Symbolic Regression Methods for Oceanic Modeling

Hernán Lira, Inria Chile

August 2025

1 Problem Introduction

Marine microbial communities drive essential ecosystem functions, including nutrient cycling, carbon export, and oxygen production. Their metabolic activity plays a central role in regulating biogeochemical processes and in shaping the ocean's response to global environmental change. Understanding how environmental gradients—such as temperature, nutrient concentrations, and oxygen availability—influence microbial metabolism is therefore key for predicting the dynamics of marine ecosystems. This work focuses on modeling the relationship between environmental variables and the abundance of molecular functions across vertical ocean layers, aiming to decode how microbial communities function and adapt throughout the water column [**?**].

At the cellular level, molecular functions—defined as the specific biochemical activities carried out by gene products—operate in a coordinated manner within metabolic pathways, which orchestrate complex reactions essential for adaptation and survival. These pathways govern fundamental processes such as energy production, nutrient assimilation, and stress response. However, accurately modeling how these pathways respond to environmental drivers remains a major challenge due to the hierarchical structure of biological systems and the nonlinear nature of molecular-environment interactions. In marine ecosystems, this complexity is compounded by ocean depth stratification, which creates distinct physical and chemical regimes that influence metabolic activity [**?**].

2 Problem Definition

Let $\mathcal{D} = \left\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(L)}\right\}$ be a collection of datasets corresponding to L different environmental views, such as oceanic depth layers (*e. g.*, surface, deep chlorophyll maximum, mesopelagic).

Each view $\mathcal{D}^{(l)} = \left\{(\mathbf{x}_i^{(l)}, \mathbf{y}_i^{(l)})\right\}_{i=1}^{N_l}$ consists of:

- Environmental features $\mathbf{x}_i^{(l)} \in R^d$ (*e. g.*, temperature, oxygen, salinity),

- Molecular response variables $\mathbf{y}_i^{(l)} \in R^k$ (*e. g.*, gene/pathway abundances).

The goal of Multiview Symbolic Regression (MvSR) [**?**] is to learn a collection of interpretable symbolic functions

$$\mathcal{F} = \left\{ f_j^{(l)} : R^d \to R \right\}_{l=1,\ldots,L;\, j=1,\ldots,k} \tag{1}$$

that approximate the mapping from environmental variables to functional molecular profiles across views, subject to two central constraints:

1. *Interpretability*: Each $f_j^{(l)}$ must be a closed-form expression composed of elementary functions $\mathcal{B}$ such as $+$, $-$, $\times$, $\div$, log, exp, etc.

2. *Cross-view Structure*: The functions $f_j^{(l)}$ should respect hierarchical consistency, promoting both shared structure across views and view-specific refinements.

2.1 Symbolic Regression Search Process

The symbolic regression task [**?**] can be formulated as a multi-objective optimization problem

$$\min_{\mathcal{F}} \quad \mathcal{L}_{fit}(\mathcal{F}; \mathcal{D}) + \lambda \mathcal{L}_{comp}(\mathcal{F}) + \mu \mathcal{L}_{hier}(\mathcal{F}) \,, \tag{2}$$

where:

- $\mathcal{L}_{fit}$ measures the prediction error (*e. g.*, mean squared error across all views),

- $\mathcal{L}_{comp}$ penalizes expression complexity (*e. g.*, total number of nodes or depth of expression trees), and

- $\mathcal{L}_{hier}$ enforces structural similarity across views, promoting reuse of symbolic subexpressions.

We define $\mathcal{L}_{hier}$ using a tree-based distance (emphe. g., tree edit distance or shared subtree frequency) across expressions corresponding to the same function j across views as

$$\mathcal{L}_{hier}(\mathcal{F}) = \sum_{j=1}^{k} \sum_{l,l'} d_{tree}\left(f_j^{(l)}, f_j^{(l')} \right) . \tag{3}$$

2.2 Biological Motivation: Hierarchical Organization of Oceanic Metabolism

Marine microbial communities mediate essential ecosystem functions, including nutrient cycling, carbon export, and oxygen production. Understanding the relationship between environmental conditions and microbial metabolic activity

is crucial for predicting the dynamics of oceanic ecosystems in response to global change [**?**].

This work focuses on predicting the abundance of molecular functions based on environmental variables across vertical ocean layers. In metagenomics, a molecular function refers to the biochemical activity performed by a gene product (*e. g.*, "ATP binding", "oxidoreductase activity"), typically annotated via systems such as KEGG Orthology (KO) [**?**]. These molecular functions are organized into metabolic pathways, which are higher-order structures that represent a series of enzymatic reactions transforming inputs into biological end-products (*e. g.*, glycolysis, nitrification, sulfur metabolism).

Oceanic Metabolism and the Biological Carbon Pump Marine microbial communities play a key role in the biological carbon pump, a process that captures atmospheric CO_2 in surface waters and exports it to the deep ocean for long-term storage [**?**]. This pump relies on depth-stratified microbial functions such as carbon fixation, organic matter degradation, and anaerobic respiration, all shaped by environmental gradients.

In surface layers, photoautotrophic plankton drive carbon fixation and primary production. At the Deep Chlorophyll Maximum (DCM), intense nutrient assimilation occurs, including nitrogen and phosphate uptake. In the mesopelagic zone, metabolism shifts toward anaerobic processes like nitrification and sulfate reduction [**?**].

This study focuses on modeling the environmental drivers of these functions across depths to better understand their role in the carbon pump. H-MvSR provides interpretable expressions that link environmental variables to pathway-level microbial functions, offering insights into carbon cycling dynamics and supporting improved predictions for marine ecosystem monitoring and climate mitigation.

3 Data

The study utilizes the Ocean Microbial Reference Gene Catalog v2 (OM-RGC.v2) [**?**], integrating metagenomic data with 30 environmental parameters. The dataset includes 9024 KEGG Orthology (KO) groups mapped into 453 metabolic pathways, enabling the quantification of key ecological processes such as carbon fixation, nitrogen metabolism, and energy production. To ensure numerical stability, environmental variables were standardized.

The dataset was stratified into three depth layers: Surface (SRF, ≈ 5 meters deep), where phototrophic processes dominate; Deep Chlorophyll Maximum (DCM, ≈ 50 meters deep), marked by peak chlorophyll concentrations; and Mesopelagic (MES, from 200 to 1000 meters deep, with a median of 550 meters), where microbial communities rely on heterotrophic metabolism. This stratification ensures that depth-dependent variations in microbial function and environmental interactions are preserved.

4 Capstone

4.1 Study Symbolic Regression Background

4.1.1 Build a solid mental model (theory → algorithms → search spaces)

- Map the SR landscape: GP-based SR (tree search), sparse/linear SR (e.g., SINDy-style), neural/differentiable SR (e.g., policy-gradient or gradient-guided search), and hybrid neuro-symbolic pipelines.
- SR as multi-objective optimization: accuracy vs. parsimony (MDL/description length), plus domain constraints. Why Pareto fronts are the right abstraction for SR model selection.
- Catalog search spaces: operator sets, terminals, protected ops, constant treatment, grammar/shape constraints, units/dimensional analysis, monotonicity/convexity constraints. Understand how each choice shrinks/expands hypothesis space and affects identifiability.

Summary

- Classic SR: Tree-based representations, Optimization criteria, Evolutionary algorithms, Overfitting control, Interpretability and extrapolation.
- Neural Symbolic Regression: Equation learner (EQL) networks, Differentiable formulation, RL-based symbolic regression, Transformer/sequence models, Hybrid methods, Interpretability constraints.

4.1.2 Implement minimal SR systems

- GP-SR toy engine: polymorphic expression tree, safe ops, subtree crossover/mutation, tournament or lexicase selection, complexity penalties, Pareto archiving. Validate on synthetic equations with noise (vary SNR, outliers, heteroskedasticity).
- Reproduce baselines with mature toolkits: PySR, Operon/PyOperon, gplearn. Align metrics, search budgets, and operator sets so comparisons are apples-to-apples.
- Add a differentiable SR variant (e.g., symbolic templates + gradient tuning of constants), then compare convergence/stability vs. pure GP.

4.2 Literature Review

4.2.1 Symbolic Regression Papers

4.2.2 Applications in Environmental Sciences and Biology

Milestone: First report/presentation with background implementations and literature review. **Decide the approach**. Could be a systematic revision of

SR algorithms applied to our problem (catalog of models) or a novel implementation.

4.3 Work on the approach

4.4 After working implementations

- Evaluation (robustness, stability, interpretability)
- Selection, diagnostics, and ablations
- Domain integration (biological + multiview)