. ser., Dit., Phaeocystis, Chae. dan., Par. sul., Tha. nit.	0.67
2	PO4, MESORG, TURB, MESINORG, NH4, NO3, SIOH, NO2, SALI, TEMP, Tha. gra., Tha. nit., Dit., Pseudo-nit. ser., Ske., Par. sul., Chae. dan., Phaeocystis, Gui. str., Gui. del., Lep. dan., Nit. lon.	0.67
3	NH4, SALI, MESORG, SIOH, TEMP, NO2, PO4, MESINORG, NO3, TURB, Gui. del., Tha. gra., Pseudo-nit. ser., Dit., Par. sul., Chae. dan., Phaeocystis, Gui. str., Nit. lon., Tha. nit., Lep. dan., Ske.	0.67
4	MESINORG, NO3, PO4, TEMP, SIOH, TURB, NH4, SALI, MESORG, NO2, Chae. dan., Pseudo-nit. ser., Phaeocystis, Par. sul., Ske., Nit. lon., Tha. nit., Gui. del., Dit., Lep. dan., Tha. gra., Gui. str.	0.68
5	SALI, SIOH, TURB, MESORG, MESINORG, NH4, PO4, NO2, TEMP, NO3, Chae. dan., Phaeocystis, Gui. del., Tha. gra., Ske., Pseudo-nit. ser., Tha. nit., Nit. lon., Gui. str., Dit., Lep. dan., Par. sul.	0.67
Aggregated model (union of rules from all runs)		**0.86**

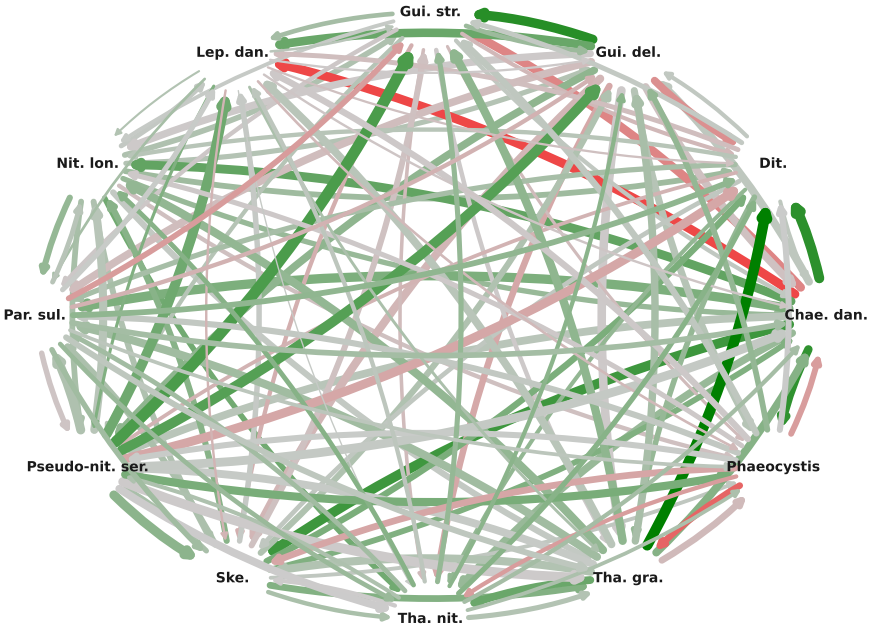


Fig. 3. Global interaction graph among phytoplankton species. Edge color indicates interaction sign (green for positive, red for negative and gray for unsure), and thickness reflects interaction strength.

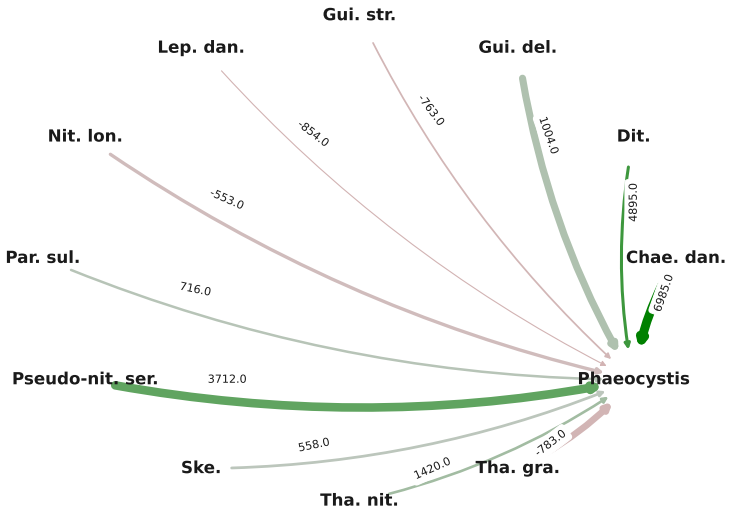


Fig. 4. Interaction graph of the biotic variables targeting *Phaeocystis*. Edge label indicates interaction sign strength (associated with edge color), and thickness reflects interaction strength.

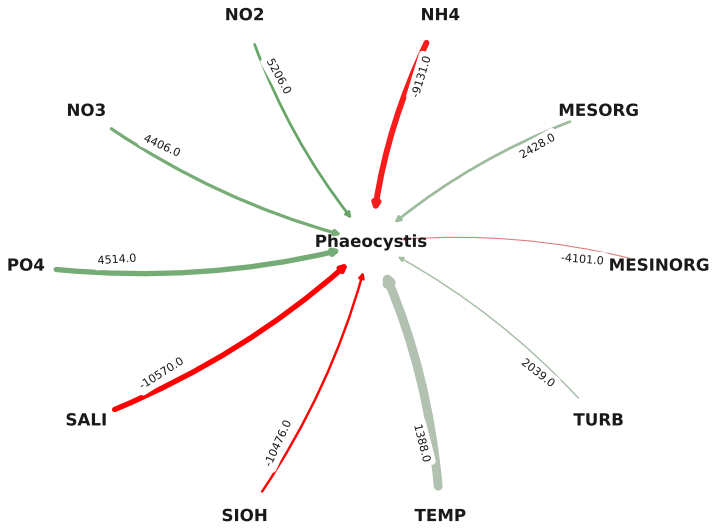


Fig. 5. Interaction graph of the abiotic variables targeting *Phaeocystis*. Edge label indicates interaction sign strength (associated with edge color), and thickness reflects interaction strength.