Exploring the nexus of omics, ecology, and phenotypes : highlighting phototrophic microeukaryotes through top-down metabolic modelling in marine plankton assessment

Marie Burel

université PARIS-SACLAY

Exploring the Nexus of Omics, Ecology, and Phenotypes: Highlighting Phototrophic Microeukaryotes through Top-Down Metabolic Modelling in Marine Plankton Assessment

Exploration de la convergence des omiques et des phénotypes écologiques: Avancées dans la modélisation métabolique ascendante pour les microeucaryotes phototrophes lors de l'évaluation du plancton marin

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Ecologie
Graduate School : Life Sciences and Health, Référent: Université d'Évry Val d'Essonne

Thèse préparée dans l'unité de recherche **Genomique Métabolique, Université Paris Saclay, Univ Evry, CNRS, CEA, Genoscope,** sous la direction d'**Eric Pelletier**, Directeur de recherche

Thèse soutenue à Paris-Saclay, le 29 juin 2023, par

Marie BUREL

Composition du Jury

Membres du jury avec voix délibérative

Chris BOWLER, Directeur de recherche, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Paris-Saclay	Président
Juan NOGALES, Investigateur principal, Consejo Superior de Investigaciones Científicas, Madrid, Espagne	Rapporteur & Examinateur
Sabine PERES, Professeur des universités, Université de Lyon 1	Rapporteur & Examinatrice
Dominique DE VIENNE, Professeur émérite, Université Paris-Saclay	Examinateur
Clémence FRIOUX, Chargée de recherche, Centre Inria de l'université de Bordeaux	Examinatrice

THÈSE DE DOCTORAT

NNT : 2023UPASL062

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Exploration de la convergence des omiques et des phénotypes écologiques : Avancées dans la modélisation métabolique ascendante pour les microeucaryotes phototrophes lors de l'évaluation du plancton marin.

Mots clefs : Biologie des systèmes, données omiques, modélisation métabolique, microeucaryote planctonique, ecologie marine

Résumé : Cette thèse se concentre sur la connexion des données omiques aux écosystèmes marins grâce à la modélisation métabolique. Le plancton marin, incluant les virus, les bactéries, les archées ou les eucaryotes unicellulaires, est essentiel à la régulation de la vie sur Terre. Ces organismes contribuent à des processus cruciaux tels que la production d'oxygène, la pompe à carbone, le recyclage des nutriments et servent de source alimentaire pour les niveaux trophiques supérieurs. Malgré cela, une grande partie de leur biologie reste peu étudiée. Les progrès en séquençage à haut-débit et en bioinformatique ont permis la reconstruction de génomes environnementaux fournissant des connaissances précieuses sur ces organismes non cultivables.

Les modèles métaboliques à l'échelle du génome (GSMs) permettent de prédire quantitativement les relations entre genotypes, environnements et phénotypes. Initialement utilisées pour modéliser la physiologie cellulaire et la croissance des organismes modèles en laboratoire, ces approches ont été étendues pour s'appliquer aux communautés microbiennes. De nombreux GSMs pertinents sur le plan écologique sont déjà disponibles pour les procaryotes. Cependant, en raison de la rareté d'organismes modèles avec des génomes séquencés disponibles et des étapes de curation manuelle laborieuses pour obtenir des modèles efficaces, les micro-organismes eucaryotes sont en retard. L'approche ascendante propose un changement de paradigme en introduisant un méta-modèle générique dont la curation n'est effectuée qu'une fois. Ce modèle générique est ensuite dérivé en modèles spécifiques prêts aux analyses sous-contraintes, tout en conservant les propriétés fonctionnelles et structurelles pertinentes. Jusqu'à présent, cette technique ne s'appliquait qu'aux procaryotes.

Dans ce travail, nous présentons PhotoEukStein, un méta-modèle générique permettant la reconstruction automatique de modèles métaboliques d'algues eucaryotes.

PhotoEukStein contient les informations biochimiques et génomiques de 16 eucaryotes phototrophes, utilisant l'énergie lumineuse pour convertir le dioxyde de carbone en composés organiques. Les modèles dérivés de PhotoEukStein capturent les propriétés métaboliques essentielles, et montrent une forte corrélation avec les modèles construits manuellement pour prédire les taux de croissance d'algues spécifiques. De plus, notre analyse suggère une étroite interconnexion des réactions qui est similaire aux modèles de référence.

À ce jour, 549 modèles ont été dérivés de PhotoEukStein en appliquant cette nouvelle méthode aux génomes environnementaux et aux transcriptomes de microorganismes eucaryotes unicellulaires phototrophes de l'expédition *Tara Oceans*, fournissant ainsi une nouvelle ressource précieuse. En effet, les GSMs offrent une représentation plus précise des caractéristiques fonctionnelles des organismes par rapport aux stratégies se basant seulement sur l'annotation des gènes, ou par proxy taxonomique. Nous accédons à une vision holistique essentielle pour comprendre de manière exhaustive comment de nouvelles fonctions émergent de l'interaction complexe des gènes avec leurs environnements, et contribuent aux caractéristiques phénotypiques. De plus, différentes techniques permettent l'intégration des GSMs aux écosystèmes planctoniques et aux processus biogéochimiques à l'échelle océanique, ouvrant les portes aux questions d'évolution ou de prédictions face au réchauffement climatique.

En permettant désormais l'intégration de la couche eucaryote pour la première fois, PhotoEukStein ouvre la voie à une exploration écosystémique approfondie des communautés planctoniques, des virus aux phototrophes unicellulaires. PhotoEukStein va ainsi contribuer de manière significative à notre compréhension du métabolisme, de la physiologie, de la biogéochimie et de l'écologie des eucaryotes phototrophes.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Exploring the nexus of omics, ecology, and phenotypes : highlighting phototrophic microeukaryotes through top-down metabolic modelling in marine plankton assessment.

Key words : System biology, omics data, metabolic modelling, microeukaryotes plankton, ecology

Summary : This thesis focuses on connecting omics data to marine ecosystems through metabolic modelling. Marine plankton, including viruses, bacteria, archaea, and single-celled eukaryotes, play a crucial role in Earth system functioning. However much of their biological processes remains understudied. They contribute to primary production, oxygen production, nutrient cycling, and serve as a vital food source for higher trophic levels, making their study essential for understanding and managing marine ecosystems and their ecological balance. Advances in sequencing technology and bioinformatics have enabled the *de novo* reconstruction of genomes from environmental samples, providing valuable insights for uncultured organisms.

Genome-scale metabolic models (GSMs) allow quantitative and computable genotype-environment-phenotype relationship of target organisms. Initially used for modelling cellular physiology and growth of model organisms, extensions of these constraint-based approaches are emerging for predicting and understanding microbial communities. To date, numerous ecologically relevant GSMs are already available for prokaryotes. However, models for marine eukaryotic microbes are lagging behind, mostly due to the paucity of model organisms with available sequenced genomes and to the time-consuming steps of manual curation required to obtain effective models. These curation steps are particularly tedious in traditional bottom-up approaches since they must be performed for each new model reconstruction. The top-down approach shifts this paradigm by introducing a generic meta-model for which curation is done only once. This meta-model is then converted to ready-to-use organism-specific models while preserving the whole manual curation and relevant structural properties. Until now, this technique was only applied to prokaryotes. In this work, we introduce PhotoEukStein, a generic model enabling fully-automatic reconstruction of eukaryotic-algae metabolic models at genome-scale.

PhotoEukStein was built from the merging of available biochemical and genomic information of 16 eukaryotic algae, and combines features of photosynthetic eukaryotic cells (using light energy to convert carbon dioxide into organic compounds). An extensive manual curation has been done to make it “simulation-ready”. We proved that PhotoEukStein-derived models accurately capture relevant metabolic properties and show high correlation with expert-based models in predicting growth rates of specific algae species. Additionally, the similarity in correlation maps suggests a close alignment in the interconnectedness of reactions.

To date, 549 models were derived from PhotoEukStein by applying this new method to *Tara Oceans* environmental genomes and transcriptomes of phototrophic marine unicellular eukaryotes, providing a brand new valuable ressource.

Indeed, GSMs offer a more accurate representation of the functional characteristics of organisms than strategies based only on gene annotation, or by using taxonomic proxy. They offer higher-level insight and a systems-level perspective which are essential to comprehensively unravel the complexities of gene function and their contributions to phenotypic trait. Moreover, metabolic niche allow connection of GSMs to biogeochemical processes at ocean-scale, opening the doors to evolutive questions or impacts of climate change on these precious ecosystems.

Overall, PhotoEukStein significantly advances our understanding and modelling of the metabolism, physiology, biogeochemistry, and ecology of phototrophic eukaryotes. By allowing now the integration of the eukaryotic layer for the very first time, PhotoEukStein paves the way for an in-depth ecosystemic exploration of plankton communities from viruses to single-cell phototrophs.

ACKNOWLEDGEMENTS

I am deeply grateful to Sabine Peres and Juan Nogales for accepting the task of reviewing this manuscript. Their expertise in evaluating my work is invaluable. Additionally, I extend my appreciation to Chris Bowler, Clémence Frioux, and Dominique de Vienne for their willingness to serve on the jury. I eagerly anticipate engaging in fruitful discussions with them.

Je tiens à exprimer ma profonde gratitude pour la relation que j'entretiens avec Damien Eveillard. Tu as été l'un des premiers enseignants à reconnaître et à nourrir ma sensibilité et ma passion pour la nature et le monde vivant. Ces interêts sont mon liant, aussi bien dans la science que dans la philosophie, la politique et la musique. Tes connaissances approfondies et tes idées futuristes ont été une véritable source d'inspiration pour moi, ouvrant de nouveaux horizons qui intègrent ces domaines fascinants. Les rencontres comme la nôtre sont véritablement motivantes dans la vie ! Et j'espère sincèrement que nos chemins continueront de se croiser, idéalement sur une île tropicale avec un cocktail à la main. Un immense merci.

Je suis particulièrement reconnaissante envers Antoine Régimbeau, qui a été mon collaborateur le plus proche. Sacrées parties de FrankenBrain-storming ! Un tronçon de PhotoEukStein te revient, et c'est un délice de le savoir. J'ai partagé avec toi l'essence de la Science (moi aussi je mets un « S » majuscule). Et je citerai Charles Pépin « Pour progresser, il faut rencontrer un autre que soi. C'est tout le sens de la dialectique hégélienne : une idée doit rencontrer une autre idée pour déployer toute sa puissance. Seule, sans antithèse, négation, à laquelle se confronter, une thèse ne peut vraiment s'exprimer. Il faut se confontrer à une autre conscience pour, contre sa différence, apprendre à se situer, connaître sa valeur et progresser». Tu as su m'offrir cette anti-thèse qui m'a parfois beaucoup manquée, et plus encore ! A la prochaine virée nocture niçoise (cette fois on ira au Tipsy Bar plutôt que de dormir sur la plage), et à toutes nos prochaines excursions !

J'en profite pour te remercier Benoît Delahaye. J'étudie secrètement dans mon coin pour te sortir un jour les arguments à l'encontre de ton déterminisme.

Merci au consortium *Tara Ocean*, comme un sentiment d'appartenance. La retraite reste probablement mon meilleur souvenir de thèse ! Lucia, c'est là bas que je t'ai rencontrée, c'est donc ici que je te remercie. Ho un bel ricordo delle nostre discussioni sotto le luci delle guinguettes o quando il mare lambiva i nostri piedi (I trust deepl for these few words).

Merci aux équipes de séquençage, sans lesquelles rien ne serait possible. Aux Catherine's et Nancy pour leur efficacité et leur bonne humeur. Merci à Patrick Wincker et Eric Pelletier de m'avoir permis de réaliser mon doctorat au sein du LAGE. Merci Eric pour ta confiance infrangible, et de m'avoir aidée à aboutir cette thèse. Tes compétences informatiques m'ont plusieurs fois sauvé la mise ! J'espère sincèrement que cette recherche a su nourrir la tienne et celles de l'équipe.

Je remercie Nina Guérin (la suite après !), Céline Orvain et Nadia Perchat, pour leurs efforts dévoués dans la mise en place des protocoles expérimentaux en laboratoire. Votre engagement a été inestimable. Je suis également reconnaissante envers Alain Perret et Adrien Thurotte pour nos discussions approfondies qui ont comblé le fossé entre mes analyses informatiques et les expériences *in vitro* menées par l'équipe.

Merci le LAGE ! Tout d'abord à la fratrie doctorale, merci aux aînés : Romuald et ton humour sarcastique qui décape, Julie pour ta détermination sans faille, à Paul et ton flegme qui m'effare. Merci aux cadets : Nina, comment ne pas être éblouie de ta lumière, estomaquée de ton authenticité? J'ai adoré pouvoir être proche de toi d'une certaine manière. Merci pour ce cadeau. Margaux, que ce soit ta sensibilité, tes engagements ou ton scepticisme, j'ai été transportée. Lucas, le gars est beau, méga intelligent, et mystérieux, la classe au naturel (et oui c'est sincère) ! Une pensée tendre pour Lucie, l'expatriée. Thibault, notre relation fut courte, mais suffisamment longue pour percevoir ta douceur. Paradoxalement, on aurait très bien pu se retrouver la tête dans un caisson de basse à taper du pied. Clément, je crois fort en toi, et je te souhaite de te surpasser ! Break the game ! J'ai également de belles pensées pour Chloé, Achraf et Victor, je vous souhaite le meilleur. Jana, merci d'avoir exposé ton visage d'enfant sur le dancefloor ! Merci Quentin (le phare même quand le LAGE est vide), Betina (pour ton coeur sur la main), Olivier (une p'tite meuf et un atout), Morgan (ton génie et ta bienveillance), Tom (et ton herpès !), merci Julie pour ton soutien et ta soif d'aventures. Force à vous pour la suite !

Merci au GenoPub ! Je me dois de te remercier Mathieu pour ta belle présence et ton humour absurde qui a su soulever la soupape de ma cocotte-minute du stress. Au GenoJazz ! À Amine et tes rythmes de guitare effrénés, à David pour cette fondue inoubliable et tes conseils précieux et bienveillants, merci Laurine, Deborah, Marion, Océane, Humbeline, Karine et Carine, Laurie, Sophie, Jean-Marc, Pedro, et toutes les personnes qui ont embelli l'aventure au sein du Genoscope. Je pense notamment à Marta, de Nyege Nyege à Bréhat, merci pour tout ce que l'on a partagé.

Merci à Valérie Chaudru et Elisabeth Petit pour ces premières expériences d'enseignements universitaires. Merci à Valérie Fortuna et Elisabeth Delbecq pour ces merveilleuses années de médiation scientifique. Ça a été particulièrement riche et captivant. Merci à mes co-équipiers préférés Mohamed, Roxane, Lydia, Heloise, Julien, Mendes, Camille et Lise. Valentin, nous avons exploré bien au-delà de la MISS et je te remercie pour ces délicieuses et extravagantes aventures (qui continueront encore je l'espère).

Merci à mes colocs Lou, Clément et Samos pour ces délires à pas d'heure, ces repas de fête, les parties de Mario Kart à n'en plus finir, la clarinette basse à en faire vibrer les murs, le piano faux mais enveloppant, le melodica à t'en péter les oreilles, la guinguette du bonheur, la tisane réconfortante, les histoires sombres et dénuées de sens de la Lisa's coloc (je m'en serais bien passée, vivement qu'on retrouve cette jupe !). Cette dernière année à vos côtés a embelli le quotidien !

Merci à tous mes potes ! Je ne peux décrire ici tous les merveilleux souvenirs ensemble, les étapes franchies, l'amour et l'humour si essentiels, votre soutien : Françoise, Moctar, Marine, Camille, Laura, Maïlys, Léa, Nadifi, Laurie, Justine, Claudia, Clémence, Isabella, Maggie, Marianna, et à toutes les personnes précieuses pour moi. Merci Ilaria, Francesco, Jade, Seb, Elena, Teresa et Léon : à vos arts et vos sciences ! Louisa et Claire, rendez-vous dans le Lot pour vivre notre vie de rêve.

Merci à ma famille pour votre soutien inconditionnel et vos richesses humaines.

Enfin, mes derniers mots sont pour toi Sarah. Toi et moi, on forme la plus belle équipe à mes yeux. On s'était promis il y a 8 ans qu'on la ferait cette thèse ! Merci d'avoir écouté et soutenu l'éventail immense de mes émotions et reflexions liées à cette aventure. Merci d'avoir partagé les tiennes. Merci infiniment pour bien plus que cela… Longue vie aux Dr Boub !

« A quoi bon en effet toute cette prospérité, si l'on ne peut y joindre la bienfaisance, qui s'exerce surtout et d'une si louable manière à l'égard de ceux qu'on aime » (Ethique à Nicomaque, Aristote).

Table of Contents

PREAMBULE

The focus of my research work was developing PhotoEukStein, a versatile meta-model that enables fully-automatic reconstruction of constraint-based metabolic models (CBMs) for eukaryotic microalgae at genome-scale. I hold a master's degree in Health Biology with a specialisation in Genetics, Genomics, and Systems Biology. Therefore, my work is mainly influenced by my background in biology, but I also possess interdisciplinary skills in computational biology. This allows me to bridge the gap between experimental biology and mathematical modelling, which is an aspect I particularly enjoy in my scientific pursuits.

Chapter 1 of this thesis serves as an introduction, providing essential concepts for comprehending this research. The initial section highlights the significance of characterising microbial planktonic communities, specifically their role in Earth system regulation. The subsequent part elucidates the available data and mathematical modelling used today to describe planktonic populations and their functions. To bridge the gap between environmental data and existing models that lack detailed descriptions of metabolic processes, we suggest to use Genome-Scale CBMs (GSMs). Although metabolic modelling has already made major advances in ecology, little has been done for eukaryotic microbes. The final section outlines constraint-based metabolic modelling and its current available model reconstruction methods.

Chapter 2 provides a comprehensive overview of the steps involved in reconstructing PhotoEukStein, a generic model enabling fully-automatic reconstruction of GSMs for eukaryotic microalgae. Beginning with meticulous manual curation and followed by validation of model predictions, the meta-model reconstruction process involves carefully refined and optimized aspects. This includes collecting and integrating genomic and biochemical information from available sources, and ensuring that the model accurately represents the metabolic characteristics of eukaryotic microalgae. This process of model reconstruction and validation not only encompasses technical aspects but also raises philosophical and epistemological considerations. The choices made during the curation process and the validation strategies employed reflect the underlying assumptions, limitations, and uncertainties inherent in the modelling approach (further discussed in chapter 4).

Chapter 3 is represented by my thesis paper and briefly restates the concepts of Chapter 1 and Chapter 2, but more importantly, it presents the brand new resource of 549 GSMs for microeukaryotes phototrophs. This paper emphasises the importance of taking a holistic approach when studying biological systems. Currently, the characterisation of planktonic functions is often limited to 1) gene annotation, 2) statistical correlations, or 3) taxonomic proxies. However, these approaches have their limitations and do not provide a comprehensive understanding of the complex interactions and emergent functions within these biological systems. 1) The reliance on gene annotation alone is reducing, promoting a view of genetic determinism and overlooking the intricate network of interactions that contribute to functional outcomes. 2) Statistical approaches that correlate gene or organism abundance with environmental parameters provide valuable insights but do not establish causal links and do not address the question of "who does what and how". 3) Modelling Planktonic Functional Traits (PFT) at ocean scale and considering temporal dynamics is very powerful. However, these models often oversimplify biological processes and associate function with taxa proxies, disregarding the intra-individual variability and the complexity of physiolocial processes.

By moving beyond simplistic associations and towards a more comprehensive and integrative approach that considers the holistic nature of biological systems, we can gain a more accurate representation of the complexity inherent in planktonic systems. By acknowledging the limitations of current methods and exploring new avenues for studying functional traits, we can uncover the intricate mechanisms that drive the emergence of functions.

Chapter 4 consists of two distinct parts, each addressing important aspects of the research. The first part delves into the inherent complexity of biological systems, addressing the sophisticated relationship between genotype and phenotype. It highlights the limitations of integrating all the parameters in a single type of model. This recognition of complexity emphasises the need for alternative modelling approaches that capture imperfections and uncertainties, which in turn can lead to the generation of new hypotheses and insights. By discussing these philosophical and epistemological concepts, the chapter fosters a deeper understanding of the underlying motivations and justifications for the chosen methodologies. It emphasises the significance of critical thinking and interpretation in scientific research, encouraging researchers to acknowledge the limitations and assumptions embedded within models.

In the second part of the chapter, the emphasis is placed on the potential of metabolic modelling in elucidating the characteristics and functions of planktonic organisms. Preliminary results are presented to demonstrate the value of integrating these modelling approaches with experimental manipulations, showcasing the synergistic effects that arise from their combination. Furthermore, the discussion extends to potential future research directions and ideas, highlighting the avenues for further exploration and investigation in this field.

I am aware that biology uses a wide range of specific vocabulary, so I provide a glossary with different definitions at the end of the document to facilitate the reading.

The various resources and scripts discussed in this work can be accessed and downloaded from the following link: https://www.genoscope.cns.fr/PhotoEukStein/

1 INTRODUCTION

« Il est difficile de ne pas être frappé par la symétrie renversée entre les gestes de Galilée et de Lovelock levant de modestes instruments vers le ciel pour y faire des découvertes radicalement opposées. (…) Tandis que Galilée, levant les yeux de l'horizon vers le ciel, renforçait la similitude entre la Terre et tous les autres corps en chute libre, Lovelock, baissant les yeux à partir de Mars dans notre direction, diminue en fait la similitude entre toutes les planètes et cette Terre si particulière qui est la nôtre.», Bruno Latour[1]

1.1 BIOLOGICAL CONTEXT

1.1.1 The building blocks of life flow through organisms

Matter can exist in various forms, including molecules, atoms, and subatomic particles such as protons, neutrons, and electrons. Molecules play a significant role in the structure and organisation of matter that we observe in the world around us. They are formed when two or more atoms bond together (Figure 1). Atoms function as the fundamental building blocks of all molecules and constitute the smallest unit of matter that retains the distinctive chemical properties of an element. Unique feature so far in the universe, life on Earth plays an important role in what we call the biogeochemical cycles. These cycles are crucial for maintaining life on our planet, as they regulate the avaibility of these essential elements.

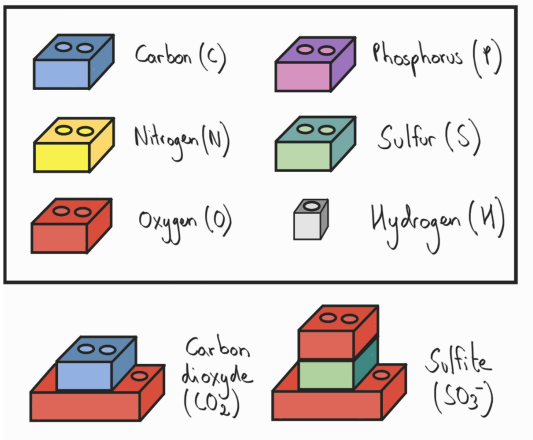


Figure 1: Atoms are the fundamental building blocks of all molecules. The six major atoms[2] are represented. Molecules are formed when two or more atoms bond together. Organisms need a constant supply of specific molecules to survive.

Molecules are always transformed, transported, and recycled. Indeed, in order for organisms to prosper, they necessitate a constant supply of specific molecules, such as nutrients and other essential elements as ions. Thus, they undergo specific metabolic processes to transform into vital molecules, including DNA, proteins, lipids, and carbohydrates[2]. Therefore, all the chemical elements in an organism are part of biogeochemical cycles (Figure 2). The chemicals flow through an organism. For example, carbon moves between the atmosphere and the biosphere through photosynthesis and cellular respiration. It moves from the biosphere to the lithosphere through decaying organisms and animal waste products, from the atmosphere to the hydrosphere through dissolution of organic and inorganic carbon, from the lithosphere to the hydrosphere through erosion and so on. Human activities such as deforestation, industrialisation, and agricultural practices have disrupted these cycles[241] leading to ecological imbalances and environmental problems such as climate change, eutrophication, and acid rain.

1.1.2 Earth system as a supraorganism ?

The concept of the biosphere, theorised by Vladimir Vernadski in 1926, is certainly one of the major points retained by James Lovelock and Lynn Margulis when they described the Gaia theory, in the 1970s. This theory suggests that Earth is a self-regulating system that maintains the conditions necessary for life to thrive[3] (example 2.2.1.1). It implies that the Earth's atmosphere, oceans, and land surface are all part of a complex feedback system that maintains the planet's environmental conditions within a narrow range that is optimal for life. They suggested that the biota and their environment are so tightly related that they function together as a single system, which Lovelock and Margulis called Gaia (in reference to the ancestral mother of all life for the ancient Greeks). Although the belief of some optimisation can be debatable* (especially if humans are included†), the fact that living organisms can change the environment in potentially drastic ways (1.1.3.2) is accepted as one of the foundations of current ecological dogma.

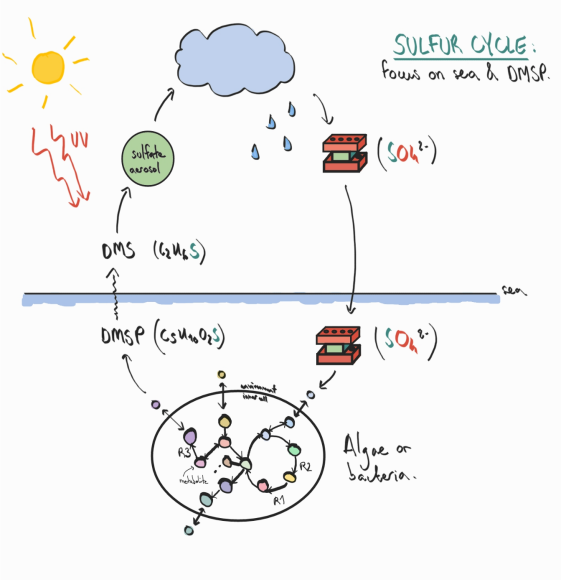


Figure 2: Molecules are transformed, transported, and recycled through biogeochemical cycles. Metabolism of organisms is part of these cycles : chemicals flow through an organism. Here a focus on sulphur cycle and dimethylsulfoniopropionate (DMSP) production. When dimethyl sulfide (DMS) is released into the atmosphere, it can act as a cloud condensation nucleus, which means it can attract water vapor to form tiny droplets that eventually form clouds. DMSP have many more functions in the ocean (see 1.1.4.2).

Understanding the complexity of the Earth system (Gaïa) and the importance of ecosystems is critical for estimating the unique conditions that have allowed life to develop and continue to thrive on our planet, and predict the impact of anthropological activities on ecosystems health. The Earth system is complex and made up of many interconnected subsystems, including the atmosphere, hydrosphere, lithosphere, and biosphere. The biosphere is composed of all living organisms and ecosystems, which play a critical role in maintaining a stable and habitable environment for life on Earth. Ecosystems are composed of both biotic and abiotic factors, including plants, animals, microorganisms, air, water, and soil. They are characterised by the complex relationships between these factors, including energy and nutrient flow, as well as competition, and predation of living organisms. Thus, they are not independent and exist in a complex web of ecological relationships, making up the biosphere.

---

* *The concept of maintaining environmental conditions for life is crucial for the sustainability and thriving of a diverse range of organisms on the planet. Optimal environmental conditions refer here to the range of conditions that allow for the greatest diversity and productivity of life. This includes factors such as temperature, water availability, air quality, and nutrient availability. By maintaining these optimal conditions, ecosystems can support a greater variety of species and promote their growth and reproduction, ultimately contributing to the overall health and functioning of the ecosystem. Are optimal conditions narrow and specific, emphasising the importance of strict requirements for certain organisms to thrive, or do they have a broader range of tolerance, allowing for greater adaptability and resilience in diverse environments ? May optimal conditions refer to maintaining a stable and unchanging environment, or conversely, they involve embracing change and variability ? Conditions necessary for the survival and growth of individual species, prioritising their specific needs, or for the overall health and functioning of the entire ecosystem? What do we mean by productivity in this context? In my opinion, concept of "optimal conditions" is complex and multifaceted, and debates can stem from different philosophical, scientific, or ethical perspectives.*

† *The relations of humans and their social organisation to the natural environment have long been studied within the framework of an opposition between nature and society[4].*

Ecology, the study of the interactions between organisms and their environment, seeks to understand the way life functions on Earth and how it is organised. Ernst Haeckel (1834-1919) recognised that living organisms could be studied at different levels of complexity, from subcellular components to ecosystems (Figure 3), and that the environment plays a critical role in shaping the development and evolution of organisms. This approach laid the groundwork for modern systems biology (1.2.2.2), which seeks to understand the complex interactions between different levels of biological organisation.

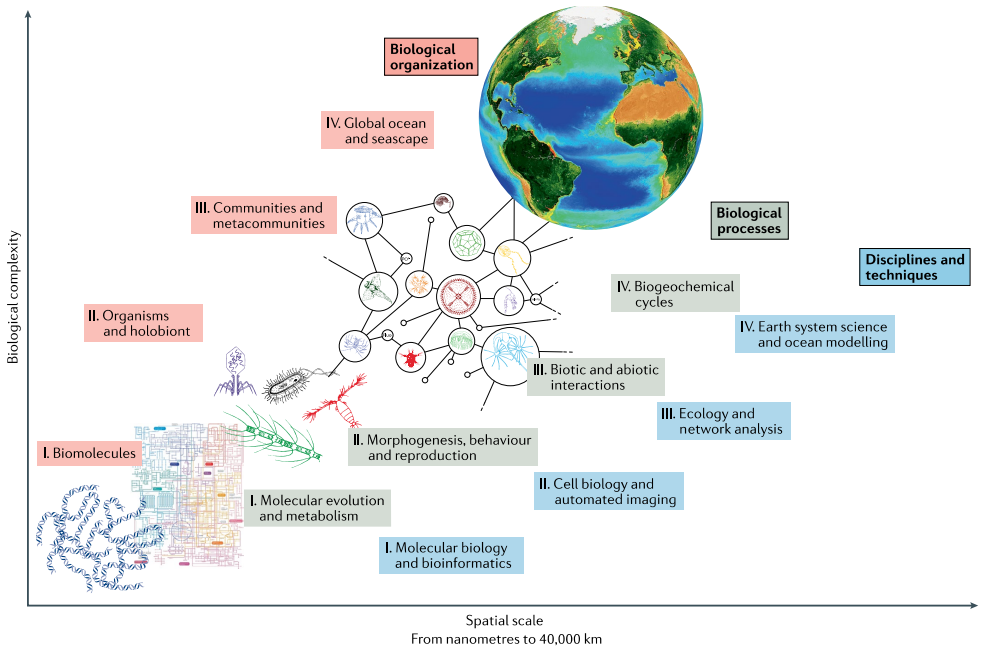


Figure 3: Ecosystems biology and integrative analyses of the global ocean. Biological functionality is multilevel. Living organisms can be studied at different levels of complexity, from subcellular components to ecosystems. Figure from[37].

1.1.3 Dynamic and essential process of photosynthesis for sustaining life

1.1.3.1 Solar energy is the primary source of energy

While materials are cycled through ecosystems, the Earth system is an open system in terms of energy. Solar energy enters the Earth system and fuels the processes that sustain life on Earth (Figure 4)**.** Ecosystems deal with energy and nutrient flow among the living organisms and their physical environment. Photosynthesis serves as the primary source of energy and nutrients for the majority of life forms on Earth (Figure 5). Indeed, phototrophic organisms, such as plants and algae, are -what we call autotrophic, meaning they are able to create their own organic matter from inorganic carbon sources such as carbon dioxyde CO_2 and bicarbonate HCO_3^-. Phototrophs use specifically solar energy through photosynthesis to achieve this process.

Global biochemical equation of photosynthesis

$$6\ CO_2 + 12\ H_2O + \text{light} \rightarrow C_6H_{12}O_6 + 6\ O_2 + 6\ H_20$$

CO_2 : carbon dioxide as inorganic carbon ; H_2O : water ;
$C_6H_{12}O_6$: glucose as organic matter ; O_2 : dioxygen.

Indeed heterotrophs, such as animals, decomposers, or heterotrophic bacteria, must obtain organic compounds from external sources whose autotrophs are the root (Figure 4; Figure 5). Organic matter provides the basic building blocks of life like carbon, which is the backbone of all living organisms, and other essential elements. They are used to form the complex molecules that make up living organisms, such as carbohydrates, lipids, proteins, and nucleic acids. These compounds are recycled as organisms die and decompose. Nutrient cycling is facilitated by decomposers, such as bacteria and fungi, which break down organic matter to extract energy and release nutrients back into the environment for reuse by other organisms. Decomposition of organic matter by microbes and other organisms also helps to recycle nutrients and maintain the balance of ecosystems.

Hence, the rate of photosynthesis places an upper bound on the overall biomass and productivity of ecosystems, and constrains the overall biological flow of energy on the surface of this planet (Figure 4; Figure 5). Without sunlight, photosynthetic organisms would not be able to produce organic food and energy, and the rest of the food chain would collapse. In my knowledge, only chemosynthetic and lithotrophs organisms (other types of autotrophs) do not rely on photosynthesis for their survival. The first category of organisms often are archaea or bacteria and survive in extreme environments, such as deep sea hydrothermal vents, where there is no sunlight (Table 1). They derive their energy from the oxidation of inorganic compounds such as hydrogen sulfide H_2S. The second category include some bacteria and archaea found in rocks and soil using inorganic minerals, such as sulphur.

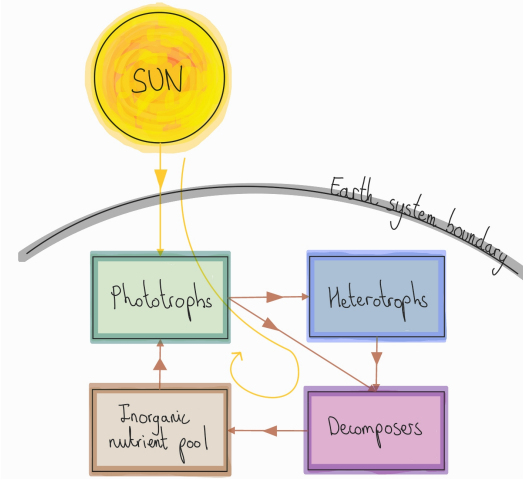


Figure 4: Simplest representation of microorganism's ecosystems. Solar energy (yellow arrow) enters the Earth system and fuels the processes that sustain life on Earth. Materials (brown arrow) are cycles through ecosystems.

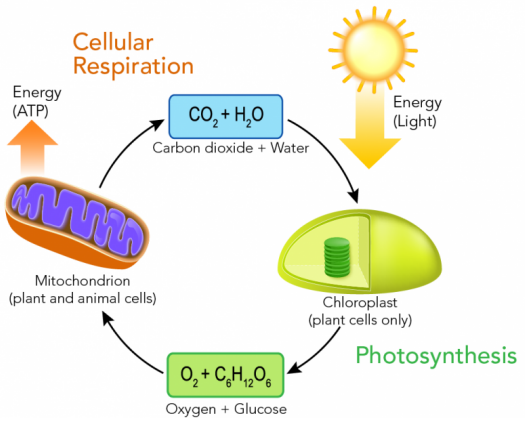


Figure 5: Chloroplasts, mitochondria and the energy cycle (from istockphoto). Photosynthesis uses light to transform inorganic carbon and water into organic carbon and oxygen (in chloroplasts of phototrophic organisms). Respiration uses organic carbon and oxygen to fuel cellular processes involving a series of chemical reactions that break down the glucose molecule, release energy, and produce waste products such as carbon dioxide and water (in mitochondria of both phototrophic and heterotrophic organisms).

1.1.3.2 Oxygen is crucial for sustaining life on Earth

For the first two billion years of the Earth's history, there was very little dioxygen O_2 in the atmosphere. Anoxygenic photosynthesis was the dominant form of photosynthesis that does not produce O_2 as a by-product. Between 2.4 – 2.1 billion years ago, in the oceans of the Proterozoic era, oxygenic photosynthesis appeared as a major evolutionary innovation, and became the source of dioxygen present in the Earth's atmosphere[242]. Over geological timescales, the drawdown of CO_2 was

not stoichiometrically proportional to the accumulation of O_2 because photosynthesis and respiration are but two of the many biological and chemical processes that affect the atmospheric concentration of these two gases. However, aquatic photosynthetic organisms, such as cyanobacteria, permanently altered Earth's atmosphere[5,6] (Table 1), a phenomenon that ultimately permitted multicellular animals, including humans, to evolve[7,12].

Indeed, one of the major benefits of an oxygen-rich atmosphere is that it allowed for the evolution of aerobic respiration, a highly efficient process for extracting energy from organic molecules (Figure 5). Through cellular respiration, organic matter is broken down into smaller molecules, such as glucose, which are then used to produce adenosine triphosphate (ATP), the primary energy currency of living cells. ATP is used for a wide range of cellular processes, including movement, growth, and reproduction. In aerobic respiration, dioxygen is used as an electron acceptor in the electron transport chain. It is much more efficient than anaerobic respiration, which uses other electron acceptors such as sulphur compounds[2]. This efficiency allowed organisms to extract more energy from their environment, leading to the evolution of more complex and energetically demanding life forms.

Phototrophs, among others, contribute to the carbon biogeochemical cycle by fixing carbon dioxide through photosynthesis and releasing oxygen into the atmosphere. With three quarters of the Earth's surface covered by water (making the ocean the largest continuous environment and home to extraordinary biodiversity)[243], oceanic photosynthesis plays as important a role in carbon capture and storage as terrestrial photosynthesis. Indeed, phytoplankton in the surface ocean accomplishes approximately 50% of the Earth's annual primary production[244]. Moreover, oceanic photosynthesis has the potential to capture and store carbon for longer periods of time than terrestrial photosynthesis. Indeed, when marine organisms die and sink to the ocean floor, they bring carbon and other nutrients with them, contributing to the deep-sea carbon cycle (Figure 7). A small fraction of the fossilised organic remains of aquatic photosynthetic organisms would become petroleum and natural gas that simultaneously fuels contemporary civilisation (Table 1).

1.1.4 Involment of marine plankton in biogeochemical cycles

1.1.4.1 The tale of marine plankton

The tale of plankton begins in the vast and mysterious oceans, where tiny organisms drift and dance in the currents. Plankton, which comes from the Greek word "planktos" meaning "wandering" or "drifting," includes a diverse group of organisms that are either too small or too weak to swim against the ocean currents. They include both unicellular and multicellular organisms, such as virus, bacteria, archaea, algae, protists, and some animals, including larval forms of various marine invertebrates and fish (Figure 6). Approximately 70 % of the biomass in marine ecosystems is microbial[244]. It is known that even a single drop of seawater can contain millions of microscopic planktonic organisms (Table 1). Therefore, a teaspoon of seawater likely contains billions or trillions of plankton (depending on the location, time of day, and other environmental factors). From an evolutionary perspective, the plankton forms what we call a polyphyletic group. Indeed, the plankton is defined by physical constraints that affects all of them, rather than by an evolutionary relationship. It thus encompasses a large number of species with extremely varied characteristics, such as size, physiology, ecological niche, form, and their position in the tree of life (Figure 6). Plankton are the

dominant life forms in the ocean and comprise highly dynamic and interacting populations[60].

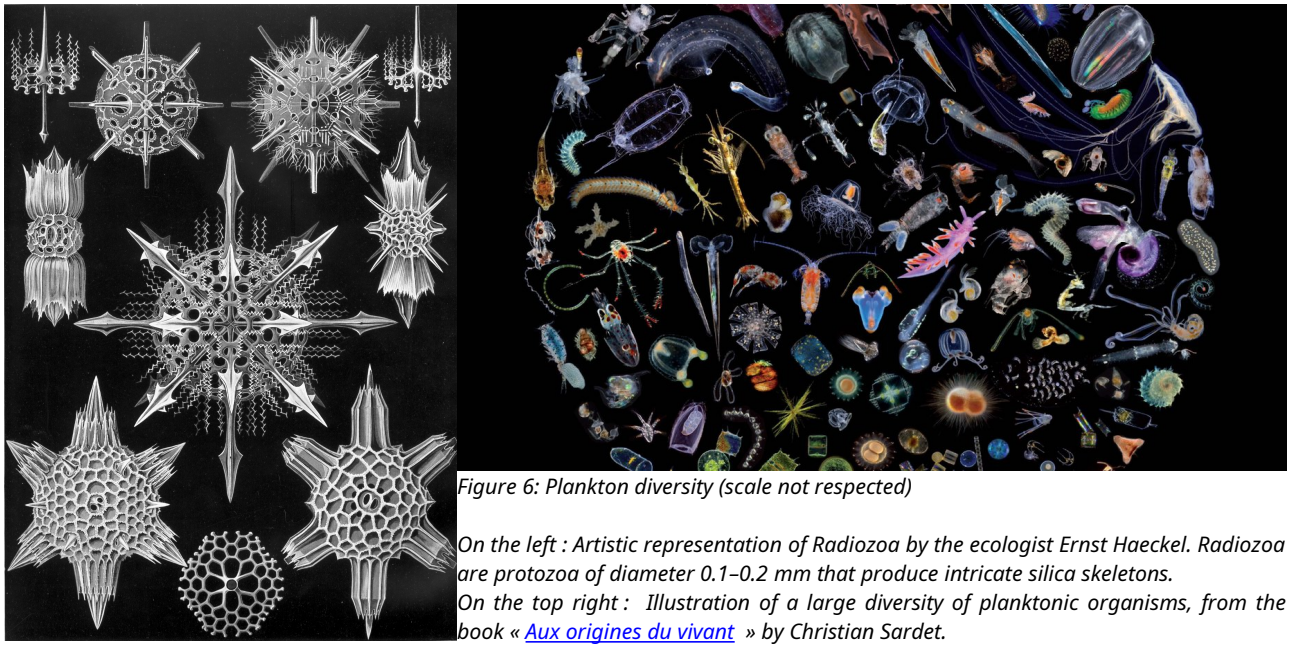


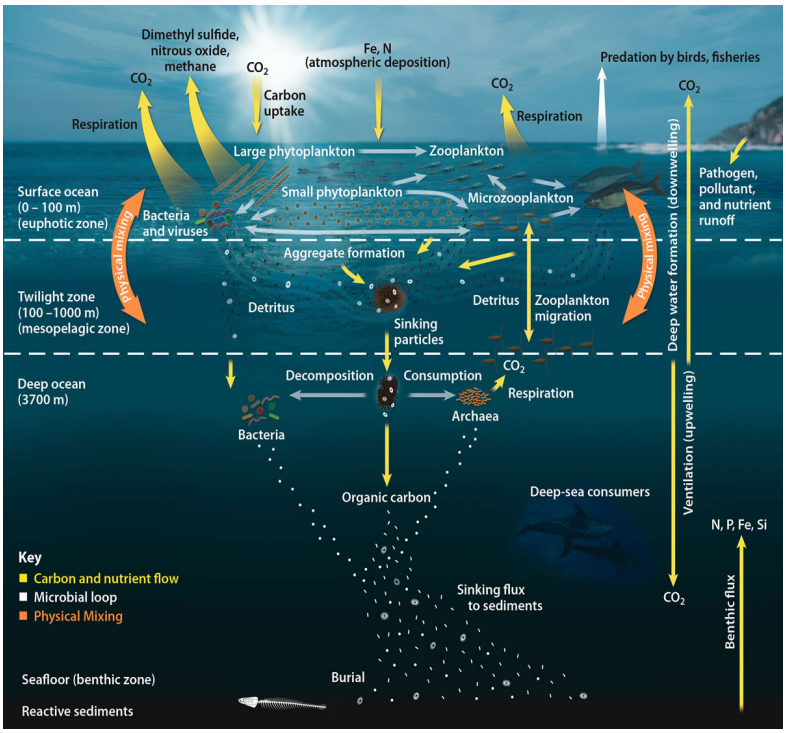
Figure 6: Plankton diversity (scale not respected)

On the left : Artistic representation of Radiozoa by the ecologist Ernst Haeckel. Radiozoa are protozoa of diameter 0.1–0.2 mm that produce intricate silica skeletons.
On the top right : Illustration of a large diversity of planktonic organisms, from the book « Aux origines du vivant » by Christian Sardet.

Planktonic organisms interact with each other and with their environment in complex ways, including commensialism, mutualism, parasitism, competition or predation relationships. They play a crucial role in marine food chains as they serve as the primary food source for many aquatic organisms[60]. In order to survive and reproduce, phytoplankton (which are the phototrophic plankton) compete for light and limited nutrients (such as nitrogen, iron, or vitamins)[8]. Zooplankton (heterotrophic plankton) feed on phytoplankton, smaller zooplankton, and detritus. They are in turn preyed upon by larger organisms, such as fish, whales, and other marine mammals (Figure 7). Moreover, viruses regulate populations of microorganisms, and play a, essential role in releasing organic matter in the environment, and transferring genetic material between species[245]. These interactions create complex food webs and nutrient cycling pathways that can vary depending on environmental conditions. For example, oceanic currents shape planktonic ecosystems by transporting planktonic organisms over long distances, influencing their distribution patterns and creating opportunities for dispersal and colonisation of new areas. These currents affect nutrient availability and can create areas of upwelling, where deep nutrient-rich waters rise to the surface, leading to increased primary production and planktonic biomass. Thus, understanding and predicting planktonic communities can be challenging, particularly in the vast and dynamic ocean environment.

Scientists have come to appreciate the importance of plankton in mediating major biogeochemical cycles of the Earth. Research in the late 1980s by geochemists and biologists contributed to a better understanding of their role in maintaining the balance of the Earth's systems[9,10] (Table 1). These organisms not only help to maintain the steady-state gas composition of the atmosphere but also respond to climate feedbacks, contribute to the regulation of the Earth's climate and weather patterns[11,12]. They interact with the atmosphere (CO_2, dimethyl sulfide...) and are then connected to all the different subsystems of the complex Earth system (Figure 7). The ecological importance of plankton in our ecosystems, particularly marine ecosystems, is a fundamental question to be addressed. They are vulnerable to environmental stressors such as pollution, ocean acidification, and climate change, which can have far-reaching effects on the health of the entire ocean ecosystem.

Figure 7: The pelagic food web.



Planktonic organisms interact with each other and with their environment in complex ways. The microbial loop starts with the production of organic matter, primarily through the growth and photosynthesis of phytoplankton. The zooplankton graze on the phytoplankton, breaking them down into smaller particles through feeding and excretion. They are in turn consumed by larger organisms such as fish, whales, and other marine mammals. The smaller particles, including organic detritus, dead cells, and fecal pellets from the zooplankton, become available as a food source for a diverse community of bacteria, archaea, and protists. Some of the microorganisms break down the organic matter and recycling nutrients back into the system. Others are heterotrophic bacteria that consume the dissolved organic matter produced during the decomposition process. The microbial loop is an essential component of marine ecosystems as it helps regulate nutrient cycling, energy flow, and carbon sequestration. It also influences the productivity and biodiversity of the ocean by supporting the growth of primary producers and providing a vital food source for higher organisms.

When marine organisms die and sink to the ocean floor, they bring carbon and other nutrients with them, contributing to the deep-sea carbon cycle. Plankton mediate major biogeochemical cycles of the Earth. They interact with the atmosphere (CO_2, dimethyl sulfide...) and are then connected to the different sub-systems of the complex Earth system.

The depths sampled by Tara Consortium (1.2.1) are mainly the subsurface (5-10 m), the Deep Chlorophyll Maximum layer (20-100 m) where the concentration of chlorophyll is maximum, and the mesopelagic zones (300-1000 m) where light is almost absent and often constituting Oxygen Minimum Zones (OMZ). (figure from public domain)

Today, planktology is an interdisciplinary field that includes (computational) biologists, oceanographers, ecologists, mathematicians, physicists, and climatologists among others. With the advent of new technologies like high-throughput DNA sequencing (see 1.2.2) and satellite imagery, researchers are able to study plankton at a level of detail never before possible, leading to new insights into the role of plankton in marine ecosystems and the global climate.

1.1.4.2 DMSP as biological example during this thesis

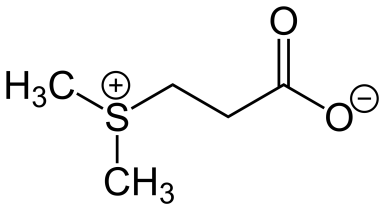


Figure 8: Topological formula of DMSP

The tertiary sulfonium compound dimethylsulfonio-propionate (DMSP) (Figure 8) has attracted particular interest as the biogenic precursor of the predominant sulphur gas, dimethylsulfide (DMS). When DMS is released into the atmosphere (Figure 2 ; Figure 7), it can act as a cloud condensation nucleus, which means it can attract water vapor to form tiny droplets that eventually form clouds. The cloud cover is important for regulating the Earth's climate and weather patterns because it affects the amount of solar radiation that is absorbed by the Earth's surface and atmosphere.

DMSP production has been observed in various planktonic organisms, including algae, bacteria, heterotrophic dinoflagellates, but also plants, and animals such as corals, among others[13–22]. Thus,

DMSP is present in all marine ecosystems and can be used for various purposes, and not only by the producing organisms, but also by other species that live in the same habitat as the DMSP producers.

(1) For example, the bacteria *Pelagibacter* lack the genes for sulphate reduction and the assimilatory sulphate reduction pathway, and has been found to use other sulphur-containing compounds, such as DMSP, methanesulfonate, and sulfonates as alternative sulphur sources[24].

(2) The chemoattraction effect of DMSP can have cascading effects on the food chain, as it can lead to the aggregation and concentration of small organisms such as krill and copepods, which in turn are important prey for larger marine animals such as fish, birds, and whales[25–27].

(3) The physiological function of DMSP as an organic osmolyte that is synthesised and accumulated under increasing salinities and under hydrostatic pressure is well proven[28]

(4) The significantly higher DMSP concentrations in many algae grown under low temperatures, compared with those maintained in temperate conditions, suggest another biological function of this compound as cryoprotectant[29]. It has been speculated that the ability to accumulate DMSP was evolved during the last ice age, when the temperature were lower and the salinity of the ocean was higher**.**

(5) DMSP can also act as an antioxidant in marine organisms, helping to protect cells from oxidative damage caused by environmental stressors such as ultraviolet radiation[30,31]. If ultraviolet radiation stress increases the production of DMSP, and DMSP is the biogenic precursor of DMS which act as a cloud condensation nucleus, then its production may protect from ultraviolet radiation. This is an example of feedback loop as described in section 2.2.1.1.

(6) Malleicyprols are known to have antibacterial and antifungal properties, and the discovery of DMSP as a precursor for their production suggests that DMSP may have a role in the chemical defense mechanisms of some bacteria[32].

(7) Finally, to finish off this Prevert-style inventory[33], DMSP may also serve as a sink for excess sulphur in response to nitrogen stress. When sulphur is in excess, microorganisms may incorporate it into DMSP, excrete it, and thus reduce the intracellular sulphur burden. This process may help maintain a balance between nitrogen, carbon, and sulphur within marine organisms and ecosystems[246]. The exact mechanisms of DMSP production and secretion in response to nitrogen stress are still not fully understood and are an active area of research. Although this thesis does not provide any major biological advances on DMSP, promising preliminary results will be presented in sections 4.2.2.1 and 4.2.3.

1.2 How to capture the complex dynamics of marine micro-organism communities ?

« The lack of numeric data describing physiology is only one of the problems for the next generation of plankton models. Inadequacies and dysfunctionality in models are not compensated for by the collection and use of data describing only part of the story. The devil is indeed in the details; nutrient- phytoplankton– zooplankton models get away with an awful lot by not exploring the details. If we are going to open Pandora's box to explore the details, then we had better be ready to handle the demons that escape from it. », Flynn[34]

1.2.1 *Tara,* the short story of the consortium

At the initiative of cell biologist Eric Karsenti, the *Tara Ocean Foundation* aims to fill some of these gaps in our understanding of complex planktonic ecosystems, among other important questions relating to ice formation in the central Arctic, microplastics and corals.

The *Tara Oceans Expedition,* in particular, was the first large-scale international scientific project which was launched by the *Tara Consortium* (Figure 9 ; Figure 10). The expedition's objective is to study the planktonic ecosystem and its diversity in the most exhaustive way possible[36,37]. It was a three-year circumnavigation of the globe that began in 2009 and ended in 2013. During the expedition, the *Tara* research vessel traveled over 140,000 km and collected more than 35,000 samples of planktonic organisms from the world's oceans.

Figure 9: Tara is the name of the schooner used for expeditions . It was initially named Antarctica by the explorer Jean-Louis Etienne in 1989, and then Seamaster by the explorer Peter Blake, who was engaged in environmental conservation. Attacked by pirates, Blake died on board in 2001 during a mission on the Amazon River35. Picture from Fondation Tara Ocean.

The *Consortium* gathers over 200 scientists from 22 countries working together through international collaborations and interdisciplinary approaches in advancing ocean-related scientific research. This research contributes significantly to our understanding of the ecological and evolutionary processes that shape marine ecosystems, as well as to investigate the potential applications of marine microbes in biotechnology, medicine, and other fields.

The *Tara Ocean Expedition* adoptes an unprecedented strategy by sampling all the micro-organisms from 0 to 2 mm encompassing the main families of plankton. The depths sampled are mainly the subsurface (5-10 m), the Deep Chlorophyll Maximum layer (20-150 m) where the concentration of chlorophyll is maximum, and the mesopelagic zones (300-1000 m) where light is almost absent[38] (Figure 7). After the water samples were collected, they were passed through increasingly fine sieves to separate the organisms of different size fractions. Sizes of plankton range from femtoplankton (<0.2 µm), mainly consisting of viruses, to megaplankton (20-200 cm), where jellyfish and salp colonies are predominantly found. In intermediate size classes, bacteria (picoplankton, 0.2-2 µm), which are mainly heterotrophic and often parasitic, but also include phototrophic cyanobacteria. As for protists

or eukaryotic microorganisms, they are primarily composed of photosynthetic algae but also many parasites. The smallest protists, such as some flagellates and ciliates, can have cell sizes in the range of a few micrometers or even smaller, making them invisible to the naked eye. These microscopic protists are typically unicellular, meaning they consiste of a single cell that performes all the functions necessary for life, including reproduction, metabolism, and locomotion. On the other hand, some protists can be much larger and more complex. For example, certain types of algae can form large multicellular colonies or even macroscopic structures like seaweed, which can be several meters in length. Finally, zooplankton, consisting mostly of small heterotrophic metazoans such as copepods, were mainly found in the size fractions of microplankton and mesoplankton.

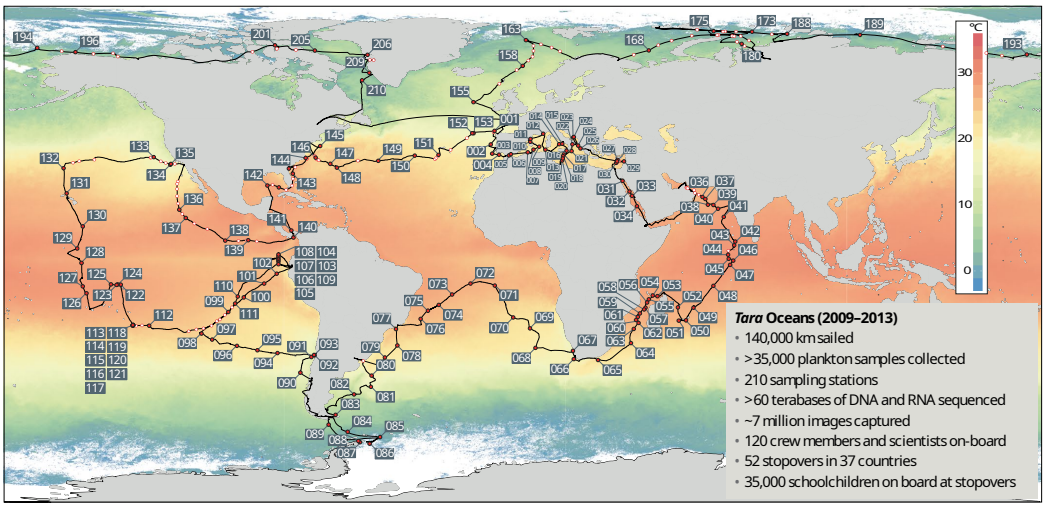


Figure 10: Sampling route of the Tara Ocean Expedition (red track) (2009-2013). After a few months of hiatus, the expeditions of the schooner continued with notably the missions Arctic polar circle (2013), Tara Mediterranean (2014), Tara Pacific (2016-2018), Tara Microplastics (2014-2019) and Tara Microbiome (2020-2022) with various scientific objectives. Figure from[37].

Some of the samples are used for -omics analyses in order to study genes, species or metabolic functions of marine microplankton communities (scope of this thesis). In parallel to the plankton sampling, a number of (bio)chemical data are measured such as pigment, nutrients, dissolved and particulate organic and inorganic matter, phosphate, nitrate or dissolved silica concentrations at the precise locations (depth, latitude, longitude) of the samples.

1.2.2 What are -omics data?

Molecular biology seeks to understand the molecular mechanisms that carry genetic information and allow the functioning of cells, organs, organisms and ecosystems. Omics data allow the holistic study of the molecules involved in these processes. What molecules are we talking about?

1.2.2.1 Molecular biology seeks to study molecules carrying genetic information

Deoxyribonucleic acid

In the second half of 20th, molecular biology developed strongly, notably with the discovery of the structure of deoxyribonucleic acid (DNA), the main carrier of genetic information, by Rosalind Franklin, Maurice Wilkins, James Watson, and Francis Crick in 1953[39]. DNA consists of a series of nitrogenous nucleotide bases, namely adenine (A), thymine (T), guanine (G) and cytosine (C), structured in a double helix[40] (Figure 11). The specific order of nucleotide bases determines the codon sequence, which in turn determines the amino acid sequence in a specific protein.

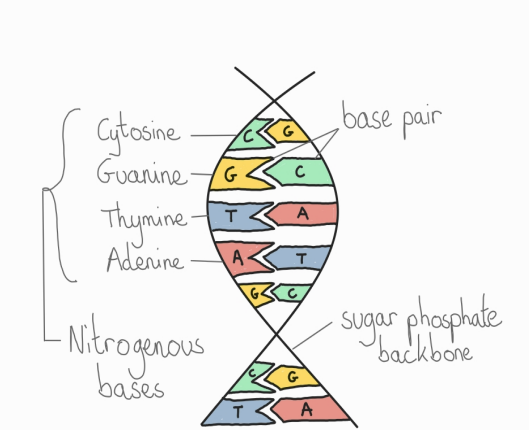


Figure 11: DNA structure.

Proteins are the main functional players in cells, and play an essential role in many biological processes. Consequently, the precise order of nucleotides is fundamental in determining the structure and function of proteins, which in turn influence the development, growth and functioning of living organisms. The synthesis of proteins from DNA is a dogma of molecular biology called "gene expression". This involves two processes referred to as transcription and translation (Figure 13). The transmission of DNA from one generation to the next allows the conservation of these instructions (is discussed in this section 4.1.2.2).

Overall, a genome refers to the complete set of genetic material (DNA) of an organism including all the genes (« coding » sequences of DNA), as well as non-coding DNA regions that play important roles in gene regulation. Genomics is a subfield of molecular biology that involve the study of the genome. Although there were several significant milestones that shaped genomics in the preceding century, it truly began in the 1970s with the development of sequencing (Figure 12). Sequencing DNA refers to the process of determining the order of the nucleotide bases (adenine, guanine, cytosine, and thymine) that make up a DNA molecule. DNA sequencing has numerous applications in genetics, genomics, biotechnology, and medicine, including identifying genetic mutations, analysing gene expression, studying evolutionary relationships, and developing personalised medicine based on an individual's genetic information.

Then, metagenomics or environmental genomics appeared[53] and involve sequencing and analysing the DNA of entire communities of micro-organisms from environmental samples such as soil, oceans (Tara, Malaspina, GOS...), rivers, wastewater, or gut microbiomes[48,49,50] (Figure 12). The sequencing targets all the DNA sequences of the sample, which is colossal if one considers the number of genomes that a sample can contain. Bioinformatics methods have been developed to assemble genomes *de novo ;* resulting environmental genomes are called "Metagenome-based Assembled Genomes" (MAGs).

Numerous collections of eukaryotic[58] but mainly prokaryotic[63,64] MAGs have been generated. They have allowed the discovery of a new class of plankton. For exemple, diazotrophic bacteria called Heterotroph Bacterial Diazotrophs which are capable of capturing nitrogen from the atmosphere although they are heterotrophs[64,65]. The MAGs collections have also made it possible to show an important biosynthetic potential in their gene collection, particularly bacterial[66], evolutionary

functional convergences[58], the geographical distribution of functions[58], but also to explore their ecological niches and characteristics in relation to their biogeography[67].

Moreover, a catalog of 40 million prokaryotic genes called Ocean Microbiome Reference Gene Catalog (OMR-GC)[60] and then 47 million[61], and a catalog of 116 million unigenes (cDNA contigs) of planktonic eukaryotes[62] called Marine Atlas of Tara Oceans Unigenes (MATOU) were published following the *Tara Oceans* expeditions. Each contains a large proportion of genes with unknown functions, having no known homologs (~60 % for the eukaryotic catalog[62] and 39 % for the prokaryotic catalog[61]). The low number of reference sequences may explain this observation for eukaryotic planktonic organisms, for which very few reference sequences are described[248].

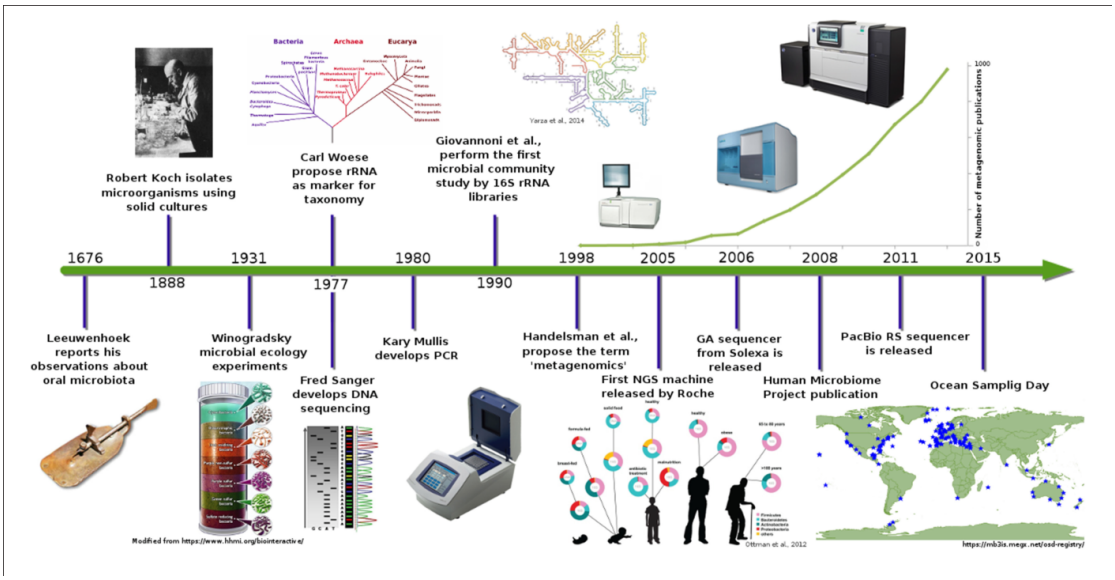


Figure 12: Metagenomics timeline and milestones, showing advances in microbial communities studies from Leeuwenhoek to next-generation sequencing. Figure from[247].

It is worth noting that metagenomic data assembly methods suffer from certain biases. For example, it is difficult to reconstruct large genomes, which is the case for eukaryotes *i.g. dinoflagellates* in marine plankton are under represented because they often have Gb-scale genomes (although they are very abundant organisms). It is also difficult to assemble genomes that are not very abundant because they are not well represented in the metagenomes. Moreover, assembly algorithms are far from assembling the entirety of sequenced DNA reads : for the eukaryotic genomes of *Tara Oceans*, only 20% of assembled reads are reached in MAGs for the best size fraction (20-180µm)[58]. Assemblers also suffer from limitations in assembling repeated sequences, highly variable sequences, regions coding for ribosomal RNAs, transfer RNAs, mobile elements or genes of unknown function. Although they suffer from biases in the estimation of the abundance of the different species present, metagenomic sequencing techniques have revolutionised our understanding of microbial communities (both prokaryotic and eukaryotic environments).

Ribonucleic acid

Transcription is the process by which a ribonucleic acid (RNA) molecule -known as messenger RNA (mRNA) is synthesised from a DNA template (Figure 13). It occurs in the nucleus of eukaryotic cells or in the cytoplasm of prokaryotic cells. During transcription, the DNA double helix is unwound, and one of the DNA strands serves as a template for the synthesis of a complementary RNA molecule. Finally, mRNA is synthesised in a process that is similar to DNA replication, but with the key difference that

RNA uses uracil (U) instead of thymine (T) as a nucleotide base.

Transcriptomics is the study of all the transcripts (mRNA) molecules in a cell or organism. The transcriptome (whole set of mRNA) provides information about which genes are being actively « expressed » in a given tissue or under specific conditions. It is worth noting that a molecule of DNA is present in every cell of an organism and yet it is not the same genes that are expressed in different cell types/tissues. Thus, it is important to keep in mind that in a cell, the transcriptome is not a reflection of all the genes contained in the genome, but rather of the expression of genes required under particular environmental conditions where the genome is expressed. The transcriptome is ultimately a subset of the « functions » encoded in the genome.

Proteins

Afterward transcription, the genetic information carried by mRNA is used to synthesise a protein by the process of translation occurring in the cytoplasm of cells (Figure 13). During translation, the mRNA serves as a template to guide the ribosome in assembling a chain of amino acids in the correct sequence to form a protein. Proteins are the primary functional molecules in cells, performing a wide range of functions. Some of the major ones include the regulation of gene expression (transcription and translation processes), the structural support to cells and tissues, or the transport of molecules within cells and across cell membranes. Proteins can also act as storage molecules, or as chemical messengers that regulate various physiological processes, and so on.

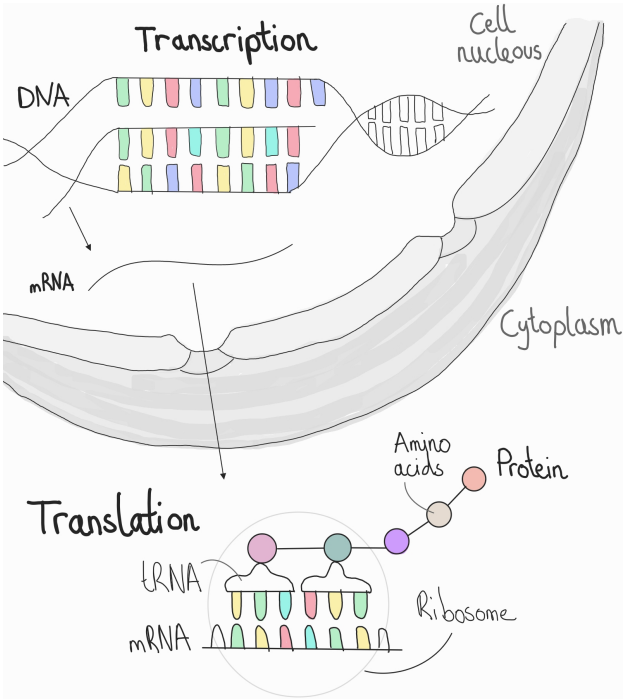


Figure 13: Gene expression in eukaryotic organisms as molecular biology dogma.

In this thesis we are particularly interested by proteins which act as enzymes. They are biological catalysts that facilitate and accelerate chemical reactions implicated in metabolic pathways - metabolism being the complex set of chemical reactions necessary for life, that occur within living organisms, and that allow them to grow, reproduce, maintain their structures, react to their environment and modify it.

To conclude, omics refers to a field of research in molecular, computational and systems biology that involves the comprehensive analysis of a large number of molecules within an organism or a system such as the genome, proteome, transcriptome, metabolome, and epigenome (omics data).

1.2.2.2 A hint of systems biology

Traditional biological approaches often rely on reductionist methods that involve breaking down complex systems into smaller, more manageable parts and studying them individually. It tends to focus on individual components and their functions, such as studying a specific gene or protein in isolation. While this approach is essential in uncovering specific molecular mechanisms and pathways, it may not capture the emergent properties and behaviours that arise from the interactions between components in a complex biological system.

The concept of emergence in biology highlights the idea that "the whole is more than the sum of its parts"[70], and that higher levels of organisation can exhibit properties that do not exist at lower levels ("more is different"[71]). For example, the behaviour of a complex system such as a cell, an organ, or a whole organism, or even an ecosystem, is the result of interactions among many different components, including genes, proteins, metabolites, and environmental factors. These interactions can give rise to emergent properties such as self-organisation, adaptation, and robustness, which cannot be fully explained by looking at the behaviour of individual components in isolation. These emergent properties are closely related to the concept of phenotypic traits, which refers to the observable characteristics of an organism or system at a particular level of organisation. Thus, the ultimate goal of systems research is to develop a systemic understanding of the whole system. This encompasses grasping not only the individual parts and their interactions but also recognising how the system itself enables or restricts certain functions and interaction. We will discuss this issue in more detail section 4.1. With the advancement of technology, it became possible to generate large amounts of biological data, including genomic, transcriptomic, proteomic, and metabolomic data, among others. These omics data sets are often too large and complex to be analysed using traditional biological approaches, which is where computational sciences come into play. Computational tools and algorithms are used to analyze and interpret these large omics data sets, allowing for a more comprehensive understanding of biological systems.

1.2.3 Mathematical models for ocean modelling

1.2.3.1 Defining ethymology of « model »

The term "model" is used in various disciplines and methodological approaches, and its meaning may differ depending on the context[74]. Let's first define ethymology used is this manuscript. In biology, "model" can refer to "model organisms", which are species or strains used as exemplars of groups of organisms due to their ease of growth, manipulation, and analysis in the laboratory. In mathematics, "models" can range from simple statistics to complex dynamic computational descriptions. For example, a statistical model is a mathematical representation of a real-world system or phenomenon, typically based on probabilistic assumptions about the relationships between different variables. It tends to find correlations based on patterns, rather than explain causality. A mechanistic model used for biological questions, on the other hand, is based on an understanding of the underlying physical, chemical, or biological mechanisms that govern a behaviour (interactions between constituent parts of a system). Mathematical models have been extensively developed and applied in the field of plankton ecology and biogeochemistry to enhance our understanding and predictive skills.

1.2.3.2 From modelling population structure…

Modelling pioneer worked in the days before computers[77]. In 1939, Richard Fleming introduced the first dynamical model of plankton populations using a differential equation to study the temporal variability of phytoplankton[249]. He used Lotka-Volterra-type, predator–prey models to interpret the blooms and seasonal cycles in phytoplankton abundance in the English Channel and at Georges Bank, Massachusetts[78]. Generating model results was laborious, at times taking 25–30 hours to solve a

single pair of equations[74]. It took so long because they had to perform the calculations manually – computers, as we know them, did not exist. At that time, part of the reason for the slow acceptance of this novel approach by the oceanographic community was that they felt it too simple to be biologically useful[79].

Beginning in the 1970s, the growing recognition of the usefulness of models was due in part to the fact that more information was available about the dependence of physiological rates on environmental factors.[80,81]. There were more examinations of how modelled dynamics were affected by different formulations and parameterisations of physiological processes. For example, the Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) models[82–84] (are very similar to the Figure 4) have been widely used to study the interactions between different trophic levels in planktonic ecosystems assuming simplified relationships, such as a direct relationship between phytoplankton biomass and zooplankton growth (which may still not accurately reflect the complexity of these interactions).

In the 1990s, increased scope and resolution of observations and increased experimental information became available. Many of these data were products of interdisciplinary research initiatives such as the Joint Global Ocean Flux Study (JGOFS), and the Global Ocean Ecosystem Dynamics (GLOBEC) program, which were born out of a growing interest in understanding the effect of climate variability on ocean production. Ocean models have become increasingly complex over time, introducing more complexity in their formulation[85–87].

I recommend the review by Gentleman *et al.*[74] for those curious about marine modelling history.

1.2.3.3 … to planktonic functional traits…

Planktonic organisms, which include a wide range of microorganisms have evolved diverse metabolic pathways and adaptations to cope with their environment. Their classification as plankton is primarily based on their shared habitat and ecological characteristics rather than a common evolutionary origin. Due to the independent evolution of metabolic pathways in different groups of planktonic organisms over millions of years, their evolutionary relationships may not always align with their functional or ecological similarities. Putting all phytoplankton in the same bag oversimplifies certain phenomena. As a result, researchers have started categorising these organisms into functional groups or biogeochemical guilds based on shared biogeochemical processes or ecological functions.

There has been a growing interest in using trait-based models to study marine ecosystems. A trait is defined as "a well-defined, measurable property of organisms, usually measured at the individual level and used comparatively across species"[88]. These models group organisms based on their functional traits, such as feeding behaviour or nutrient uptake strategies, in order to better understand their role in biogeochemical cycles. Indeed, taxonomy or size assumptions are still typical proxies used to associate organisms with biological functions. For example, *diatoms* is a separate group[89], because they have larger cells with silica frustules, thus need silice and are therefore connected to the global silicon cycle[90]. It is assumed that small phytoplankton doing calcification (*coccolithophore*) are also included into a specific class[91]. Indeed, dense calcium carbonate platelets enhance export of organic matter to the deep and modulate alkalinity, surface ocean carbonate chemistry, and the air-sea equilibrium of CO_2[92]. These phenotypic characteristics allow to create groups that represent aggregates of many species with common biogeochemical functions such as the atmospheric gaseous nitrogen (N_2) fixation into a more usable form such as ammonia, or the

dimethylsulfide (DMS) production[84]. While early global models used only a single compartment to represent biogeochemical dynamics[93,94], they are now capable of simulating the dynamics of dozens[95–97] or even hundreds of planktonic compartments called Plankton Functional Type (PFT) in 3D (vertical and horizontal ocean fluxes)[98], at ocean-scale. And this is these models that are now used to predict the past and the future dynamics of ocean biogeochemical cycles and their feedback on climate[99].

However, many biogeochemical processes are performed by organisms with very different functional traits and ecological roles. For example, calcification can be performed by calcifying phytoplankton such as *coccolithophores*, but also by other diverse planktonic organisms such as *foraminifera*, which are *protozoa*, or *ostracods*, which are small crustaceans. It is also known that the production of DMSP is not restricted to phytoplankton (1.1.4.2). Conversely, having only one generic box, for example for *diatoms*, cannot perfectly represent this whole large group and does not take into account individual variability. Thirdly, while current biogeochemical models, and in particular ocean-climate models, are relatively well constrained in terms of physics and chemistry, they are still based on very simplified representations of biology.

Ocean modelling has maintained consistent goals throughout its timeline, which is to gain a better understanding of ocean processes across space and time[74]. The main objective is to study the interactions between organisms and the environment and to identify the fundamental principles that clarify how ecosystems function, thereby improving predictions of ecosystem change[75,76]. The increased complexity and sophistication of ocean models have allowed for more detailed and accurate simulations, enabling researchers to make more informed precisions about the ocean, its ressources and its future. However, these models do not resolve the immense planktonic taxonomy, physiology and functional diversity, and still struggle to take into account intra-species molecular processes and thus capture the individual variability.

1.2.3.4 …extended with omics data ?

We need to gain a deeper understanding of the biocomplexity of plankton with more accurate representations of the complex interactions between different biological processes and environmental factors (such as biological feedback processes…) which affects the production of key metabolites. We also need to better understand the dynamic emerging properties of marine plankton and their impact on ocean biogeochemistry. Techniques centered on omics approaches could allow to go beyond taxonomic classification, and has the potential to greatly enhance our understanding of plankton diversity and its impact on Earth system functioning. Incorporating genomic data into models can help to identify the functional roles of different genes in plankton metabolism, while transcriptomic data can provide insights into how gene expression changes in response to environmental stressors. Proteomic and metabolomic data, on the other hand, can provide information on the functional proteins and metabolites that mediate key biological processes in the ocean[100].

Bridging the gap between biogeochemical processes and the genome scale is a challenging task for models. Omic-based models like Species Distribution Models (SDMs) provide valuable insights into how the environment shapes the distribution of species or communities, they typically focus on statistical relationships without explicitly considering underlying biological functions or mechanisms[67]. SDMs answer the question "who is where?" but do not explicitly incorporate genomic, metabolic, or physiological information, limiting mechanistic understanding. The question is then, can we find mechanistic models with a connection to omic data ?

This question has been addressed in very recent works. Constraint-based metabolic modelling at genome scale (GSMs) study and predict the behaviour of an organism's metabolism based on its genome information (fully explain in section 1.3, scope of this thesis). The Ocean System Model, Nemo-PISCES, provides information on the concentration of nutrients available for planktonic growth across the global ocean. By connecting GSMs into Nemo-PISCES, it becomes possible to gain an integrated understanding of how gradients in resource stress, as indicated by nutrient concentrations, modulate metabolic reactions and molecular physiology in planktonic organisms. In turn, the GSM estimates the maximal theoretical growth rates for each grid point in the global ocean and the associated internal metabolic fluxes. These preliminary results address the challenge of estimating functional traits while considering biogeochemical information from Nemo-PISCES and the organism's metabolism in a holistic manner. This new integration is presented in the forthcoming paper "modelling genome-scale knowledge in the global ocean" by Regimbeau *et al*., and will obviously be discussed in section 4.2.3, when all the pieces of puzzle of this thesis are assembled.

1.3 CONSTRAINT-BASED METABOLIC MODELLING AT GENOME-SCALE

"Because biological information is incomplete, it is necessary to take into account the fact that cells are subject to certain constraints that limit their possible behaviours. By imposing these constraints in a model, one can then determine what is possible and what is not, and determine how a cell is likely to behave, but never predict its behaviour precisely", Palsson[250]

1.3.1 Mechanistic modelling of the physiology at molecular-scale

Metabolic networks refer to the collection of all metabolic reactions and pathways (within the limits of current knowledge) that occur within a cell or organism (1.3.2). These reactions are the set of life-sustaining biochemical transformations in organisms (photosynthesis, respiration, DMSP anabolism...) that allow them to grow, reproduce, maintain their structure, and respond to their environments (1.2.2.1). These networks can be reconstructed from genomic data using bioinformatic tools (1.3.4), and are the cornerstones of constraint-based models (CBMs).

CBMs are mathematical representations of metabolic networks that take into account the constraints imposed by thermodynamics, stoichiometry, and other physiological factors (1.3.3). These models use optimisation algorithms to predict metabolic fluxes or growth rates under different environmental or genotypic conditions (1.3.3.6). It assumes that the metabolic system is in quasi-steady-state, which means that the rates of production and consumption of all intracellular metabolites are balanced. This assumption allows for the calculation of metabolic fluxes without the need for detailed kinetic data (1.3.3.4). Initially CBMs are used for modelling cellular physiology and growth of model organisms[102], however, extensions of these constraint-based approaches are emerging for predicting and understanding microbial communities[103–107].

The following sections will describe in more detail the nature of CBMs and how they are reconstructed to model the metabolic behaviours of target organisms.

1.3.2 From genomes to metabolic networks...

Metabolic networks contain the metabolic capabilities encoded in organism's genomes. Indeed, from a genome, it is possible to predict the encoded genes[108] and thus, identify the corresponding enzymes and their associated metabolic reactions (Figure 14 ; 1.2.2.1). The correspondence from metabolic genes to enzymes to reactions is not straightforward and requires lots of genetic and biochemical knowledge[109].

It is possible to construct a set of Gene-Protein-Reactions (GPR) rules that will gather the requirements for the production of each enzyme in terms of the presence/absence of genes. By applying those rules to a genome, one can extract the enzymes that the genome can produced, and thus deduct the set of metabolic reactions which can occur (Figure 14).

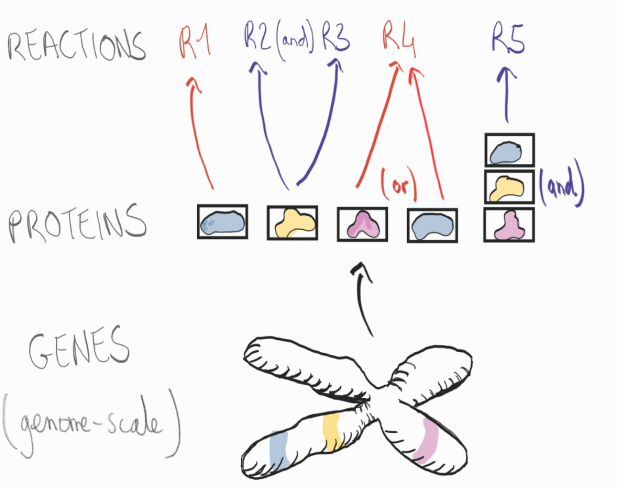


Figure 14: Logical conjunction of genes coding for enzymatic reactions depicting the biological process of gene expression (1.2.2.1).

Each protein (sub)unit is encoded by a gene. From left to right : Monomeric enzymes consist of a single protein subunit, and are generally small in size and have a simple three-dimensional structure. There are multifunctional enzymes that contain several active sites that are able to recognise and bind to different substrates, allowing the catalysis of several different biochemical reactions (R2 and R3). Or different isoenzymes that can catalyse the same biochemical reaction (R4). Sometimes several polypeptide subunits, identical or different, are needed to define an oligomeric enzyme, e.g. ATP synthase (ATPS), 2.2.2.1.

Once individual components are identified, the data is integrated to study the behaviour of the biological system as a whole. A metabolic network can be graphically represented as a bipartite graph (Figure 15), with reactions (red squares) and metabolites (blue circles) as nodes. Edges (arrows) connect reactions to metabolites involved in the catalysed biochemical transformations. If the metabolite is a product, one edge should be directed toward the metabolite, and if it is a substrate, another should depart from it. Reversible reactions, like R1, can produce (pink arrow) or consume (grey arrow) a metabolite like M5. Cooperative interactions between reactions are evident as some products serve as substrates for others (e.g., M5 in R1, R2, R4, and R5). Functional pathways can be depicted as paths from one metabolite to another such as the pink or green arrows.

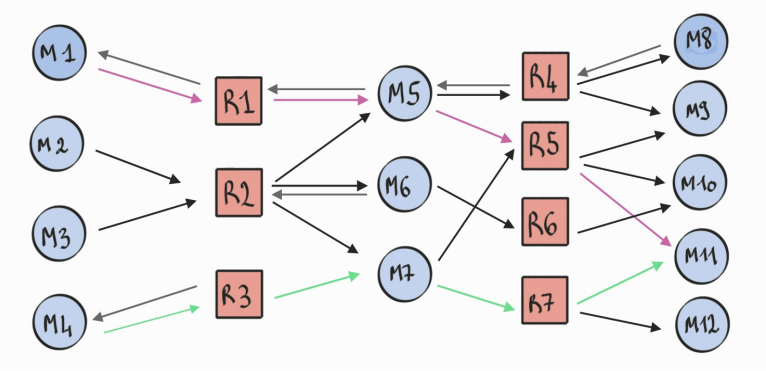


Figure 15: Bipartite graph depicting a metabolic network. Reactions are red squares; metabolites are blue circles.

In summary, a metabolic network piles up biological and chemical knowledge. In order to study the physiology of microorganisms through mechanical metabolic processes, it is necessary to transform the network into an adequate model, which implies modelling assumptions.

1.3.3 ...to constraint-based metabolic models

Metabolic models are used to infer reaction rates, also known as fluxes, without using kinetic parameters. The mathematical wrap is the last thing we need to make our model. It is what we are going to explore in the following parts (1.3.3 ; 2.1.2).

1.3.3.1 Exchange reactions

The continuous supply of metabolites from and to the media is facilitated by exchange reactions (blue edges in Figure 16). They are are responsible for uptake or secretion of nutrients, waste products, or signaling molecules by cells, thus exchange of metabolites between the environment and the system we are modelling.

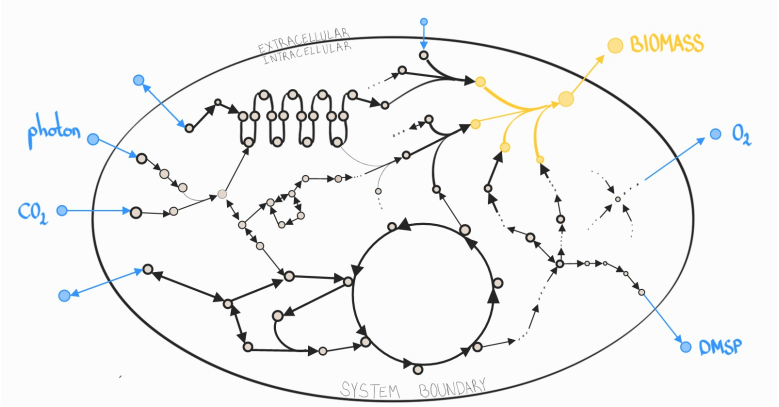


Figure 16: Unipartite graph depicting a metabolic network. Arrows are reactions; metabolites are nodes as round shapes. Reactions link metabolites. The biomass reaction is in yellow (as explained in 1.3.3.2). The external metabolites are blue circles (characterising the system environment as explained in Figure 26), and blue arrows are transport reactions. The internal metabolites (intern to the system) are in beige.

Exchange metabolites are important in metabolic modelling because they represent the interface between the metabolic network and the external environment, and they can have a significant impact on the behaviour and properties of the network. Their uptake or secretion rates can be constrained based on experimental measurements or estimated using optimisation methods. If metabolite exchange were not possible, then for each reaction the only possible state would be the chemical equilibrium, with all net fluxes equal to zero[110].

1.3.3.2 Biomass reaction

Finally, modellers developed the fictive biomass reaction (yellow arrows in Figure 16) to model the growth rate of organisms[111]. This reaction encompasses the needs of the modelled system and also the energy cost of cellular division or cell maintenance[112–119]. It often accounts for constituents of the five major cellular macromolecules (carbohydrate, DNA, lipid, protein, and RNA), and their fractional contributions to the overall cellular biomass[119].

The functionality of a metabolic model is typically characterised by its capacity to traverse the graph, starting from source metabolites (such as nutrients available in the environment) and reaching targets (such as biomass constituents or DMSP production for example), resulting in interdependencies between uptake and secretion reactions that are intricately connected to downstream metabolic processes.

1.3.3.3 Stoichiometric matrix

A metabolic network is formally described by its stoichiometric matrix $\mathbf{S} \in \}n,\ m$ describing the relationship between the m metabolites and the n reactions (Figure 17). Each row of the matrix represents a metabolite Mi, and each column represents a reaction Rj. The entrie Si,j of the matrix is the stoichiometric coefficient of the metabolite Mi in the reaction Rj. By convention it is negative if the metabolite is a substrate, positive if the metabolite is a product and null if the metabolite is not implicated in the reaction. This matrix shows the relationships between the reactants and the products of a set of reactions, and is a fundamental tool in metabolic network analysis to study the behaviour of metabolic systems.

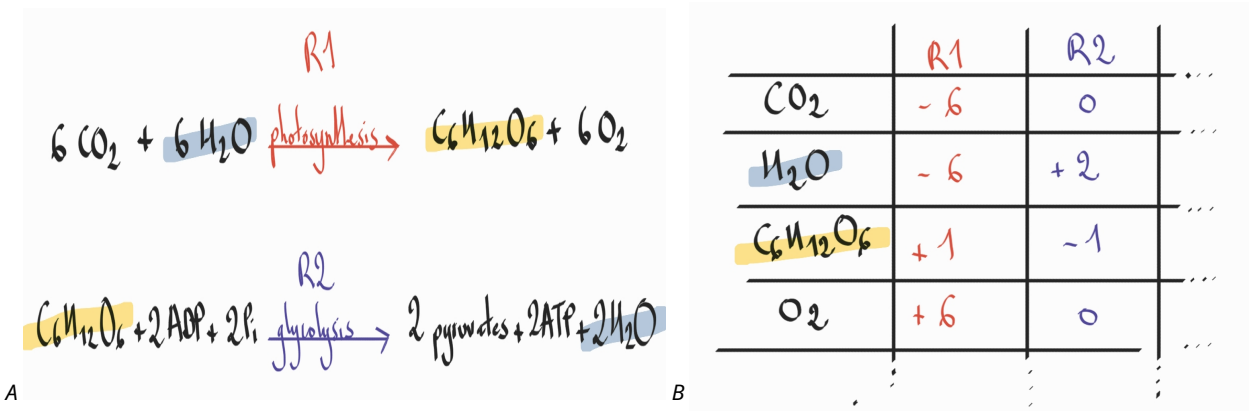


Figure 17: Stoichiometric matrix connect metabolites and reactions through stoichiometry ; A) Examples of biochemical reactions ; B) Stoichiometric matrix of metabolic networks. Each row of the matrix represents a metabolite M_i, and each column represents a reaction R_j. The entrie $S_{i,j}$ of the matrix *is the stoichiometric coefficient of the metabolite M_i in the reaction R_j.*

1.3.3.4 Quasi-steady state assumption constraint

The change over time of the concentration of the metabolite M_i is given by the mass-balance equation:

$$\frac{d[M_i]}{dt} = S_{i1}v_1 + \ldots + S_{in}v_n = \sum_{j=1\ldots n} S_{ij}v_j ,$$

where v_j is the reaction rate or flux associated to reaction R_j . The concentration of a metabolite over time depends on its rate of consumption and production in all the reactions in which it is involved with its respective stoichiometries. Using a vector notation, the above equation can be written as:

$$\frac{dM}{dt} = Sv \quad , (1)$$

where $\mathbf{M}$ is the vector composed of the concentration of each metabolite M_i, and $\mathbf{v}$ is the flux vector composed of each flux v_j. In general, the rate of reactions depends on metabolite concentrations and other parameters influencing enzyme kinetics, such as temperature, or pH. However, determining these parameters and the function of reaction rate are complex experimental tasks. Moreover, these parameters are in general very sensitive to biochemical conditions, so *in vitro* determinations may not correspond with *in vivo* values[122]. Thus solving Eq. 1 is a daunting task for genome scale systems.

Indeed, metabolic reactions within living organisms occur at high rates, allowing for rapid adjustments and responses to changes in the external environment. This inherent rapidity enables organisms to efficiently adapt their internal biochemical processes to counteract external disturbances. Thus, CBMs assumes that organisms maintain homeostasis by regulating internal concentrations to remain as constant as possible. This is achieved by ensuring that the rate of formation of internal metabolites is equal to the rate of their consumption. Consequently, the internal system is considered to be in a quasi-stationary state (QSSA), leading to:

$$Sv = 0. \quad (2)$$

QSSA is a simplifying assumption preventing the integration of kinetic parameters and is used to reduce the complexity of large-scale metabolic models because it allows linearity of the equations. In general, the QSSA is assumed to be valid because of the timescale separation between (fast) intracellular metabolic conversions and (slow) genetic regulation[124–126]. However, it is important to note that the assumption may not always hold true in real biological systems, and therefore the results obtained from models using the QSSA should be interpreted with caution.

1.3.3.5 Thermodynamic constraints

In addition to this system of linear equations, we also consider thermodynamic constraints of reactions fluxes, expressed in mole of product formed by gram of dry weight of the considered organism by hour (mol.gDW^{-1}.h^{-1}). The fluxes are not infinite, and we assume the following inequalities for each v_j:

$$lb_j \leq v_j \leq ub_j \text{ , (3)}$$

where lb_j represents the lower bound of the flux v_j, and ub_j represents its upper bound. A positive flux means that the reaction is occurring in its forward direction, whereas a negative flux means that it is occurring in the reverse direction (Figure 18). For instance, if the reaction is known to be direct and irreversible[‡], it means that the flux cannot be negative. Eq. 3 becomes:

$$0 \leq v_j \leq ub_j \text{ . (4)}$$

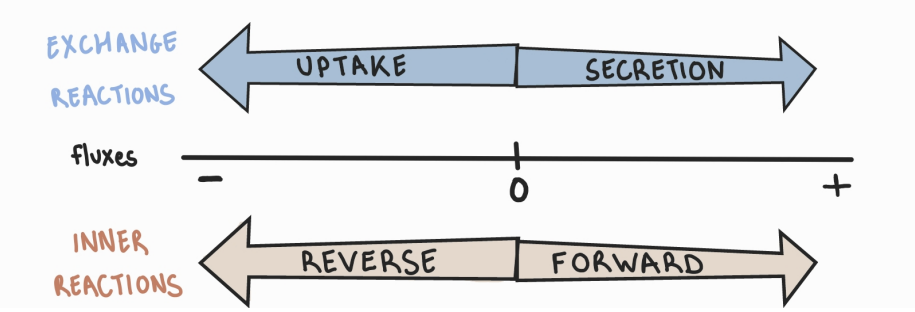


Figure 18: Definition of reaction fluxes (in mol.gDW^{-1}.h^{-1}).

For inner reactions, a positive flux means that the reaction is occurring in its forward direction, whereas a negative flux means that it is occurring in the reverse direction. For boundaries reactions (Exchanges, Sinks, Demands), a positive flux means that the metabolite goes to the environment, whereas a negative flux means that the system uptake from the environment.

[‡] *The reversibility of a reaction can be checked experimentally or inferred computationally. Reversibility of reactions can be check experimentally, e.g. measuring the rate or equilibrium constant of the forward and reverse reactions under different conditions such as pH, temperature, substrate and product concentrations, or using isotopic labeling techniques to trace the flow of metabolites through a reaction. Alternatively, computational methods such as thermodynamic analysis can be used. Gibbs free energy is a measure of the energy available for the reaction to occur. In thermodynamics, a reaction is considered to be spontaneous and feasible if ΔG° is negative, indicating that the reaction can occur without external energy input. Conversely, if ΔG° is positive, the reaction is considered non-spontaneous and cannot occur without an input of energy. When ΔG° is zero, the reaction is at equilibrium. The standard Gibbs free energy change (ΔG°') of a reaction can be calculated using the following equation: ΔG°' = ΣnΔG°f(products) – ΣmΔG°f(substrates), where ΔG°f is the standard Gibbs free energy change of formation for the reactants and products, n and m are the stoichiometric coefficients for the products and reactants, respectively.*

Altogether, Eq. 2 and 3 form a CBM of the corresponding organism and can be resumed as :

$$\mathbf{Sv = 0}, \quad (5)$$

$$\mathbf{lb \leq v \leq ub.}$$

1.3.3.6 *Exploration of the solution space*

All solutions of Eq. 5 define the flux space of the system (Figure 19). The allowable solution space represents all possible metabolic states of the system that satisfy all the constraints.

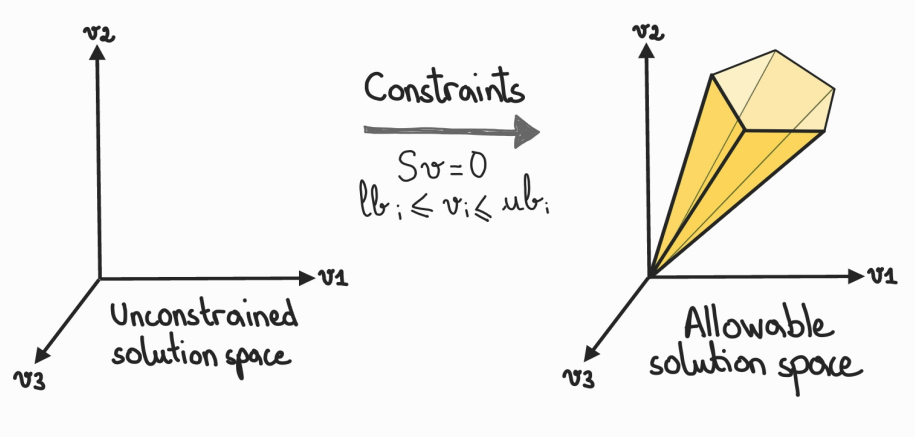


Figure 19: Definition of the flux space. In an n-dimensional space, with n the number of reactions in the model, the allowable solution space is defined by the steady-state assumption and the thermodynamic constraints. Each point of this space corresponds to a possible metabolic state (in terms of reactional fluxes) of the network, satisfying all the described constraints. Figure adapted from [271].

This flux space may be analysed through several state-of-the-art approaches (Figure 20). I will briefly describe some techniques used during this thesis. For a detailed review of these methods, the reader may wish to refer to[127–129].

Flux Balance Analysis (FBA) is a commonly used approach, as depicted in (Figure 20.A). FBA aims to optimise the flux of an objective reaction, typically by maximising or minimising it, often representing the growth rate of the organism. In linear programming, it is known that if an optimal value for the objective reaction exists, it is unique. However, the same cannot be guaranteed for the flux vector. As a result, there can be multiple flux distributions that could potentially yield optimal objective function values. To address the existence of multiple optimal flux distributions, the Flux Variability Analysis (FVA) technique was developed[251]. FVA aims to explore the range of feasible flux values that satisfy a given optimal objective value. By applying FVA, we obtain a range of values for each flux, providing a comprehensive understanding of the solution space surrounding the specified conditions.

On the other hand, sampling the flux space (Figure 20.B) involves exploring the feasible solution space of metabolic flux distributions, enabling the generation of a representative set of flux distributions that adhere to the given constraints. Sampling procedures provide a comprehensive understanding of the metabolic landscape under the defined constraints. It allows researchers to explore different potential metabolic states and assess the range of phenotypic behaviours exhibited by the organism or community and offers a valuable tool for studying the functional potential and plasticity of metabolic systems.

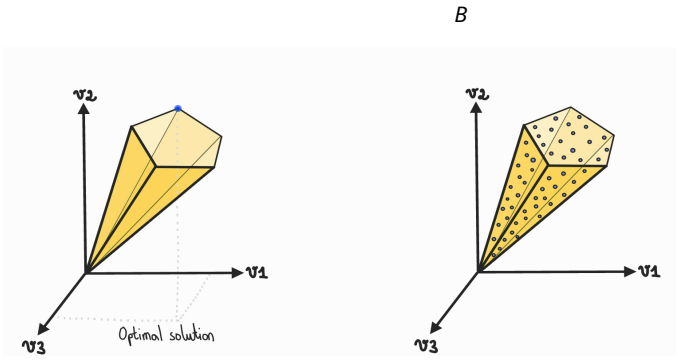


Figure 20: The allowable flux space, satisfying all constraints, may be analysed through several state-of-the-art approaches. A) Flux-Balance Analyses (FBA) optimises i.g. maximising or minimising the flux of an objective reaction by optimisation (most often growth) ; while B) the random sampling of the flux space allows us to obtain several thousand possible metabolic states of the model.

Initially, the concept of metabolic niche[105] is a fundamental technique which allow the projection of the allowable solution space to conceptualise a volume whose dimensions correspond to the a set of metabolic environmental conditions (in terms of reaction fluxes) in which the organism can growth. This is why this concept is called *metabolic niche*. By optimising computational efficiency, this technique facilitates the integration of Genome-Scale Models (GSMs) with Earth System Models (ESMs) (Regimbeau *et al*. under review), as introduced in (1.2.3.4). In the context of ESM, these chosen axes represent sets of environmental parameters in the form of fluxes of available metabolites provided by the ESM. The authors establish a connection between metabolic requirements and survival, while considering the inherent biological complexity of the metabolic network.

In this manuscript, we re-use the metabolic niche technique for characterising metabolic functions, with a specific focus on key internal or external reactions. We represent this function as a reduced flux space, which provides insights into the organism's flexibility and adaptability concerning these critical reactions. By utilising the metabolic niche concept, we aim to describe the continuous phenotypes exhibited by the organisms under study, enabling us to capture the nuanced variations in their metabolic capabilities and adaptations.

To conclude, constraint-based metabolic modelling allow quantitative and computable genotype-phenotype relationships of target organisms. They aim to study the effects of environmental perturbations or genetic modifications on metabolic fluxes, and to understand how individual components interact to give rise to emergent properties and behaviours at the systems level.

1.3.4 Two approaches for the reconstruction of metabolic models

To date, numerous ecologically relevant GSMs are already available for prokaryotes and archaea (BiGG[130], EcoCyc[252], CyanoCyc[not published]). However, models for marine eukaryotic microbes lag behind, mainly because of the scarcity of "model organisms" whose genomes are sequenced, and also because of the tedious manual curation steps required to obtain effective models.

In practice, as genomic knowledge is not exhaustive, simply adding up metabolic reactions coming from metabolic genes and modelling reactions (boundaries reactions, biomass reactions…) might end

up in incomplete or absent pathways. For a metabolic network to be complete, one needs to add yet other reactions. To date, two main conceptual approaches are used to solve this problem (Figure 21) : the traditional bottom-up approach using gap-filling, and the top-down one using graph refinement ("carving"). The top-down approach is the most appropriate one for our biological questions and data.

1.3.4.1 Traditional bottom-up approach

Until now, only bottom-up approaches were available for eukaryotes (aureme, aucome, metadraft, merlin, modelseed, pathwaytools or raven). However, bottom-up approaches are not the best suited for the modelling of marine planktonic microeukaryotes from environmental omics data.

To reconstruct organism-specific metabolic networks, a bottom-up approach is commonly employed, wherein pathways are filled by adding reactions one by one (Figure 21). This process can be carried out manually or semi-automatically : either by extensively reviewing literature and culturing the organism (first, most planktonic organisms are still undescribed and not cultivable, and secondly we can't do this work for hundreds of organisms) ; or by utilising an evolutionary approach that incorporates pathways from related organisms based on taxonomic proximity[274]. However, recent studies have shown that the gene content of planktonic communities is more informative in relation to biogeochemical gradients than taxonomic information, particularly for microbiomes[100, 272-273]. So this approach is contrary to our scientific convictions. The reconstruction process involves incorporating quality control and validation procedures to ensure the accuracy and biological relevance of the resulting models. However, their criteria are based on correlating the topology of the networks with the taxonomy[274], raising questions about the potential circularity in the process.

Bottom-up approaches often require substantial computational resources and time, especially for large-scale network reconstructions. Although dedicated workspaces like AuReMe[275] have been developed to facilitate model reconstruction, a certain level of familiarity with computational modelling and programming is still necessary to utilise them effectively. One of the difficulties of bottom-up approaches is to obtain high quality input data, which is not really appropriate when one wants to use MAGs and knows their limitations (1.2.2). While bottom-up reconstructions allow for the construction of metabolic networks, they may not accurately predict metabolic fluxes within those networks, which are crucial for understanding metabolic function. They usually focus on topological approaches which are suited for analysing the network's structure. Constraint-based approaches, on the other hand, focus on predicting flux distributions in metabolic networks. In my opinion, this is an additional layer of information that is not negligible.

1.3.4.2 Top-down approach

The top-down approach as implemented in CarveMe[133] shifts these paradigms by introducing a generic meta-model for which curation is done only once. This meta-model considers all chemical knowledge (from the whole gene repertoire including both core and variable genes) of a set of organisms (prokaryotes Gram+, Gram-, or archae so far) into one extensive network. The generic model is manually curated to make it ready-to-use for constraint-based analyses (Chapter 2 is entirely devoted to this process).

This model is then converted to organism-specific models while preserving the whole manual

curation and relevant structural properties. This carving process aims to maximise the number of reactions with genetic evidence while importing the minimum of reactions that allow the organism to keep growing. In other words, CarveMe adds the missing reactions based on fluxes (to maintain growth) rather than on taxonomy or topology. In my opinion, it has the right philosophy by directly considering biological functions at higher levels. CarveMe is the tool with the easiest handling, and the reconstruction of organism-specific CBMs (directly ready for constrained-based analyses) takes only about 3 minutes (once the time-consuming task of producing the meta-model is done).

It has been shown a good reproducibility of results : the performance of CarveMe models is close to experimental phenotypes compared to manually curated models[133,253, this work]. Moreover, the annotated genome does not need to be complete, and the environment do not need to be known beforehand.

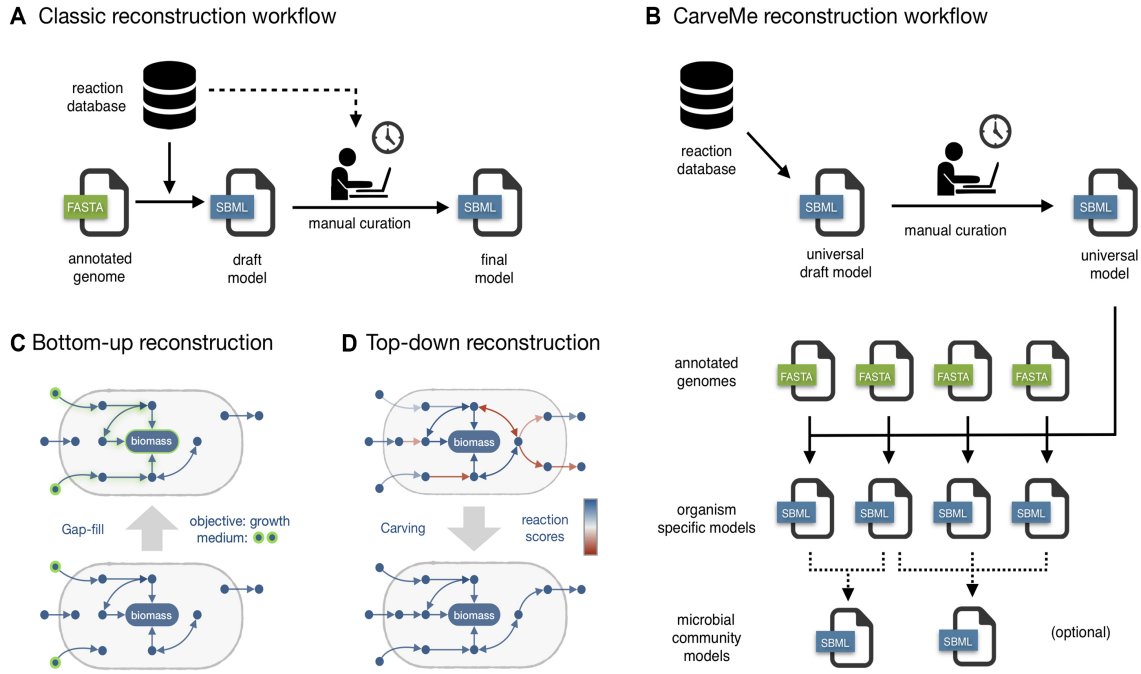


Figure 21: Top-down versus bottom-up approaches for metabolic model reconstruction. The bottom-up approach considers the reactions associated with genes, and then tries to fill the gap between reactions in order to make pathway usable and the network functional. Many formulations of this problem have been proposed[253]. The top down approach considers all chemical knowledge into one extensive network and then removes the maximum of reactions that are unused or without omic evidences, while keeping a functional network (Figure from [133]).

Globally, CBMs reconstruction is particularly tedious in traditional bottom-up approaches since they must be performed for each new model reconstruction. Moreover, these approaches rely on a reductionist philosophy that focuses on the smallest parts of the system, while the top-down approach focus on the system as a whole in an holistic way. Until now, top-down technique was only applied to prokaryotes. Prokaryotes have been studied for a long time and are the best described organisms, mainly because they are much more easily cultivated. There is an urgent need to introduce eukaryotes into our analyses and ecological studies.

1.4 AIMS OF THE THESIS

Biogeochemical cycles are essential processes that involve the transformation, transport, and recycling of molecules on our planet. These cycles play a vital role in sustaining life by regulating the availability of essential elements. Indeed, organisms rely on a continuous supply of specific molecules, such as nutrients and essential ions, to thrive and prosper. Biochemical compounds undergo metabolic processes within organisms to either incorporate or transform them into vital molecules like DNA, proteins, lipids, and carbohydrates. These metabolic processes and material transformations contribute to the overall functioning and survival of organisms. Microbial ecosystems, in particular, play a crucial role in sustaining a stable and habitable environment. These ecosystems interact with the environment, shaping the development and evolution of organisms in intricate ways. Marine planktonic organisms play a pivotal role in shaping major biogeochemical cycles, influencing Earth's climate and weather patterns. Additionally, they occupy a critical position in marine food chains, serving as the primary source of nourishment for numerous aquatic organisms. However, these essential organisms face vulnerabilities due to various environmental stressors, including pollution, ocean acidification, and climate change. The impacts of these stressors can extend throughout the entire marine ecosystem, affecting its overall health and stability. To enhance our comprehension of plankton diversity and its profound significance in Earth's system dynamics, a collaborative effort encompassing diverse fields of research is imperative.

Currently, the availability of environmental metagenomes and metatranscriptomes provides valuable insights into the vast diversity and functional roles of both prokaryotic and eukaryotic plankton within complex ecosystems, directly from environmental samples. However, it is important to recognise that omics data alone cannot address all the challenges at hand. While these datasets offer a wealth of information, their integration into mathematical models holds great promise for advancing our understanding. Genome-scale metabolic models (GSMs) provide a mechanistic approach by establishing quantitative and computable genotype-environment-phenotype relationships for target organisms.

Planktonic Functional Trait (PFT)-based models, on the other hand, emphasise how the environment shapes specific functional traits (and the opposite too). Ocean system models (OSMs), for instance, describe biogeochemical phenomena at the ocean scale. However, they do not incorporate omics data or account for the complex biological mechanisms underlying these phenomena. Therefore, the long-term goal include predicting physiological processes, such as planktonic organism growth or the production of key molecules, by considering the full range of biochemical reactions rather than simplifying them to physical equations. Additionally, the aim is to move away from systematic associations between taxa and functions, recognising that functional traits can be present in diverse organisms, and to integrate intra-individual variability and plasticity.

The integration of GSMs with OSMs show considerable potential in achieving these goals. This integration, which has been recently explored, aims to bridge the gap by incorporating omics data and considering the complex biological mechanisms underlying biogeochemical dynamics. Although still in its early stages, the preliminary results of this integration show promising potential for advancing our understanding of planktonic organisms and their ecological and biogeochemical functions.

Obviously, there is still room for improvement, particularly in obtaining a sufficient number of GSMs that can accurately represent the vast taxonomic and functional diversity of plankton. Currently, there are numerous ecologically relevant GSMs available for prokaryotes, but models for eukaryotes are lagging behind. This lag can be attributed to several factors, including the limited availability of model organisms with fully sequenced genomes for eukaryotic plankton. Additionally, the manual curation required to construct effective GSMs can be particularly tedious and time-consuming, especially in traditional bottom-up approaches where curation must be performed for each new model reconstruction. To address these challenges, the top-down approach offers a promising solution. This approach involves the development of a generic meta-model that undergoes curation only once. From this meta-model, ready-to-use organism-specific models can be derived, preserving the manual curation and important structural properties. Until now, this technique was only applied to prokaryotes.

The main objective of my research was to develop PhotoEukStein, a novel generic meta-model designed specifically for the fully-automatic reconstruction of eukaryotic-algae metabolic models. This meta-model represents a significant advancement in the field by streamlining the process of model reconstruction for eukaryotes. Furthermore, as part of the thesis, a comprehensive database was created, consisting of 549 GSMs derived from environmental genomes and transcriptomes. These GSMs provide a valuable resource offering new opportunities for understanding the complex metabolic networks and ecological implications of these eukaryotic organisms in various environmental contexts.

By transcending simplistic associations and embracing a comprehensive and integrative approach, we can attain a deeper understanding of the intricate complexities present within planktonic systems. It is crucial to recognize the limitations of current methodologies and explore novel avenues that enable the study of functional traits in a more nuanced manner. By doing so, we can unravel the elaborate mechanisms that underlie the emergence of diverse functions within these systems.

Rather than relying on simplistic correlations or isolated observations, adopting a holistic perspective would allow us to capture the multifaceted nature of biological systems. It empowers us to explore the intricate interplay between genetic diversity, environmental dynamics, and ecosystem functioning, ultimately leading to a more nuanced and accurate portrayal of the complexities inherent in these vital ecosystems. In order to achieve this, it is imperative to acknowledge the existing limitations of current research methodologies and to actively seek innovative approaches.

2 PhotoEukStein allows fully automatic reconstruction of GSMs for phototrophic microeukaryotes

In the novel "Frankenstein" by Mary Shelley (1818), Victor Frankenstein collects body parts from various sources, including graveyards and slaughterhouses, to assemble a creature's body. He selects the parts that he believes will create the perfect human form, and then uses his knowledge of biology and chemistry to bring the body to life through a process of galvanism, which involves using electricity to stimulate the muscles and create movement.

In the shoes of Victor Frankenstein, I merged the available biochemical and genomic information of 15 eukaryotic algae and 1 land plant mainly from BiGG[130] and BioCyc databases (Figure 24) to assemble PhotoEukStein. PhotoEukStein is an hypothetical meta-organism (Figure 22) that combines metabolic features of photosynthetic eukaryotic cells (using light energy to convert carbon dioxide into organic compounds). The draft metabolic network have been brung to a curated constraint-based meta-model using knowledge in biochemistry, cell biology and computer modelling. Combined with top-down technique[133] (Figure 21), this new generic model enables fully-automatic reconstruction of constraint-based models at genome-scale (GSMs) for phototrophic microeukaryotes.

Figure 22: PhotoEukStein by DALL-E.

2.1 PHOTOEUKSTEIN RECONSTRUCTION

2.1.1 From the merging of reference metabolic networks…

2.1.1.1 Generating a draft network of PhotoEukStein

Input data for PhotoEukStein reconstruction

Historically, most of the detailed biochemical, biophysical and molecular biological information about eukaryotic photosynthetic processes comes from studies of higher plants and a few model algae, including *Synechocystis*, *Chlamydomonas*, *Chlorella, Thalassiosira* and *Phaeodactylum* (Figure 23). Traditionally, most model organisms have been chosen because they are easily grown or can be genetically manipulated rather than because they are ecologically relevant.

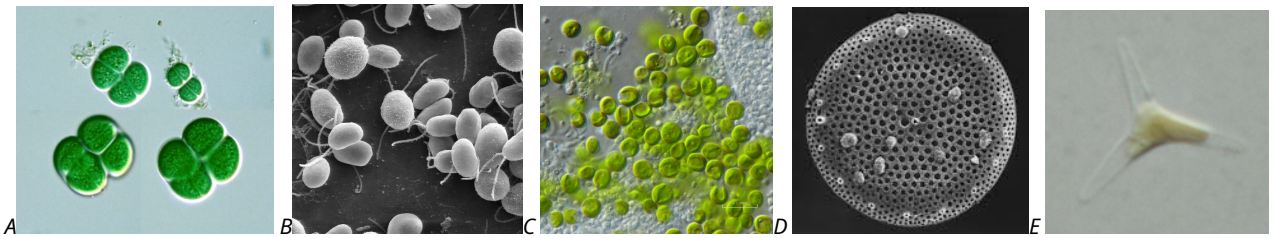


Figure 23: Few model organisms as proxy for phototrophic plankton (scale not respected).

A. Synechocystis sp.~2µm, B. Chlamydomonas ~10µm, C. Chlorella variabilis ~2-6µm, D. Thalassiosira levanderi ~8-16µm, E. Phaeodactylum tricornutum ~3µm. Pictures from nordicmicroalgae.

While several databases store biochemical and genomic data about phototrophic eukaryotes, including Eukprot[134], Phytozome[135], KEGG[136], AlgaeBase[137], and Diatomics[138], only a few provide easy access to the logical conjunction of genes that guarantee the existence of a metabolic reaction. These gene-protein-reaction associations (Figure 14) are directly available in Pathway-Genome-DataBase (PGDB) or metabolic model files facilitating the reconstruction process and the use of the generated models.

Metabolic networks for eukaryotic algae can be located in databases such as BiGG[130] and BioCyc[140], or directly from literature sources[129,141–146]. Within these two databases, I specifically targeted organisms that are photoautotrophic. To further narrow down the selection, I excluded models related to terrestrial plants, except for *Arabidopsis thaliana*, which is extensively studied and well-documented. Additionally, I excluded parasitic organisms that likely possess unique metabolic pathways associated with their adaptive strategies. After careful review of all available data, the biochemical and genomic information of 15 eukaryotic algae and 1 land plant was chosen as raw material for the construction of PhotoEukStein (Figure 24). These data benefit from (i) the curation efforts that have been applied to the biological entities that constitute the biological networks, and/or (ii) the curation efforts that allow these entities to cooperate dynamically and to bring out interesting properties.

Components of metabolic networks (reaction, metabolites, genes) are represented in multiple file

formats and also using different markup languages, with varying levels of annotations ; this leads to inconsistencies and increases the complexities in comparing and analysing reconstructions[147]. For example, SBML (Systems Biology Markup Language) and XML (Extensible Markup Language) are two file formats used to encode and share metabolic models, and differ in their specific structure and syntax. In order to merge them, a formatting step is necessary.

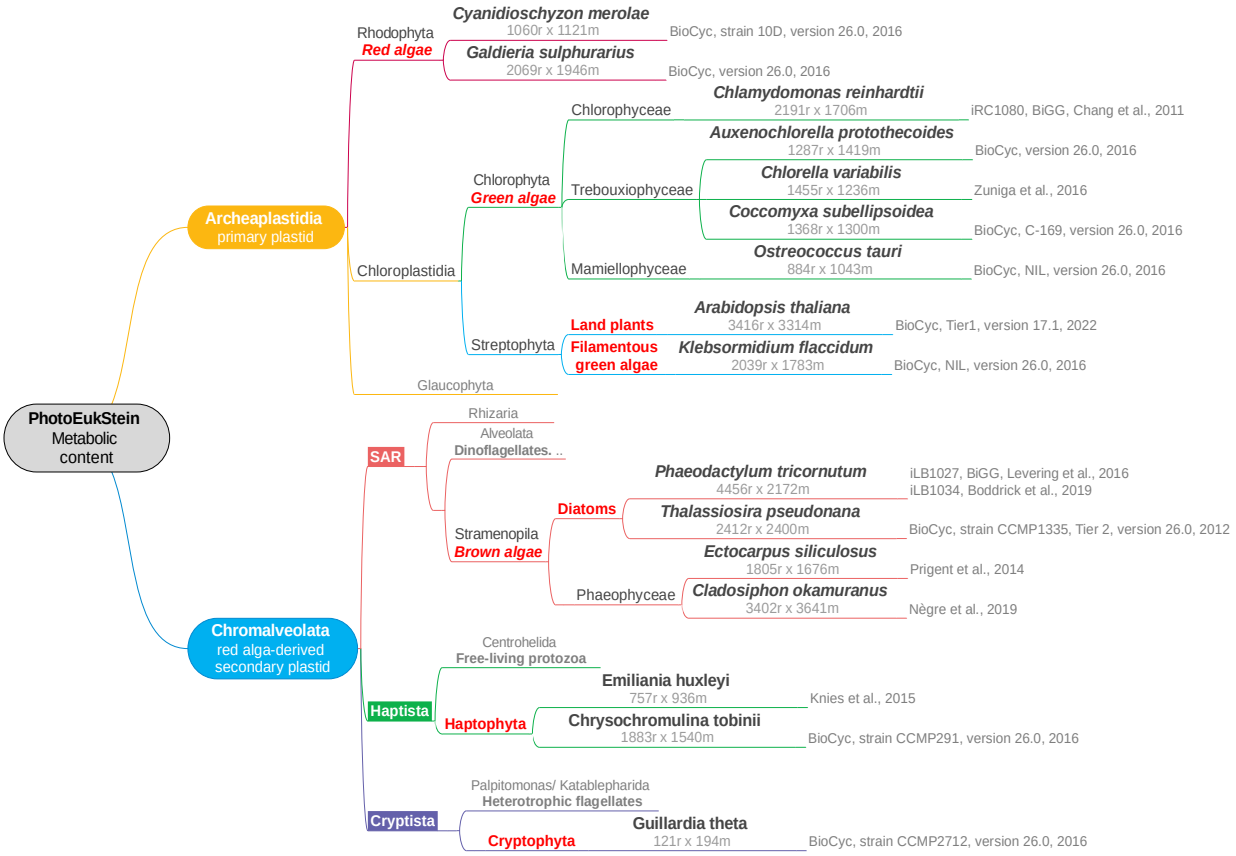


Figure 24: Genomic and biochemical informations from 15 eukaryotic algae and 1 land plant are merged to create PhotoEukStein.

A namespace for standardising identifiers between databases

In the context of biological databases and data integration, using different identifiers for the same entity can create confusion and make it difficult to merge data from different sources. In addition, these duplicates are present within the databases themselves. In order to ensure that different databases or sources use the same identifier for a given entity (such as a gene, a protein, reactions, metabolites...), one can create a namespace that refers to a system of unique identifiers assigned to entities in order to standardize them. Thus, researchers can more easily integrate data from different sources, perform cross-database queries, or analyze datasets across multiple studies.

MetaNetX[148] is an online platform that provides tables for mapping identifiers for metabolites and enzymatic reactions. Despite efforts to reconcile metabolic databases, such heterogeneity still requires meticulous manual curation. The management of identifiers within and between databases presents significant challenges, and this cleaning process is time-consuming. However, through my work a more comprehensive table for converting identifiers from BiGG and MetaCyc, to BiGG is now available,

complementing the one on MetaNetX for these two databases. With this new table, I further identified 2,870 duplicated metabolites along the PhotoEuStein reconstruction process. By integrating diverse data sources and standardising identifiers, this table streamlines the analysis and comparison of metabolic pathways for phototrophic microeukaryotes. Fellow curators, please find this valuable resource at https://www.genoscope.cns.fr/PhotoEukStein/photoeukstein_manual_curation/.

Compartmentalisation

Algae exhibit intra-cellular compartmentalisation differences in various biological processes. For instance, glycolysis occurs in the mitochondria of diatoms, but in the cytosol of other eukaryotes[149,150]. To enable seamless integration of the metabolic pathways of 16 different organisms into a single supraorganism, all enzymatic reactions were assumed to occur in a single compartment. To achieve this process, transport reactions between compartments were eliminated, and all duplicated reactions were removed (Figure 27 for example of duplicated reactions).

2.1.1.2 Cleaning loop for a mass-balanced PhotoEukStein network

Essentiality of mass balance for stoichiometry-based models

The ultimate aim of the curation process is to prepare a metabolic model for constraint-based analysis. Constraint-based models rely on the mass conservation law (Antoine Lavoisier, 1789), which assumes that the metabolic system is in a quasi-steady state (1.3.3.4). According to this law, the total mass of a closed system remains constant over time and cannot be created or destroyed. This principle is crucial for metabolic models because it ensures that the reaction stoichiometry is correctly balanced. This balance is essential for the accurate modelling of metabolic pathways and the prediction of metabolic fluxes, without relying on detailed kinetic data. A reaction is mass-balanced if the elements counts are the same on the left- and right-side of the reaction. Because protons and water are often omitted from ressources (2.1.2.2), unbalanced reactions can generate false proton gradients, leading to energy (ATP) synthesis from out of nowhere[109](2.2.2.1). Therefore, the first step in manual curation is to mass balance all reactions using the chemical formulas for all metabolites. However, at this stage of the process, we consider that a reaction is balanced even if an atom of hydrogen or a molecule of water (H_2O) is missing (2.1.2.2). We mainly consider here the backbone atoms of molecules (such as carbon, nitrogen, phosphorus, sulphur…) (Figure 1).

Mass balancing procedures

I added missing chemical formulae (3406 missing formulae/ 7467 metabolites) using a combination of methods including MetaNetX, manual curation, and a 'home-made' prediction algorithm. To predict the chemical formulae of metabolites, the algorithm begins by identifying all balanced reactions (in which all metabolites have obviously a known formula). The algorithm tags the involved-metabolites formula as correct if they are only involved in balanced reactions. Next, the algorithm looks for reactions having only tagged-metabolites except one missing formula. We call these reactions 'predictable' and the missing formula is then predicted.

If the formula predicted does not balance all 'predictables' reactions containing it, the formula and these 'predictable' reactions need to be checked manually. Otherwise, the formula is tagged as correct. As it goes along, the set of balanced reactions and correct formula increase, allowing the prediction of new formulas. The algorithm keeps going until all metabolites have a formula or no

more formulae can be predicted without error.

At this point, the algorithm outputs (1) all metabolites (A) without a formula, or (B) with uncertainly predicted formulas (the one which finally unbalance other 'predictable' reactions), and (2) any (A) unbalanced reactions or (B) 'predictable' reactions which failed. Afterward, curators can focus on the metabolites that were outputted by the algorithm. Chemical formulas need to be manually added for these metabolites (File4 in Figure 25). This process can also lead to the discovery of new duplicated metabolites (same compound, different identifier) (File1). If curators look at the list of reactions outputted by the algorithm, they determine if a reaction is not balanced because a metabolite is missing rather than because the predicted formula is wrong. In this case, they can suggest a modified reaction (File2). Or they can decide that the reaction should be deleted[§] (File3).

Once the curator has updated the four files, metabolite identifiers are mapped (File1), reactions are modified (File2), reactions are deleted (File3), duplicated reactions are removed, formulas are added[**] (File4), and the formulas prediction process starts again. After a while (several days to several weeks), all metabolites have a chemical formulas and all reactions become mass-balanced.

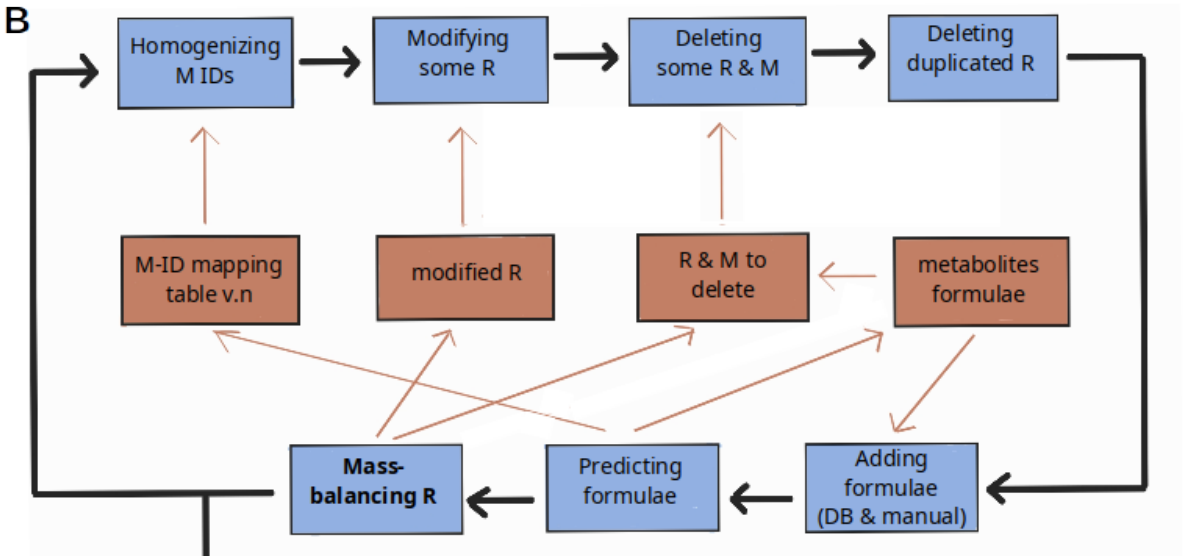


Figure 25: Semi-automatic curation loop in order to mass-balance all reactions. R stands for reaction and M for metabolite.

The curator has access to four files (brown). From left to right : (file1) the namespace ; (file2) modified reactions ; (file3) reactions and metabolites to delete ; (file4) formulae.

These files are utilised by various algorithms (blue) to modify the network in multiple ways. Starting from the top left, then clockwise : update the identifiers assigned to metabolites, apply the specified modifications to the reactions, removes the specified reactions and metabolites, identify any duplicated reactions in the network and eliminate them, update the chemical formulas of metabolites, predict the chemical formulas. Finally, ensure that the total mass of elements is conserved in each reaction.

This phase of the reconstruction process proved to be exceptionally demanding. The most time-consuming aspect involved manual verification and retrieval of missing formulas, as well as

[§]*Metabolite formula prediction algorithm identifies reactions whose manual reassessment brings network maintenance one step closer. I always looked if these reactions were associated with some information (associated gene, link to biochemistry or genetics databases). When the reaction had no associated metadata, no associated genes either or genes found only in Arabidopsis (multicellular terrestrial plant), the reaction is most often deleted.*

[**]*During the mass-balancing process, 7 'dummy' metabolites were added. When considering very complex molecules like starch (which is a polyoside, composed of chains of n D-glucose molecules), the length of the chain can vary according to the models. Thus, in PhotoEukStein, two different starch molecules are considered.*

identification of duplicate metabolites. Although I managed to predict 672 formulas, at least the same number of internet searches had to be done manually on the different databases throughout this whole step. It is worth noting that I have collected 24945 metabolite identifiers with their respective chemical formulas, in order to avoid this tedious work next time (see the resource at the link above).

The search for duplicate metabolites posed a persistent challenge, as they were discovered at various points throughout the whole PhotoEukStein reconstruction process, and there may still be some remaining. To identify potential new duplicate metabolites, all metabolites with the same formula were retrieved for examination to determine if they were either isomers (possessing the same number of atoms of each element but differing spatial arrangements) or genuine duplicates. While I am uncertain of the precise number of manually identified duplicated metabolites, a total of 2870 duplicates were ultimately recognised within PhotoEukStein. It is important to note that whenever metabolite identifiers are modified, the algorithm responsible for identifying duplicated reactions must be restarted (as illustrated in Figure 27). Moreover, 123 reactions were modified, 160 unknown metabolites and 250 reactions (either not found in the database or deemed fictitious entities) were deleted during this process.

2.1.2 ...to a constraint-based generic metabolic model

While our network is now quite clean, it is not yet appropriate to refer to PhotoEukStein as a CBM *per se*. These next steps describe how to make PhotoEukStein ready for constraint-based analysis, as per the protocol of Thiele and Palsson[109]. Constraint-based methods are mathematical approaches for analysing the fluxes through a metabolic network (1.3.3). Thereby it makes possible to predict the growth rate of an organism, or the rate of production of a biotechnologically or ecologically important metabolite, or even infer the metabolic dependencies of small communities[8,103].

2.1.2.1 Biomass objective function

As first approach, I tried to generate a generic biomass objective function (BOF) (1.3.3.2) of phototrophic microeukaryotes. However, the formulation of a BOF is usally dependent on knowing the composition of the cell and energetic requirements necessary to generate biomass content from metabolic precursors, therefore an accurate formulation need experimental data[111,119,151–155]. One may estimate the relative fraction of each precursor from genomes[30] (e.g., by using the Comprehensive Microbial Resource database[156]). But knowing that about a third of the diversity of eukaryotic plankton remains a black box[157]; not to mention the ~60% of the eukaryotic gene catalogue whose function is unknown[62] ; organisms that are uncultured and for which no literature is yet available ; it would have been ambitious to continue trying to estimate a generic composition of eukaryotic marine plankton (especially since GSMs for these organisms did not yet exist). How to consider, or not consider, the N:P Redfield ratio variation within eukaryotic algae[121]? Does a growth objective correspond to that of organisms embedded in complex environments[103]?

In the end, I thought it was better to use the BOFs already present in the reference models. PhotoEukStein sticks to a set of 15 biomass reactions obtained from a range of sources and reflect different types of metabolism. These reactions mainly include autotrophic biomass reactions, from *Chlamydomonas reinhardtii* iRC1080[141], *Chlorella variabilis*[143] and *Phaeodactylum tricornutum* iLB1034[145],

as well as specific reactions for biomass production during light or dark periods, and for various metabolites such as DNA, RNA, lipids, and carbohydrates.

2.1.2.2 Charge balance reactions and protonation

A reaction is charge-balanced if the charge counts are the same on the left- and right-side of the reaction. The charge of molecules depends of the pH, and the pH of organelles in eukaryotic algae may be different. For example, the luminal pH (inside a thylakoid Figure 31) has been estimated at 5.8–6.5 under normal light conditions, and 4.5–4.8 under high light conditions[158,159]. Therefore rather acidic. While the pH of the matrix (inside a mitochondrion) pH values range from 7.2 to 8.2 in different cell types[160,161]. Therefore rather alkalin. Adjusting metabolites to a particular pH may change their charged formulae and thus may require correction of the reactions network.

In modelling, the creation of compartments within the cell makes it possible to dissociate a metabolite into several entities according to the compartment within the cell, and to give each the appropriate charge formula (E.g. « CO_2 » in Figure 31). However, PhotoEukStein is a soup model[78], containing over 9000 reactions of 16 species in only one compartment (except 48 reactions, see 2.2.2 and Figure 31). Thus modifying the protonation of a metabolite can balance one reaction in mass and charge, but can also unbalance other reactions that should take place in other hypothetical compartments. It is a daunting task while considering a pangenome-scale metabolic network as PhotoEukStein. In the end, all the 9162 enzymatic reactions of PhotoEukStein are mass balance. 90% of them are also charge balanced, 1 % have missing charge, the remainings are not charge-balanced. This can surely be improved by separating reactions into specific compartments in further version of PhotoEukStein.

2.1.2.3 Directionality of reactions

The directionality of a reaction is important in constraint-based metabolic models because it determines whether the reaction can proceed in a forward or reverse direction (1.3.3.5). In other words, it determines whether the reaction can produce or consume a particular metabolite (Figure 15). For example, if a reaction is irreversible in the forward direction, then the flux through that reaction must be non-negative (i.e., it can only proceed in the forward direction) (Figure 18). The directionality of reactions impacts the set of allowable flux distributions (Figure 19) in the network and may affect the feasibility and optimality of metabolic phenotypes (Figure 20).

Most of input « models » used to reconstruct PhotoEukStein come from BiGG[130] and BioCyc database[140] relying on MetaCyc database[162]. The Gibbs free energy (‡) of reactions in theses databases has been checked and manually curated by experts in the field, based on a combination of experimental data[141,143–145], available literature, and bioinformatics approaches. It is then assumed that most of the directionality of the PhotoEukStein reactions has already been addressed. However, it's important to note that the Gibbs free energy values for some reactions may not be accurate due to limitations in the available thermodynamic data, and in these cases, additional experimental or computational validation may be necessary.

2.1.2.4 Heuristics constraints

Nevertheless, heuristic rules of thumb are applied to prevent the generation of ATP by futile cycles or false proton gradients. Indeed, ATP is a molecule that serves as the primary energy currency of cells (Figure 5 ; 2.2.2). We make sure that only those reactions that are known to produce ATP are allowed for ATP synthesis (Table 2), whereas all other reactions are set irreversible (can only consume it).

Also, reaction involving quinones are generally irreversible[109]. Quinones are a class of organic compounds that contain two carbonyl functional groups, typically in a cyclic six-membered ring. They are widely distributed in nature and play important roles in biological processes, such as electron transport in photosynthesis and respiration. Examples of quinones include plastoquinone, which is an important electron carrier in transport chain of photosynthesis in the thylakoid membranes of plants and algae (Figure 29 ; Figure 31).

2.1.2.5 Blocked, Sink and Demand reactions

In order to maintain the stationary state of the network, all inner metabolites consumed have to be produced and *vice versa (1.3.3.4)*. If it is not the case, these metabolites are called orphan (H, I and J in Figure 27) or dead-end metabolites (G). The associated reactions can not carry any flux in any simulation conditions because they lack a pathway for the uptake/anabolism or secretion/catabolism of the orphan metabolites. In other words, they do not participate in any optimisation solution ; they are blocked. Some blocked reactions can be reactivated, however it is advisable to delete the others, because they can give false-negative analysis regarding gene deletion on flux redistribution[163]. By including a demand/sink reaction (DM, SK) for a particular dead-end metabolite, one can turn otherwise blocked reactions into active reactions (can carry flux). Demand reactions (DM) are unbalanced network reactions that allow the accumulation of a compound (blue arrow in Figure 26). In PhotoEukStein, 1033 DM were added for metabolites produced but never consumed.

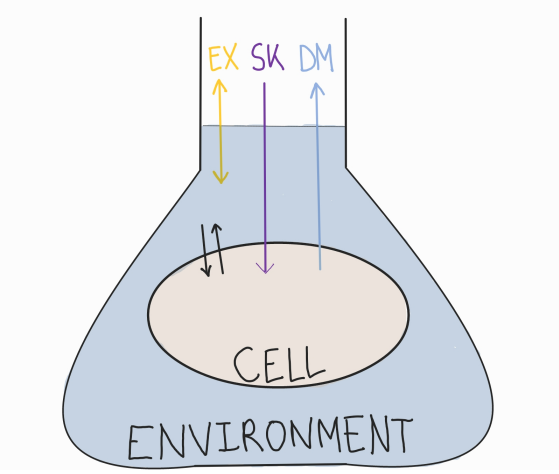


Figure 26: Definition of systems boundaries.

Exchange reactions (yellow arrow) define the medium/ environment. They are coupled with transport reactions (black arrows).

Demand reactions (blue arrow) (DM) are unbalanced network reactions that allow the accumulation of a compound (e.g. DMSP).

Sink reactions (purple arrow) (SK) are similar to DM but provide the network with metabolites. Adding too many SK may enable the model to grow without any resources in the medium/environment (Figure 30B as example).

Sink reactions (SK) are similar to demand reactions but provide the network with metabolites (purple arrow). 674 SKs were added for metabolites consumed but never produced (with hard-constraint on the uptake flux ($v_{lb}=-0{,}5$ and $v_{ub}=0$). Adding too many SKs may enable the model to grow without any resources in the medium/environment (Figure 26 for theory, and Figure 30.B for practice).

Although it is advisable to add SK and DM reactions temporarily, for debugging and network evaluation processes only[109], I consider they are biologically meaningful. They allow the inclusion of

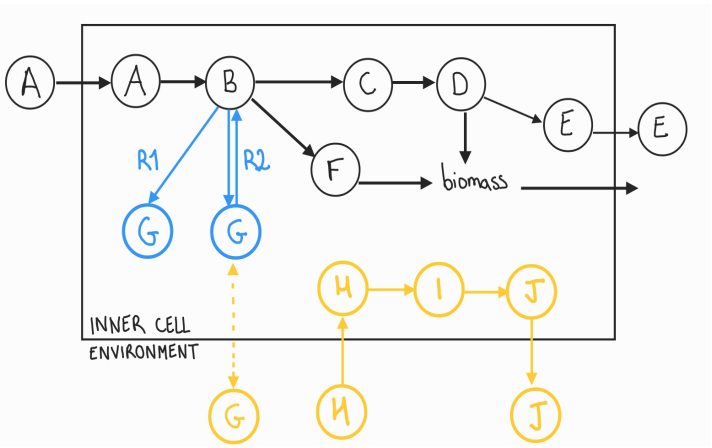
compounds in the metabolic network that are suspected to be anabolisable or catabolizable by the cornerstones of PhotoEukStein (the reference models used :Figure 24). It is possible that their metabolic function has not yet been elucidated, and the underlying reactions, enzymes, and genes are yet to be discovered.

Perhaps these compounds are graphically a leaf of the metabolic network because they have a function in intracellular storage, or in the phycosphere, or perhaps they play a key role in symbiotic interactions. For example, DMSP play a critical role in climate regulation and impact the entire marine food chain. Identification of DMSP-transporter enzymes is overdue for eukaryotic phytoplankton[164], and none of the models used to reconstruct PhotoEukStein (Figure 24) incorporate them. Therefore, there is a DM for DMSP in PhotoEukStein, allowing the study of its production and secretion rates (see example in 4.2.2.1).

On the other hand, when it is impossible to block SKs in order to maintain growth, it surely indicates specific needs of the organism, like a key metabolic pathway or the highlighting of a possible mandatory symbiosis.

When we remove all the 674 SK and 1033 DM reactions, 2,554 enzymatic reactions are disabled (thus 4,261 blocked reactions / 11,229 total reactions). Yet, we would have activated even more reactions if my orphan metabolite detection technique was based on constraints rather than topology. In a directed graph, a node that only has incoming edges is called an "in-degree" node (Figure 27, G in reaction R1), and a node that only has outgoing edges is called an "out-degree" node. The C metabolite has an arrow towards it (producible) and away from it (producible). With this logic, a metabolite that has both an arrow towards it and an arrow away from it is therefore producible and consumable (like metabolite C). However, this consideration is in fact not always true. Consequently, this topological approach did not unblocked all the desirable dead-end reactions. For example in Figure 27, I considered that the metabolite G was not orphan metabolite because the R2 reaction is reversible and therefore G seemed producible and consumable topologically. Regardless R1, R2 is indeed blocked.

Figure 27: Suggestions for improving the detection of duplicated reactions and orphan metabolites.



1. Duplicated reactions (blue rarrows) are those that propose similar metabolic mechanistic transformation. The reaction R2 is reversible and therefore includes 2 reactions: the direct reaction and the reverse reaction. The forward direction is equivalent to the irreversible R1. In this case, only the R2 reaction is kept, and R1 is removed. The genes associated with R1 are recovered and are associated with R2 (2.1.3). Without the addition of a SK and a DM (dotted yellow arrow), the R2 reaction is blocked.

2. Orphan metabolites were not perfectly identified. R2 is one reversible reaction. The metabolite G was considered producible and consumable while the R2 reaction cannot be active without the addition of a SK and a DM (yellow dotted line).

3. Adding SKs and DMs does not activate all inactive reactions. H, I and J metabolites are disconnected from the main network.

By performing an FVA with a low optimum fraction it is possible to identify blocked reactions and dead-end metabolites. For all orphan metabolites, we could have added a reversible exchange reaction (with a recognition tag), which we will call here $EX_{unblocked}$. Then, after running a second

FVA, we consider the flux interval, with v_{lb} the lower bound of the considering $EX_{unblocked}$, and v_{ub} its upper bound :

$$if\ v_{lb} < 0 \quad \text{and} \quad v_{ub} \leq 0,$$

the metabolite cannot be produced and requires a SK ;

$$if\ v_{lb} \geq 0 \quad \text{and} \quad v_{ub} > 0,$$

the metabolite cannot be consumed and requires a DM ;

$$if\ v_{lb} = 0 \quad \text{and} \quad v_{ub} = 0,$$

the addition of SK or DM will neither consume nor produce the metabolite. As the H and J metabolites belong in fact to a subnetwork disconnected from the main network. Do nothing ;

$$if\ v_{lb} < 0 \quad \text{and} \quad v_{ub} > 0,$$

the metabolite can neither be produced nor consumed and requires both a SK and a DM, or a reversible EX (similar to the yellow dotted arrow).

2.1.3 PhotoEukStein-associated data

2.1.3.1 Logical conjunction of genes to ensure enzymatic reactions

In its initial version, PhotoEukStein encompass 5,831 metabolites and 11,229 reactions. Two types of reactions are distinguished : 2067 boundary reactions (including 360 exchanges reactions, 674 sink reactions, 1033 demand reactions), and 9162 internal biochemical transformations. The number of sink and demand reactions may be reduced in future versions of PhotoEukStein as new enzymatic reactions and/or associated genes are discovered. For each internal reaction in the curated universal model, we identify all those that are equivalent in the input models (i.e., duplicates Figure 27) to recover the maximum number of logical gene conjunctions (and their identifiers) (Figure 8).

For information, detecting duplicated reactions within photoeukstein is rather easy since everything is formatted in the same way. However, when it comes to comparing PhotoEukStein reactions with some reference models, it is necessary to reuse the formatting algorithms as explained at the beginnig of section 2.1.1.1.

Thus, 7,599 PhotoEukStein reactions (/9162) are associated to 20,468 protein sequences, from reference genomes[††], by their respective logical associations (Figure 14).

2.1.3.2 Anabolism of DMSP in eukaryotic algae

DMSP synthesis from methionine (Met) has been shown to take place via three pathways in various organisms[164] : a transamination pathway in some marine bacteria and algae[19,20,165,166] (Figure 28 right), a Met methylation pathway in angiosperms and bacteria[167,168] (Figure 28 left), and a decarboxylation pathway in the dinoflagellate *Crypthecodinium,* which is still not well described[169]. The transamination

[††] mostly retrieved from NCBI, UniProt, Diatomics and TAIR (*Arabidopsis thaliana* database), or added by hand because of defective identifiers (about 100 sequences).

pathway consists of four reactions (Figure 28 from bottom to top right) :

(1) The Met aminotransferase (MAT) activity, yielding 4-methylthio-2-oxobutyrate (MTOB) from Met,

(2) the MTOB reductase (MR) activity, yielding 4-methylthio-2-hydro-oxybutyrate (MTHB) from MTOB,

(3) the methylation of MTHB to 4-dimethylsulfonio-2-hydroxybutyrate (DMSHB) catalysed by the MTHB methyltransferase (MHM),

(4) Finally, the DMSP production from DMSHB via DMSHB decarboxylase.

MHM (step 3) is the key enzyme of the Met transamination synthesis pathway. Indeed, it has been shown that MAT (step 1) and MR (step 2) enzyme activities exist, although at reduced levels, in non-DMSP-producing algae, whereas MHM activity is specific to DMSP producers[170]. Moreover, MHM would be the rate-limiting and committing step in the transamination DMSP synthesis pathway, thus its activity and DMSP production are correlated.

As enzyme associated to MHM, DsyB protein was first identified in marine Alphaproteobacteria[19]. Today, two enzymes encoding for MHM in eukaryotic algae have been identified : (i) DSYB gene encoding DSYB enzyme is a eukaryotic homologue of DsyB[20] , and (ii) TpMT2 whose the function was confirmed in *T. pseudonana*[21].

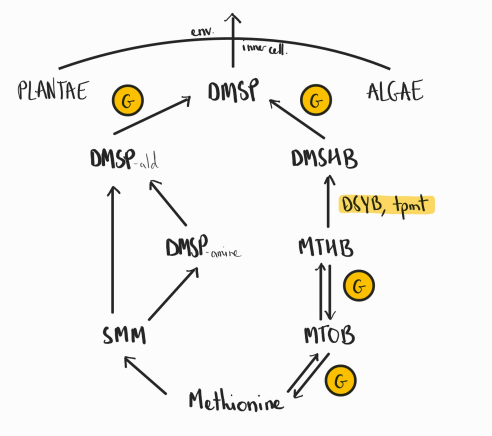


Figure 28: Anabolism pathway of DMSP in phototrophic eukaryotes. Yellow circles indicate the presence of protein sequences associated with the reactions.

More recently, a third enzyme DSYE with MTHB S-methyltransferase activity, would have been identified in diverse and environmentally abundant *Chlorophyta, Chlorachniophyta, Ochraphyta, Haptophyta* and *Bacillariophyta* algae[171] (including *Pelagomonas calceolata*, amongst the most abundant eukaryotic species in the ocean[172]). Although some of the models that make up PhotoEukStein had the DMSP synthesis pathway (*Thalassiosira*[140], *Okamuranus*[146], ou *Phaeodactylum*[145]), none had a gene associated with MHM. This has had some consequences that will be discussed in the section 4.2.2.1.

We added 135 sequences for DSYB, and 6 for TpMT2 (from [19,21,173]) in the protein sequences database of PhotoEukStein. This also shows that it is rather easy to add information to PhotoEukStein.

2.1.3.3 Metadata of enzymatic reactions

Using KofamKOALA[174] , I annotated the 20,468 protein sequences from PhotoEukStein. Thus, with the logical conjunctions of genes that ensure the existence of an enzymatic reaction, I added metadata to 7,599 reactions. It indicates in which metabolic pathways (according to the KeGG maps [136]) the reaction occurs. The different metabolic pathways present in PhotoEukStein are indicated in the Table 4. This simply gives some clues about the function of the reaction.

2.2 PHOTOEUKSTEIN'S VALIDATION AND REFINEMENT LOOP

"All models are wrong, but some are useful", Box[175]

2.2.1 A hint of epistemology

This part is much more discussed in this one 4.1.

2.2.1.1 Daisyworld and DMSP

Some models never will be (in practice) able to compare predictions with empirical data because the future is too distant[67]. Other one, because it will never have a proxy for the phenomenon of interest[11,176]. The Gaia theory, proposed by James Lovelock and Lynn Margulis in the 1970s, suggests that the Earth is a self-regulating system that maintains conditions that are favorable for life[3]. According to the theory, the physical and biological components of the Earth, including the atmosphere, oceans, and living organisms, interact to form a complex, interconnected system that regulates the environment (1.1.2). The Gaia theory proposes that life and its environment are in a constant state of feedback, with living organisms altering the environment and the environment shaping the evolution of life.

The model which tried to compute the phenomenon has been criticised for beeing too abstract[176]. Indeed, the Daisyworld model is a simplified theoretical model proposed by James Lovelock and Andrew Watson in 1983 to explore the concept of Gaia theory[11]. The model describes a hypothetical world inhabited by two types of daisies, black and white, which have different albedo, or reflectivity, and affect the temperature of the planet. The daisies grow and reproduce based on the temperature of their environment, creating a feedback loop that influences the planet's climate. In the model, if the temperature is too low, the black daisies are favoured, as they absorb more solar radiation and warm up the planet, allowing more white daisies to grow. If the temperature is too high, the white daisies are favoured, as they reflect more solar radiation and cool down the planet, allowing more black daisies to grow. The model shows how the interactions between the daisies and the planet's climate can lead to self-regulation of the planet's temperature.

Daisyworld is not intended to be a precise representation of Earth's climate. It is a simplified, fictive and abstract model that omits many important factors that contribute to the regulation of planetary climate (atmospheric composition, ocean currents, geological processes, trophic interactions...). However, it is still a useful tool for exploring the potential consequences of different feedback mechanisms and for developing a better understanding of how complex systems interact.

It is important to note that models are not perfect representations of reality. They are simplifications or abstractions of reality. They are built to represent a facet of a particular system or phenomenon and are based on assumptions and approximations, which can introduce errors or uncertainties. Their limitations should be acknowledged[177]. However, models can be extremely useful tools for making predictions, testing hypotheses, and gaining insights into complex systems.

2.2.1.2 Strategies for PhotoEukStein's validation

In an epistemological context, model validation and sensitivity analyses are critical steps to ensure the robustness and reliability of the model's predictions. The validation process generally consists of comparing the model's predictions with observed data, the literature or other reliable models, and thus assessing the model's ability to simulate known phenomena. Then, one can use this model to test new hypothesis and predict future outcomes for which one does not yet have empirical values.

These validation steps of PhotoEukStein required a lot of back and forth between hypothesis testing, new ideas, adjustment, and refinement. I would even qualify them as endless since it is based on the prediction of phenotypes whose number could be infinite by definition[178,179]. In this manuscript, we primarily characterize the metabolic function as a flux space projected on some axes defined by key reactions. It provides a unique way to assess continuous biological phenotypes *per se* as it differs from the sole identification of functional genes, and show model plasticity for specific functions evaluated. The mechanisms underlying a phenotype are as complex as the system we are studying. Therefore, exploring and evaluating behaviours of metabolic models requires strategies, time and perseverance. PhotoEukStein's reliability is demonstrated in three steps :

(1) When a generic model is converted to ready-to-use organism-specific models using CarveMe, the whole manual curation and relevant structural properties are preserved[133]. Therefore, we ensure that PhotoEukStein can grow under photoautotrophic conditions with adapted physiological strategies like the ability to fix inorganic carbon. We expected a coupling between light uptake and CO_2 uptake from the environnement, and the underlying synchronisation of photosystem reactions, ATP production by the chloroplastic ATP synthase, as well as inorganic carbon assimilation by the ribulose-1,5-biphosphate carboxase/oxygenase (RuBisCo), the key enzyme of the Calvin cycle (2.2.2).

(2) Once these mechanisms were established, we derived metabolic models from PhotoEukStein for 3 eukaryotic algae (*Phaeodactylum tricornutum, Thalassiosira pseudonana, Chlorella variabilis*) and compared these PhotoEukStein-derived models to their respective manually-curated metabolic models (2.2.3.1). Only three species are taken for comparison because A) they are the only available reference models ready-to-use for constraint-based analysis, and B) the predictions of these models have been validated by culture experiments[143,145,180]. In our sake of validation, we compared predicted growth rates of both princeps and PhotoEukStein-derived models across under 10^4 photoautotrophic environmental conditions.

(3) To further scrutinise the internal consistency of PhotoEukStein-derived GSMs, we compared the distribution of reaction fluxes as predicted by both models for *Phaeodactylum tricornutum (2.2.3.2).* We considered inter-reactions fluxes correlations within each model when sampling the whole metabolic space with 10^4 iterations.

2.2.2 Photoautotrophic phenotypes of PhotoEukStein

In this step, it should be tested if basic capabilities of photoautotrophic organisms can be reproduced by the model PhotoEukStein. This first version of PhotoEukStein focuses on oxygenic photosynthesis and its ability to growth on photoautrophic conditions.

The ultimate goal of photoautotrophic organisms is to use light energy to convert water and carbon dioxide into oxygen and energy-rich organic molecules such as glucose. It occurs in several

steps, which can be broadly categorised into two main stages : the light-dependent reactions (photosynthetic apparatus) and the light-independent reactions (also known as the Calvin cycle). Even though the Calvin cycle occurs during the light period, it is still considered light-independent because it does not require direct energy from light to proceed. Rather, it relies on the ATP (adenosine triphosphate) and NADPH (Nicotinamide adenine dinucleotide phosphate) molecules that are produced by the light-dependent reactions. These whole reactions take place in the chloroplasts of phototrophic eukaryotes.

2.2.2.1 *Photosynthetic apparatus and chemical energy production*

Physiological context

The photosynthetic apparatus is a highly organised structure in the thylakoid membrane of the chloroplasts that facilitates the transfer of electrons and protons in response to light stimulation. Light energy is captured by pigments like chlorophyll. The absorption of light by photosystem II (PSII) excites electrons in chlorophyll molecules, which are then passed through a series of electron carriers (electron transport chain), ultimately resulting in the production of NADPH (Figure 29).

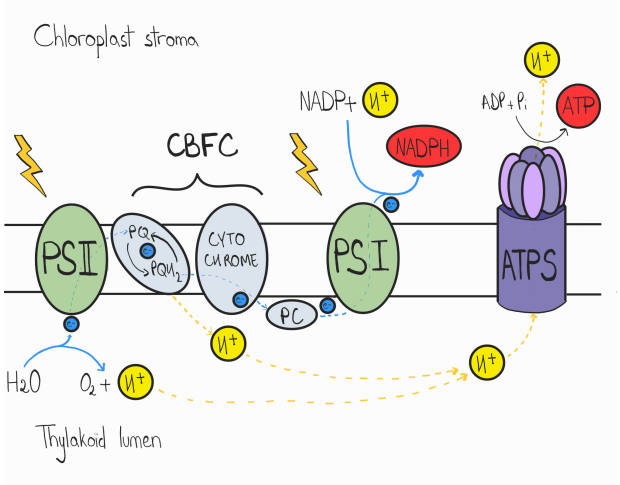


Figure 29: photosynthetic apparatus (PA) coupled with ATP synthase (ATPS), converts light energy into chemical energy. PA facilitates the transfer of electrons and protons in response to light stimulation, ultimately resulting in the production of NADPH. The protons released create a proton gradient across the thylakoid membrane, which drives the synthesis of ATP by ATPS. Plastoquinone (PQ/PQH$_2$) is an important electron carrier in transport chain of photosynthesis.

During the light-dependent reactions of photosynthesis, water is split into oxygen, protons (H+), and electrons through a process known as photolysis or water splitting. The electrons produced by the splitting of water are used to replace those lost by the photosystems as they reduce the primary electron acceptors NADP. Additionally, the protons released create a proton gradient across the thylakoid membrane, which in turn drives the synthesis of ATP by the chloroplastic ATP synthase (ATPS) through the process of chemiosmosis. In other words, the main role of the photosynthetic electron transport chain is to convert light energy to chemical energy in the form of ATP and NADPH.

At this stage of the reconstruction process, PhotoEukStein had only one 'cell' compartment (without organelles) (2.1.1.1). However, it is worth to note that biological membranes serve many purposes. One is to control the fluxes of solute between compartments within cells and between cells. A second is to separate electrical charges across the membrane. Finally, membranes facilitate spatial organisation of chemical reactions. In the context of the photosynthetic apparatus, specific products of biochemical reactions accumulate on only one side of a thylakoid membrane (*e.g.* proton), thereby forming concentration gradient accross the membrane. The translocation of ions and electrons helps establish an eletrical field. Protons are then transported from one side of a membrane to other and produce ATP. Thus, a thylakoid compartment was added to PhotoEukStein to synchronise the reactions of the synthetic apparatus (based on iLB1034[145]) and ATPS (Figure 31). Despite the great diversity of aquatic photosynthetic organisms, most of the molecular structures and functions that are essential for photosynthesis are highly conserved.

Metabolic niches to explore photoautotrophic phenotypes

A photoautotrophic medium containing inorganic sources of sulphur, nitrogen, carbon, phosphate, but also iron, magnesium, and photon (*n.b*. is considered as metabolite in CBMs) is designed. Concerning the exchange reactions, only the uptake of these « nutrients » is allowed to test the phototrophic phenotype of PhotoEukStein. In the first simulation (Figure 30), sink and demand reactions can be used by PhotoEukStein to maintain as many active reactions as possible (2.1.2.5). Then, we observe the relationships between the photon uptake (« EX_photon_e »), and the reactions of the photosynthetic apparatus (« PSII_u » and «PSI_u ») (Figure 30.A) ; the relation with ATP production by ATPS, and the growth rate of PhotoEukStein (Figure 30.B). The more photons enter the system, the more the photosystems are stimulated with a synchronisation of the two PS (Figure 30.A). We also see that the ATP production rate is coupled to the photosynthetic activity and fuels the growth reaction (B). Interestingly, even when the photosynthetic apparatus is off ($v_{PSII_u}=0$ as proxy in B), the growth rate is at 15.15 mol.gDW^{-1}.h^{-1}, meaning that PhotoEukStein can grow without light. The SK are the cause of this growth by feeding the network with organic molecules, and not allowing fully photoautrophic conditions (2.1.2.5). To test the metabolic adaptation strategy of PhotoEukStien under purely autotrophic conditions, access to any other carbon molecules other than CO_2 is not allowed. Thus, SK reactions are blocked for the next simulations (Figure 32 ; Figure 33 ; Figure 34).

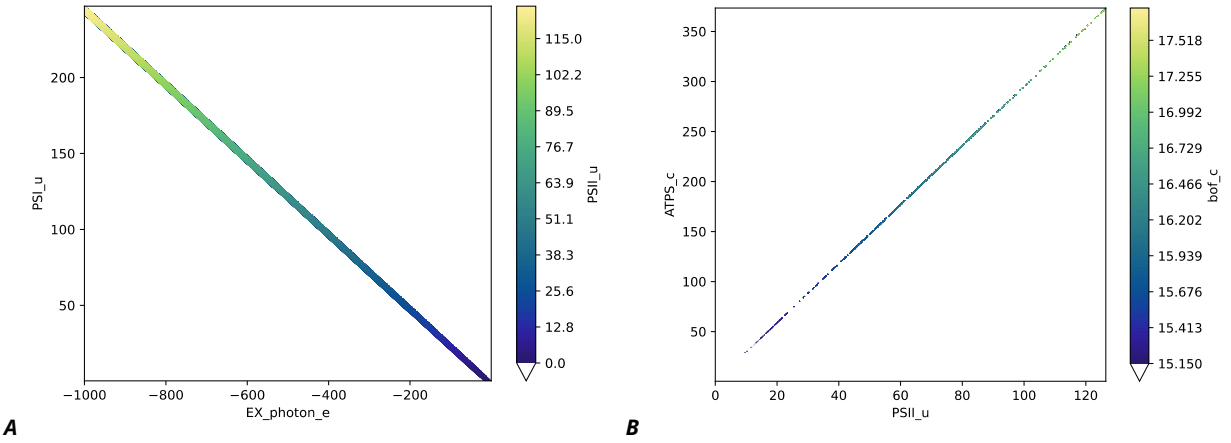


Figure 30: Photoautrophic phenotypes of PhotoEukStein. As a reminder, for exchange reaction (EX), a negative flux (mol.gDW-1.h-1) means that the system uptakes from the environment.

A) Relationship between the photon uptake (EX_photon_e), PSII and PSI fluxes (photosynthetic apparatus) ;

B) Relationship between photosynthetic apparatus stimulation (PSII as proxy), the ATP production (ATPS) and growth rate (bof : biomass objective function).

2.2.2.2 Autotrophy and inorganic carbon assimilation

Physiological context

ATP is a molecule that serves as the primary energy currency of cells. ATP is composed of an adenine base, a ribose sugar, and three phosphate groups. The energy stored in the chemical bonds between the phosphate groups is used by cells to power various cellular processes. When ATP is hydrolysed (broken down), it releases energy. ATP is constantly being regenerated in cells through processes like cellular respiration and photosynthesis (Figure 5). The ATP produced from the light-

dependant reactions are used for different purposes fuelling biosynthetic processes, such as the polymerisation reactions implicated in the synthesis of macromolecules (synthesis of amino acids, nucleotides, lipids...), or the translocation of many ions and solutes through membranes.

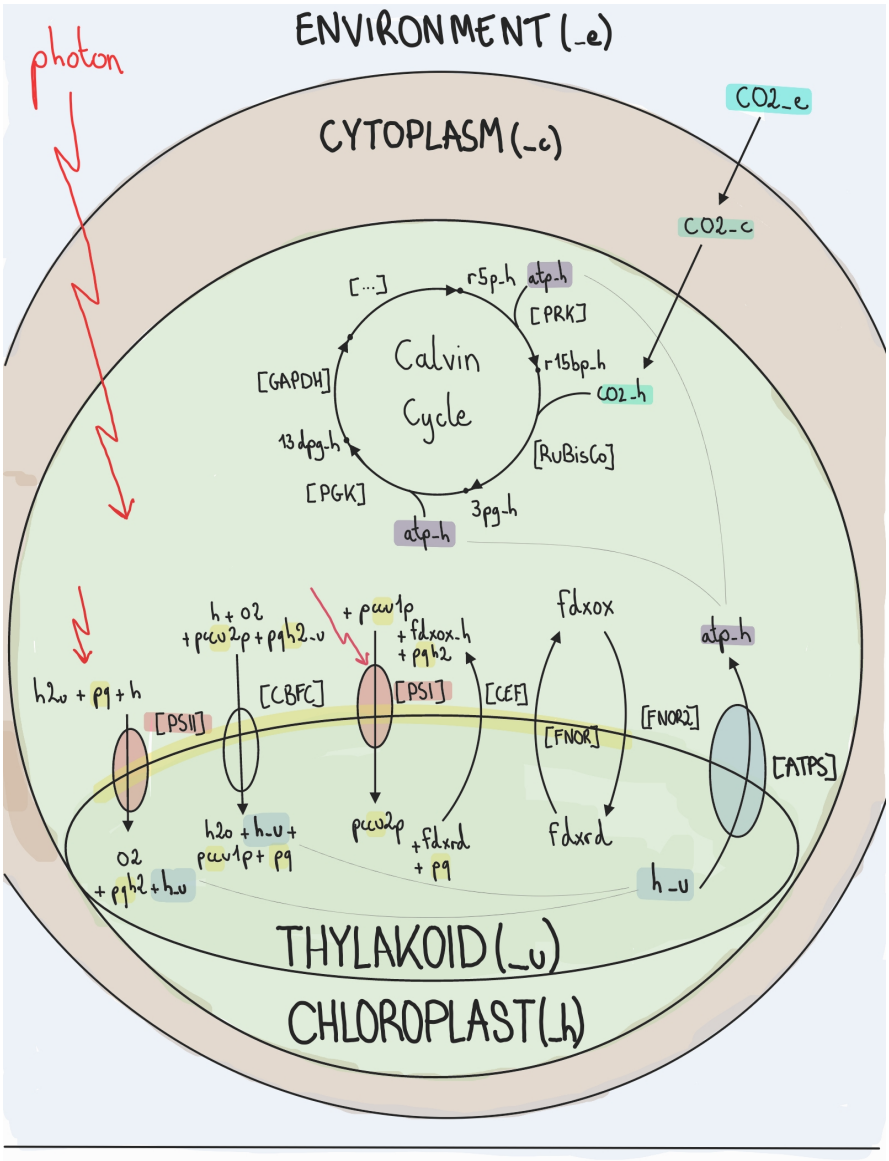


Figure 31: The compartments allow a spatial organisation of certain molecules and thus allow the emergence of specific functions in PhotoEukStein. For example, photosynthesis has been set up thanks to the compartments (schematic representation).

PSII and PSI are couple to photon absorption (red). The photosynthetic apparatus is coupled thanks to specific molecules in strategic compartments (yellow). The ATPS is coupled to the photosynthetic apparatus by proton gradient (blue) through the thylakoid membrane. Calvin cycle is coupled to ATPS by using specific ATP molecules from ATPS (purple). Finally the Calvin Cycle fixes environmental CO2 dependently to photon absorption and fuel the biomass reaction.

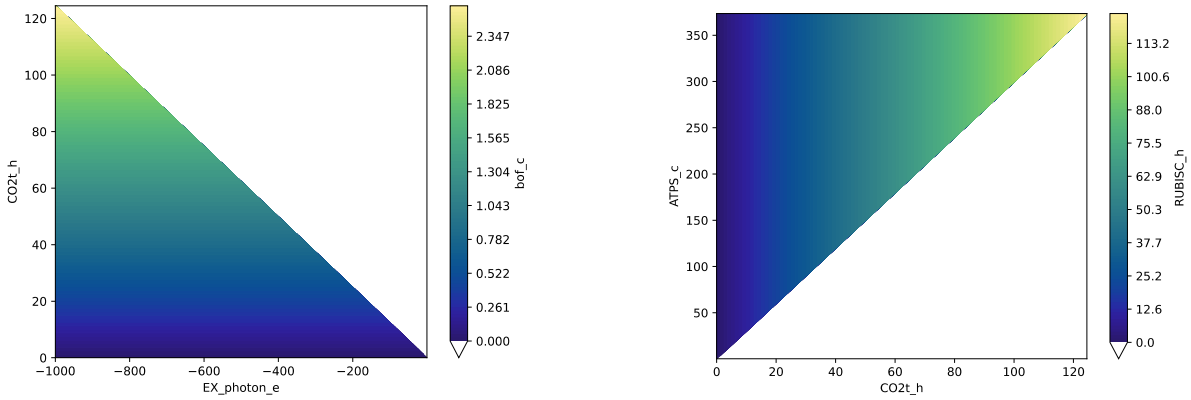
CO2 _e, CO2 _c, CO2 _h are three different entities that all three represent the CO2 molecule in different compartments of the system (environment, cytoplasm, chloroplast). In this case, the three identifiers are associated with the same formula.

However, for a microalga growing photoautotrophically, more than 60 % of the photosynthetically generated ATP are used to assimilate and reduce inorganic carbon[181]. This reducing-process is known as the Calvin cycle and invariably involves the enzyme RuBisCo (Ribulose-1,5-bisphosphate carboxylase/oxygenase). RuBisCo is a key enzyme which catalyses the carboxylation of ribulose-1,5-

bisphosphate (« r15bp_h » in Figure 31), a five-carbon molecule, with CO_2 to form two molecules of 3-phosphoglycerate (« 3pg_h »), a three-carbon molecule, which is then used to synthesise energy-rich organic carbon molecules such as glucose, starch, sucrose or other organic compounds[‡‡]. Carbonyl groups, such as those found in CO_2, have a double bond between carbon and oxygen. In order for carbon to be assimilated into organic molecules, this double bond needs to be broken and the carbon must be reduced, through consumption of chemical energy coming from ATP hydrolysis. In order to couple the mechanisms that build up this photosynthetic function, the chloroplast compartment has been added and houses the Calvin cycle (Figure 31).

Metabolic niches to explore inorganic carbon assimilation

We observe that the CO_2 uptake flux into the chloroplast correlates well with the action of RuBisCo, which itself depends on the supply of ATP from chloroplast ATP synthase (Figure 32.A). We also see that, when CO_2 is the only carbon source, PhotoEukStein can not growth without light (B). When the maximum biomass is reached, the maximum photosynthesis is also reached.



A ***B***

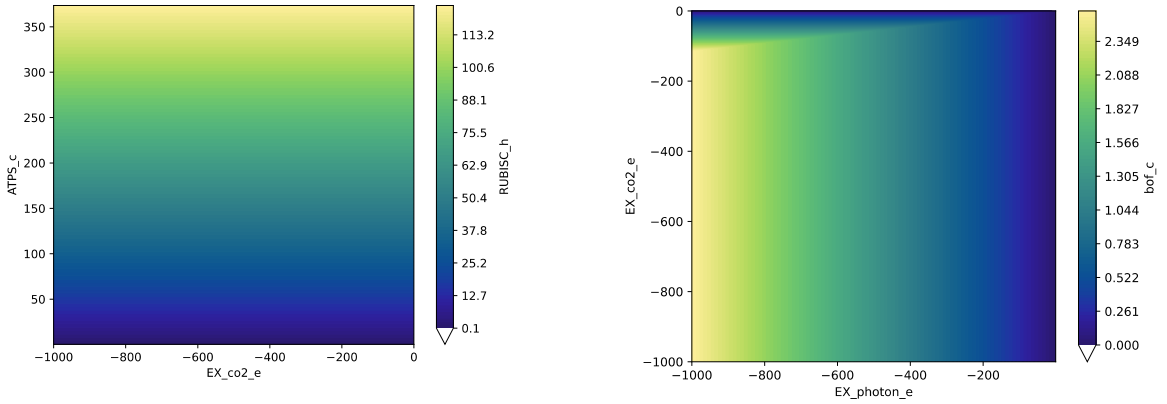
Figure 32: Photoautrophic phenotypes of PhotoEukStein. For exchange reaction (EX), a negative flux (mol.gDW-1.h-1) means that the system uptake from the environment (Definition of reaction fluxes Figure 18). SK are blocked.

A) Relationship between the uptake of photon, the import of CO2 into the chloroplast (CO2t_h) and the growth rate (bof) ;

B) Relationship between CO2 import into the chloroplast (CO2t_h), ATP production (ATPS) and CO2 fixation (RUBISC_h).

However, we also observe that when the maximum growth rate is reached, the CO_2 from the environment (EX_co2_e) keeps entering the system (Figure 33).

[‡‡] *The high-energy molecules (sugars) produced by photosynthesis can then be used as an energy source during periods of darkness or when the demand for energy is high.*



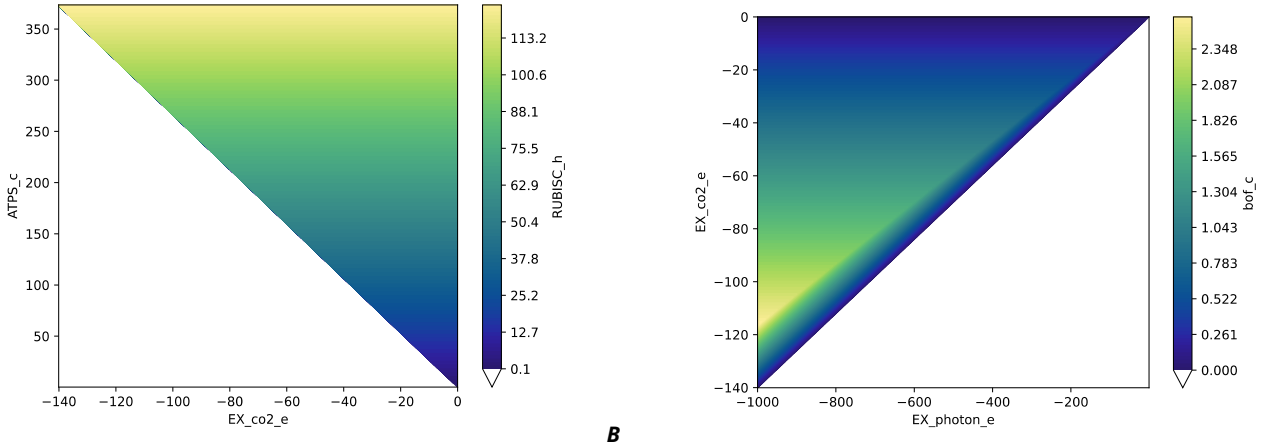
A **B**

Figure 33: Photoautotrophic phenotypes of PhotoEukStein. For exchange reaction (EX), a negative flux (mol.gDW-1.h-1) means that the system uptake from the environment (Definition of reaction fluxes Figure 18). SK are blocked.

A) Relationship between ATP production (ATPS_c), environmental CO2 uptake and CO2 fixation by RUBISCO.

B) Global phenotypes of photosynthesis showing photon and CO2 uptake and growth rate.

When we close the export of all carbonaceous molecules (EX or DM), we see that the system can no longer import excess CO_2 (Figure 34. A and B). However, we see that if the system imports more carbon (v_{lb} < -120), the growth rate drops sharply. Closing the molecule export prevents the regulation of the stoichiometric balance of the different elements (especially carbon, nitrogen, phosphate). This is a phenomenon that will be observed again (4.2.2.1). I imagine that the model must manage this excess of carbon internally, bypassing the needs for growth. Moreover, closing the export of carbon molecules does not seem to me to be biologically relevant (e.g. lost of DMSP production).



A **B**

Figure 34: Photoautotrophic phenotypes of PhotoEukStein. For exchange reaction (EX), a negative flux (mol.gDW-1.h-1) means that the system uptake from the environment (Definition of reaction fluxes Figure 18). SK and all carbonaceous molecules (EX or DM) are blocked.

A) Relationship between ATP production (ATPS_c), environmental CO2 uptake and CO2 fixation by RUBISCO.

B) Global phenotypes of photosynthesis showing photon and CO2 uptake and growth rate.

2.2.2.3 The photoautrophic phenotype of PhotoEukStein : a conclusion

The photosynthetic apparatus of PhotoEukStein is based on the photosynthetic system of iLB1034 model[182]. The reaction fluxes of the photosynthetic apparatus as well as the production of ATP by the chloroplastic ATPS are coupled to the uptake of photon into the system (Figure 30). The RuBisCo

enzyme of the Calvin cycle fixes CO_2 dependently to light stimulation, and supplies the energy and carbon needed for the growth of PhotoEukStein (Figure 32). These flux couplings are possible thanks to the structuring capacity of biological membranes (Figure 31). Intracellular compartments, such as the chloroplast and its thylakoids, give rise to metabolic functions that are only possible through the organisation and spatialisation of specific reactions[183,184].

Nevertheless, as long as there are carbon molecules available for secretion reactions, PhotoEukStein continues to absorb CO_2 (the sole carbon source) even when the photosynthetic system is saturated (Figure 33). Alternatively, we can view it as a continuous carbon influx requiring an equivalent efflux (as long as there is carbon coming in, it must come out). Notably, because the photosynthetic apparatus and carbon fixation by RuBisCo are already saturated when the environmental CO_2 uptake rate is at -120 mol.gDW-1.h-1, importing additional carbon compels PhotoEukStein to employ alternative metabolic strategies beyond the normal regulation of excess import observed in typical photoautotrophs.

It was surprising to conceive of the ability of PhotoEukStein to incorporate CO_2 into complex molecules, considering that this process requires energy. In order to understand this phenomenon, I thoroughly examined all reactions involving ATP and CO_2, imposing the necessary constraints (Table 2 and Table 3). However, the constraints alone were insufficient to restrict this effect. Either PhotoEukStein switches to a "cellular respiration mode," generating energy through glycolysis and the Krebs cycle, which enables the incorporation of CO_2 into complex molecules, or the CO_2 is rapidly expelled through a minimal number of reactions.

Thus, to maintain phototrophic phenotypes, one can limit the uptake of CO_2 ($vlb_{EX_{CO2}} = -120$).

2.2.3 PhotoEukStein-derived models validation

2.2.3.1 Comparing growth rates under similar environmental conditions

To assess the validity of PhotoEukStein-derived GSMs, we reconstructed 3 models (*Chlorella variabilis, Phaeodactylum tricornutum, and Thalassiosira pseudonana*[§§]) in order to compare them with those of expert-based GSMs[143,145,180] (hereinafter referred to as « references »). We extensively sampled the photoautotrophic metabolic niches for the 6 GSMs (i.e., approximately 10^4 randomly generated environmental conditions each), and compared their predicted growth rates (3.2, Supplementary Figure S1, left pannel). In all cases, both predicted growth rates are highly correlated, showing that PhotoEukStein-derived GSMs are as efficient as expert-based models to capture fundamental biological knowledge, and thus correspond to observations made with cultures. Therefore, PhotoEukStein-based GSMs are prone to provide useful biological insights based on integration of gene content of organisms into GSMs, even if no further knowledge but their gene content is available. Moreover, exploring metabolic niches with GSMs allows to assess the metabolic exchange fluxes differentiating growth rates between references and PhotoEukStein-based GSMs. On the right

[§§] *The genomic and biogeochemical information of Thalassiosira pseudonana included in PhotoEukStein comes from the PGDB of BioCyc (2012). The constraint-based model used to make the comparisons comes from a fairly recent publication[185]. Thus, unlike the other GSMs, the reference used to validate the PhotoEukStein-derive model of Thalassiosira is not included in PhotoEukStein (and therefore its BOF is also different). This may explain the greater differences in growth rates between the two models in 3.2, Supplementary Figure S1). However, the R^2 being very high, we can see that the two models adapt to their environment in a rather similar way.*

(Supplementary Figure S1) are indicated the metabolites uptake the most correlated with predicted growth rates differences between reference and PhotoEukStein GSMs. For the three reference species we studied, metabolite exchange fluxes between both models vary, and a given metabolite can favour growth in either reference or PhotoEukStein GSM, depending of the organism. It may therefore indicate missing metabolic reactions leading to the emergence of somewhat different functional strategies.

2.2.3.2 Comparing the coupling of internal reactions fluxes

To further scrutinise the internal consistency of PhotoEukStein-derived GSMs, we compared the reaction fluxes as predicted by both models for *Phaeodactylum tricornutum*. We considered inter-reactions fluxes correlations within each model when sampling the metabolic space. The resulting correlation maps are highly similar in both reference and PhotoEukStein-derived GSM, indicating that both models connect very similarly the various fluxes (3.2, Supplementary Figure S2). We confirmed this visual analysis by plotting compared intra-model correlations values distribution (3.2, Supplementary Figure S3). When pairs or reactions are highly correlated within one GSM, they are as highly correlated within the other GSM, as is visible in the top-right and bottom-left cells, and loosely connected pairs of reactions have similar characteristics in both models (central area). Automatised top-down approach applied to *P. tricornutum* is therefore capturing the same essentials as the expert-based model, and represents the same biological features even when considering distribution of metabolic fluxes within the GSM[†].

2.3 STATE-OF-THE-ART OF PHOTOEUKSTEIN

PhotoEukStein encompasses available biochemical and genomic information of 15 eukaryotic algae and 1 terrestrial plant (Figure 24). Combined with top-down technique[133] (Figure 21), this new generic model enables fully-automatic reconstruction of constraint-based metabolic models (CBMs) at genome-scale (GSMs) for microeukaryotic-algae (3).

PhotoEukStein contains 5831 metabolites and 11229 reactions (2.1). Two types of reactions are distinguished : 2067 boundary reactions (including 360 exchanges reactions, 674 sink reactions, and 1033 demand reactions), and 9162 internal biochemical transformations (Figure 35). Of the 9162 enzymatic reactions of PhotoEukStein, 7599 of them are associated with 20468 protein sequences from reference genomes (2.1.3.1). As for the other 1563 internal reactions of PhotoEukStein that have no associated genes, either they are "spontaneous" (occur without outside influence or intervention), or no genes have been found to catalyse the reactions. A third case would be that the genes are known, but PhotoEukStein's backbone models have not incorporated them.

In PhotoEukStein, 15 biomass objective functions (BOFs) have been incorporated from *Chlamydomonas reinhardtii* iRC1080[141], *Chlorella variabilis*[143], and *Phaeodactylum tricornutum* iLB1034[145]. The included reactions primarily consist of autotrophic biomass reactions, along with specific reactions dedicated to biomass production during light and dark periods. Currently, the GSM biomass reactions represent the molecular content of specific model organisms. It is helpful for bio-

[†] *Some reactions have their reactants reversed and therefore some correlations would be similar if we consider their absolute. E.g. R1 (model 1) : A → B (positive flux) ; R1 (model 2) : B ← A (negative flux) ; R2 (both models) : B → C (positive flux). In Model 1, the flux correlation between R1 and R2 would be positive, while in Model 2 it would be negative. However, the biochemical transformations are similar in both models.*

engineering work, but not really for the modelling of wild-type organisms. As an improvement, one could strip down the biomass reaction so that it only represents the strict minimum for the organism's growth needs. Other molecular contents would then be produced independently. Such a modelling scheme allows considering molecular contents as stock.

PhotoEukStein combines metabolic features of photosynthetic eukaryotes i.e. the photon absorption allows ATP production by ATPS, which fuels both the fixation of CO_2 by the RuBisCo and its integration into organic components essential for growth (2.2.2). The chloroplast (12 metabolites and 44 reactions) and thylakoid (3 metabolites and 4 reactions) intracellular compartments allows this fine synchronisation between these key reactions of photoautotrophic metabolism. Indeed, biological membranes serve many purposes. One is to control the fluxes of solute between compartments within cells and between cells, a second is to facilitate spatial organisation of chemical reactions, and thus carve out new emerging phenotypes.

Until now, our discussion has focused on the carboxylase function of RuBisCO. However, it is important to note that this enzyme also possesses another enzymatic activity known as oxygenation. Both carboxylation and oxygenation take place within the same active site of RuBisCO, resulting in a competitive relationship between the two activities. The dominant activity depends on the relative concentrations of substrates in the immediate vicinity of the enzyme. The oxygenation activity primarily occurs during light exposure when the ratio of O_2 to CO_2 is high around RuBisCO. This process, which involves the uptake of oxygen and release of carbon dioxide, is referred to as photorespiration.

Cellular respiration is another intricate metabolic process comprising of three primary stages : glycolysis, the Krebs cycle, and oxidative phosphorylation. In oxidative phosphorylation, energy is liberated from electrons carried by reduced molecules, such as NADH. These electrons are transported across a sequence of electron transport proteins situated in the inner mitochondrial membrane, generating a transmembrane proton gradient. Ultimately, this proton gradient facilitates the production of ATP by the mitochondrial ATP synthase. However, this ATP synthase necessitates the presence of both mitochondrial and peroxisome compartments to enable the proper functioning and coordination of the reactions involved in cellular respiration. This continuous process occurs throughout both day and night since algae require the energy derived from cellular respiration to sustain their cellular functions even in the absence of light for photosynthesis. While key pathways of respiration, such as glycolysis and the Krebs cycle, are already incorporated into PhotoEukStein, the validation of respiratory metabolism during night condition would enable the modelling of algae based on their circadian clock. Photoautotrophic organisms rely on the utilisation of organic molecules stored during the day in the absence of photosynthesis during the night.

I believe that there is room for further improvement in the primary and secondary metabolism of eukaryotic algae within PhotoEukStein. By delving deeper into specific key reactions, we can ensure that the system's emergent properties align with biological principles. However, despite these limitations, PhotoEukStein demonstrates the ability to replicate expected physiological phenotypes, as evidenced by predicted growth rates and metabolic flux distributions across approximately 10^4 environmental conditions (2.2.3). The genome-scale models derived from PhotoEukStein for various algae species, such as *P. tricornutum, C. variabilis,* and *T. pseudonana* exhibit comparable efficiency to expert-based models in capturing essential biological knowledge[143,145,180].

Moreover, PhotoEukStein can easily be extended to incorporate new metabolic knowledge to cope

with the development of eukaryote phototrophs unicellular organisms studies, either through identifying new metabolic reactions, or accumulating reference protein sequences associated with a given reaction (2.1.3). For example, DMSP is a zwitterion and this charge means that it cannot cross cell membranes without a specific transporter[186]. There are two main families of the transporter that are known to be used by the bacteria Roseobacter, SAR11 clade bacteria, cyanobacteria, and also phytoplankton[187] : ABC (ATP binding casette) transporters[188], a commonly used primary transporter that can be found in all three domains of life[189]; and BCCT (betain, choline, carnithine transport) proteins[190]. Indeed, similarity in structure and properties between DMSP and its nitrogen analogue glycine-betaine (GBT) was noted[187,191]. We could then replace the DMSP DM reaction by a transporter reaction associated with some protein sequences found in the literature. However these transporters exist almost ubiquitously in microorganisms, and are not specific to DMSP producers.

While it is acknowledged that PhotoEukStein may contain inaccuracies in various aspects, it remains the most comprehensive and refined generic model currently available for phototrophic eukaryotes. Its development and curation involved the integration of diverse experimental data and literature sources, as well as extensive manual curation and refinement, resulting in a model that captures a broad range of metabolic processes and interactions. Therefore, PhotoEukStein represents an important step towards understanding and modelling of metabolism, physiology, biogeochemistry and ecology of phototrophic eukaryotes, and thus provides a valuable resource for researchers. This paves the way for an in-depth ecosystemic exploration of plankton communities from viruses to single-cell phototrophs.

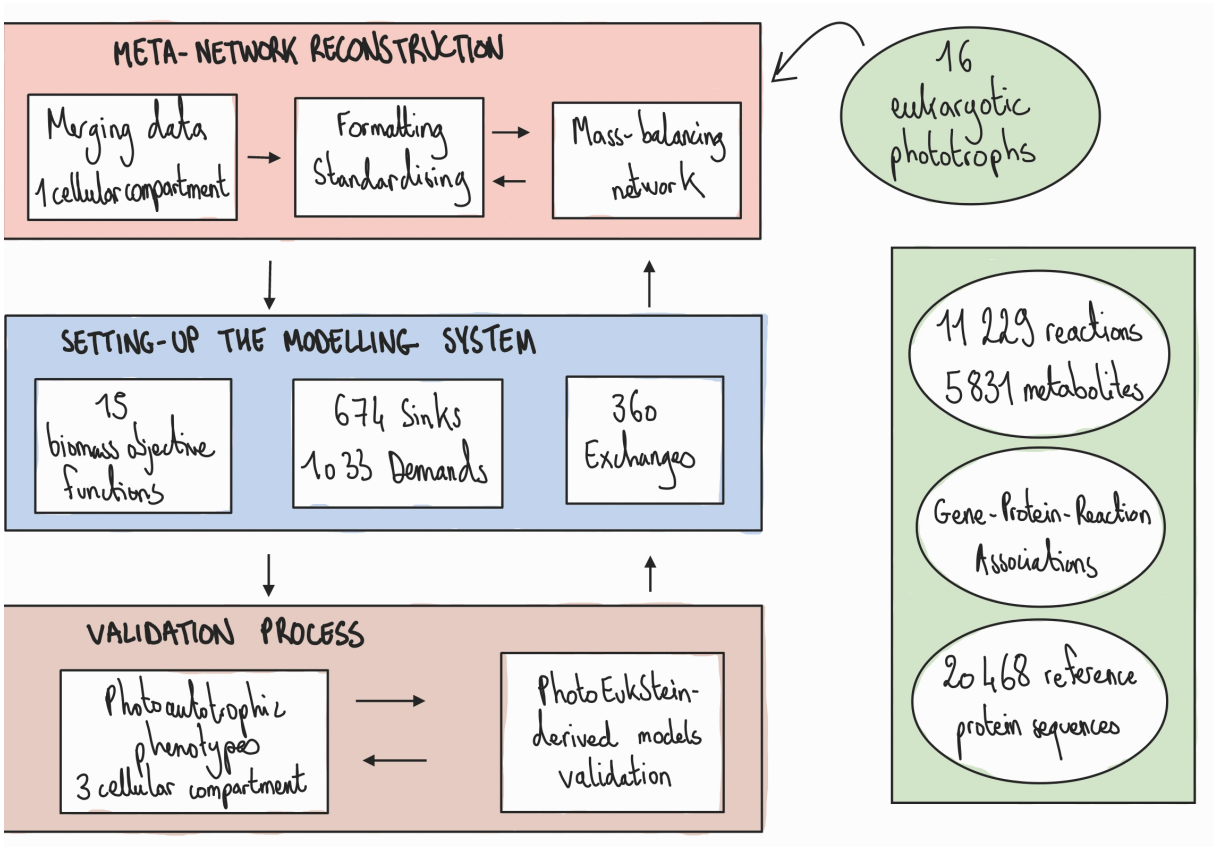


Figure 35: PhotoEukStein reconstruction workflow

3 A DATABASE OF MARINE PHOTOTROPHIC MICROEUKARYOTE METABOLIC MODELS

3.1 INTRODUCTION AND SUMMARY

PhotoEukStein enabled the fully automatic reconstruction of 549 constraint-based metabolic models from *Tara Oceans* environmental genomes and transcriptomes of phototrophic marine unicellular eukaryotes, providing a brand new valuable resource.

When focusing solely on the functional annotation of genes within the genomes, a phylogenetic signal is apparent[58]. An intriguing observation is that when we exclude the structural annotations and consider only the functional annotations that code for enzymes, a portion of this signal is lost. However, when we examine the content of reactions among the models, the phylogenetic signal remains relatively consistent. These three steps collectively indicate that the decrease of the phylogenetic signal is not attributed to PhotoEukStein's prediction of reaction content, but rather to the inherent taxonomic specificity of cell structure.

What adds further interest is the examination of how the different components of the systems, specifically the reactions, are interconnected. The phylogenetic signal is completely absent, and no specific pattern emerges. Instead, each network appears to be unique, resulting in a scattered distribution across the space. It is possible that the absence of compartments in PhotoEukStein contributes to a loss of information, and the structure of the metabolic networks may align with a phylogenetic pattern. Although it is important not to exclude this hypothesis, I firmly believe that interpretation can extend beyond these aspects. By focusing on functional aspects, this dispersion can also indicate a significant level of adaptability when considering all the networks collectively.

What is really powerful is to consider the fluxes dynamics of the models. Organisms have been tested for growth and DMSP production in many random environments. This time, clear clusters are emerging. These functionnal patterns respond similarly to environmental conditions and are not at all phylogenically related. Consequently, closely related organisms with a similar repertoire of metabolic reactions may exhibit dissimilar functional profiles, while distantly related organisms with different sets of metabolic reactions can mask metabolic similarities. Profiling organisms based on specific functional traits leads to distinct classifications that cannot be reduced solely to taxonomy or the presence/absence of a gene. We advocate for considering PhotoEukStein and its derived GSMs as a resource to highlight improved categories of omics-driven phenotypes that can be considered as potential traits in future ocean system models.

3.2 THESIS PAPER. PHOTOEUKSTEIN: TOWARDS AN OMICS-BASED DEFINITION OF UNICELLULAR EUKARYOTE PHOTOTROPHS FUNCTIONAL TRAITS VIA METABOLIC MODELLING

PhotoEukStein: Towards an omics-based definition of unicellular eukaryote phototrophs functional traits via metabolic modelling

Marie Burel [1,3*], Antoine Régimbeau [2,3], Samuel Chaffron [2,3], Damien Eveillard [2,3], Eric Pelletier [1,3,*]

[1] Génomique Métabolique UMR8030 Genoscope, Institut de Biologie François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
[2] Combi Team, LS2N, UMR6004 CNRS, Université de Nantes, Centrale Nantes, Nantes, France
[3] CNRS Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France

[*] To whom correspondence should be addressed: eric.pelletier@genoscope.fr - marie.burel@genoscope.fr

Abstract

Defining biological functional traits for unicellular organisms relies on comprehending the set and combination of the biochemical reactions their genomes encode for. This network of biochemical reactions defines the metabolic strategy organisms and communities used to grow in a given environment. While prokaryotes have been the ideal target for reconstructing and analysing these metabolic networks, eukaryotes lagged behind due to the complexity of their genomes and the paucity of knowledge on their metabolism.

Here, we developed PhotoEukstein, a meta-metabolic model for unicellular phototroph eukaryotes allowing a fast and automated top-down derivation of Genome-Scale Metabolic models directly from genomes. We applied it to a diverse collection of 559 environmental genomes and transcriptomes of marine eukaryote unicellular plankton.

We show these models allow to predict functional traits that cannot be purely deducted from taxonomic information or listing of metabolic reactions encoded by the genome. They provide the opportunity to build connections with Earth System Models to pinpoint environmental parameters needed to parametrise specific functional traits.

Introduction

Marine plankton are the dominant life form in the ocean, covering a broad diversity of organisms from viruses up to meter-size cnidarians via archaea, bacteria and single-celled eukaryotes, and have highly dynamic interactions. Together, these organisms play an active role in maintaining the Earth system, carrying out almost half of the net primary production on our planet [1] and exporting photosynthetically fixed carbon to the deep oceans [2]. Yet, a large part remains elusive to in-depth laboratory investigations. While ocean ecosystems biology investigates how biotic and abiotic processes determine emergent properties of the ocean ecosystem as a whole, the only biological knowledge we have from a large part of plankton comes from environmental genomics data.

With the ability to generate a vast amount of sequencing data out of environmental samples at ever-decreasing costs and the improvement of bioinformatics methods to reconstruct high-quality genomes from metagenomic data, several hundreds or thousands of Metagenome-Assembled Genomes (MAGs) have been reconstructed for viruses, bacteria, archaea, and eukaryotes, covering a large fraction of the biological diversity in several environments [3–12] . These environmental genomes greatly expand genomic and transcriptomic knowledge of cultured organisms [13,14]. Furthermore, most of these genomes correspond to organisms without cultured representatives, although they represent species playing essential contributions in the global biomass and cycling of nutrients. For example, while green algae and protists represent a third of the total marine biomass [15], eukaryote genomes recovered from marine environments are all differing from reference sequences and can even describe putative new phylum [8]. Omics-based approaches can hence significantly contribute to gain knowledge about the biology of these uncultured organisms.

However, one cannot mechanistically understand these organisms and how they interact with their environment through those sole prism of omics data. Functional annotation and phenotypic characterization are essential to allow us to gain further insight about "who is doing what" rather than answering the question of "who is here." In particular, systems

biology approaches have been instrumental to acquire a detailed stoichiometric representation of metabolic phenotypes via constraint-based reconstruction and analysis [16].

As biological features (traits/phenotypes) of organisms are primarily driven by their metabolic abilities [1,17], reconstructing metabolic networks from environmental -omics data provides a unique way to study the biology and ecology of these organisms and communities, as well as to draw a better picture of their influence on the environment. Indeed, metabolic networks are the cornerstone of Genome-Scale Metabolic Models (GSMs), which have demonstrated numerous applications in various field, such as biotechnology or synthetic biology [18]. GSMs are constraint-based models that use -omic knowledge in metabolic networks and reformulate it into linear inequalities. GSMs regroup all the metabolic reactions encoded in a genome or transcriptome, and their intertwining (cf. Methods, section 1). Using tools such as flux-balanced analysis or flux variation analysis [19] (see [18] for a review) to explore the solution space (i.e. the ensemble of possible solutions in the n-dimensions space defined by all the metabolic fluxes and that satisfy the constraints imposed on each flux in the GSM) [20], we can compute and predict metabolic phenotypes [21] through the optimisation of an objective function of interest, usually the growth rate. Even though GSMs do not take into account various biological regulations that modulate enzymatic activities within a cell (i.e. regulation of genes expression or protein synthesis, post-traductional modifications of proteins, or protein-protein interactions).

However, behind its benefit, reconstructing a metabolic network from -omics data analysis is a tedious task initially performed only for reference genomes, mobilising tedious laboratory experiments and metabolism experts for long periods and requiring expertise dedicated to a single genome [22,23] for a review). With every new genome sequenced, the traditional bottom-up approach requires that these time-consuming tasks are to be performed again. Metabolic modelling for eukaryotes has primarily been restricted to well-studied model organisms (*Homo sapiens*, *Arabidopsis thaliana, Phaeodactylum tricornutum, Saccharomyces cerevisiae* being the most complete examples). Few efforts have been devoted to unicellular phototroph organisms, even though they represent half of Earth's net primary production [24].

As a recent alternative, top-down semi-automated approaches deriving GSMs from a global reference pan-genomic collection of described reactions have been proposed [25]. The curation of such a generic model is performed only once, and is then converted into ready-for-use organism-specific models while preserving all manual curation and relevant structural properties. Among the most efficient algorithm for metabolic modelling, both in terms of computational time and quality of resulting models, is CarveMe [25,26], which has been used in various studies (for examples [27–31]) to derive prokaryotic GSMs. The EMBL-GEMs database (https://github.com/cdanielmachado/embl_gems) encompasses more than 5500 bacterial or archaeal simulation-ready GSMs.

Here, we report the development of PhotoEukStein, a reference-based metabolic meta-model for unicellular eukaryotic phototrophs, and its use with the CarveMe algorithm to reconstruct constraint-based metabolic models on a collection of 259 MAGs and 274 transcriptomes, plus the 16 references. The analysis of this resource revealed that, while there is a taxonomic imprinting of the repertoire of metabolic reactions across the 549 organisms considered, metabolic network topologies of resulting GSMs suggest that distantly related organisms can display similar metabolic phenotypes in a given environmental context. While a metabolic framework is particularly well-suited to formalize and analyse an organism ecological niche, deriving GSMs for unicellular (phototroph) eukaryotes paves the way to better describe their

metabolic phenotypes and functional traits and ecology.

Results

PhotoEukStein combines 16 existing available metabolic models of marine unicellular eukaryote phototrophs species ranging from Rhodophytes (red algae), Chlorophytes (green algae), Streptophytes, Stramenopiles (brown algae), and Haptophytes to Cryptophytes; along with *Arabidopsis thaliana* and *Klebsormidium nitens* (see Supplementary Table S1). These reference organisms cover most of the described taxonomical diversity of phototrophs eukaryotes (Figure 1a and Figure 1b). Out of these 16, 4 results from experts-curated bottom-up annotation, namely *Chlamydomonas reinhardtii* iRC1080 [32], *Chlorella variabilis* [33], *Phaeodactylum tricornutum* iLB1031 [34] and *Thalassiosira pseudonana* [35]. After the merging of these models as per the protocol described in [23], a manual curation phase was performed to make PhotoEukStein ready for constraint-based analyses (see Material and Methods for details). PhotoEukStein encompasses 5831 metabolites and 11229 reactions, 7599 of the later being associated with 20468 protein sequences from reference genomes. Two types of reactions are distinguished: 2067 boundary reactions (including 360 exchanges reactions) accounting for the transport of metabolites from or to the environment, and 9162 internal metabolic processes (see Supplementary Table S3).

For all 16 model organisms used for the construction PhotoEukStein, we derived GSM using the CarveMe algorithm. Nearly systematically (14 cases out of 16) more reactions and metabolites were included in PhotoEukStein-derived GSMs as compared to reference GSMs (from 1.5 to 6.7 times more reactions), a notable exception being *Guillardia theta* for which the reference GSM is only composed of 121 reactions (Supp. Table 2). Only two PhotoEukStein-based GSMs contain slightly fewer reactions than the reference one (*Cladosiphon okamuranus* - 86% and *Arabidopsis thaliana* - 98%). Interestingly, these two organisms are multicellular.

Several reasons can be invoked to explain this different number of considered reactions: i) ad-hoc expert-based models often focus on specific aspects of the metabolism (e.g., the lipids metabolism in *C. reinhardtii* [32]), ii) specific choices made during PhotoEukStein construction phase, were, for example, reactions only appearing in the *A. thaliana* GSM were discarded from PhotoEukStein, as potentially representing specific terrestrial multicellular plants-specific reactions, and iii) only reactions with identified corresponding protein sequences encoded in the genome were considered during the expert-led curation of the reference GSMs. Indeed, the CarveMe process allowed to include the minimal set of reactions from the meta-model that are mandatory to maintain a functional GSM (i.e. with gap-less pathways), even if they lack the identification of associated proteins. Thus, the process may overcome incomplete gene predictions or partial MAG completions. Moreover, it can point to critical reactions with yet unidentified associated genes in the genome.

For 3 out of the 16 reference GSMs (*Chlorella variabilis*, *Phaeodactylum tricornutum,* and *Thalassiosira pseudonana*), predicted growth rates were compared initially with culture-based data [33–36] (cf. Methods for details). In order to assess the validity of the PhotoEukStein-derived GSMs, we extensively sampled the metabolic niches and compared growth rates predicted from expert-based GSM with those obtained from PhotoEukStein-based GSMs (Supplementary Figure S1). Both predicted growth rates were highly correlated in all cases, showing that

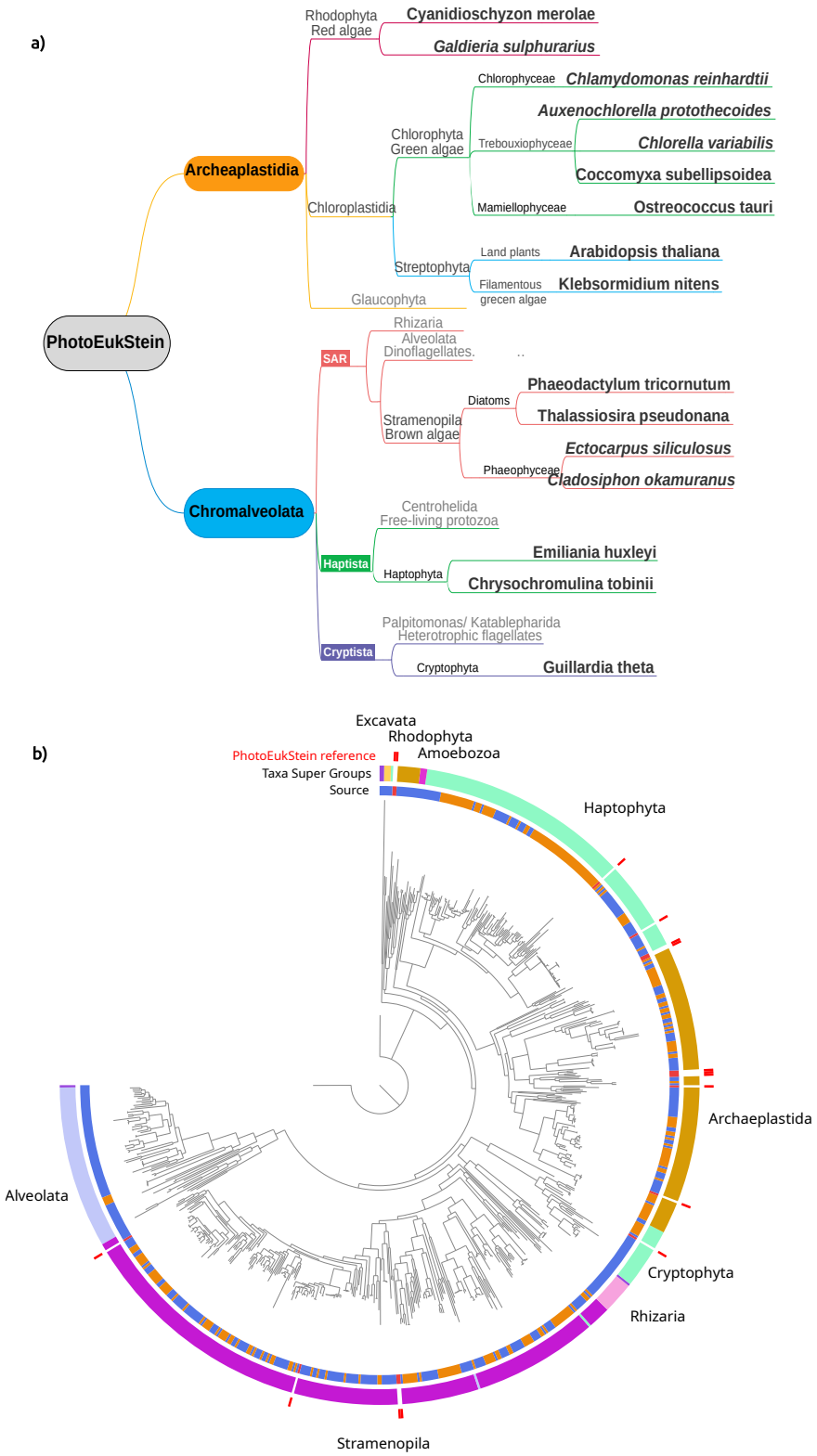


Figure 1: a) PhotoEukStein construction. Taxonomic diversity of the 16 existing GSMs that were combined to generate PhotoEukStein. b) Taxonomic diversity of the 553 PhotoeukStein-derived GSMs applied on 259 MAGs from Tara Oceans data (Delmont 2022, in orange in the inner circle) and 274 transcriptomes from METdb (Niang 2018, in blue in the inner circle). The taxonomic distribution of the 16 reference GSMs is indicated in ref (outer circle). Main taxonomic groups are indicated in the medium circle. Center is a dendrogram representing the taxonomy.

PhotoEukStein-derived GSMs are as efficient as expert-based models in capturing fundamental metabolic knowledge and corresponded to observations made with cultures. Therefore, PhotoEukStein-based GSMs are likely to provide valuable biological insights solely based on the genomic information of organisms.

Moreover, exploring metabolic niches with GSMs allows for assessing the metabolic exchange fluxes associated with higher growth rates, either for reference or PhotoEukStein-based GSMs, therefore delineating at environment's metabolic limitations for organisms or missing metabolic reactions. For the three reference species we studied, metabolite exchange fluxes differentiating growth rates between both models vary, and a given metabolite can favour growth in either reference or PhotoEukStein GSM, depending on the organism.

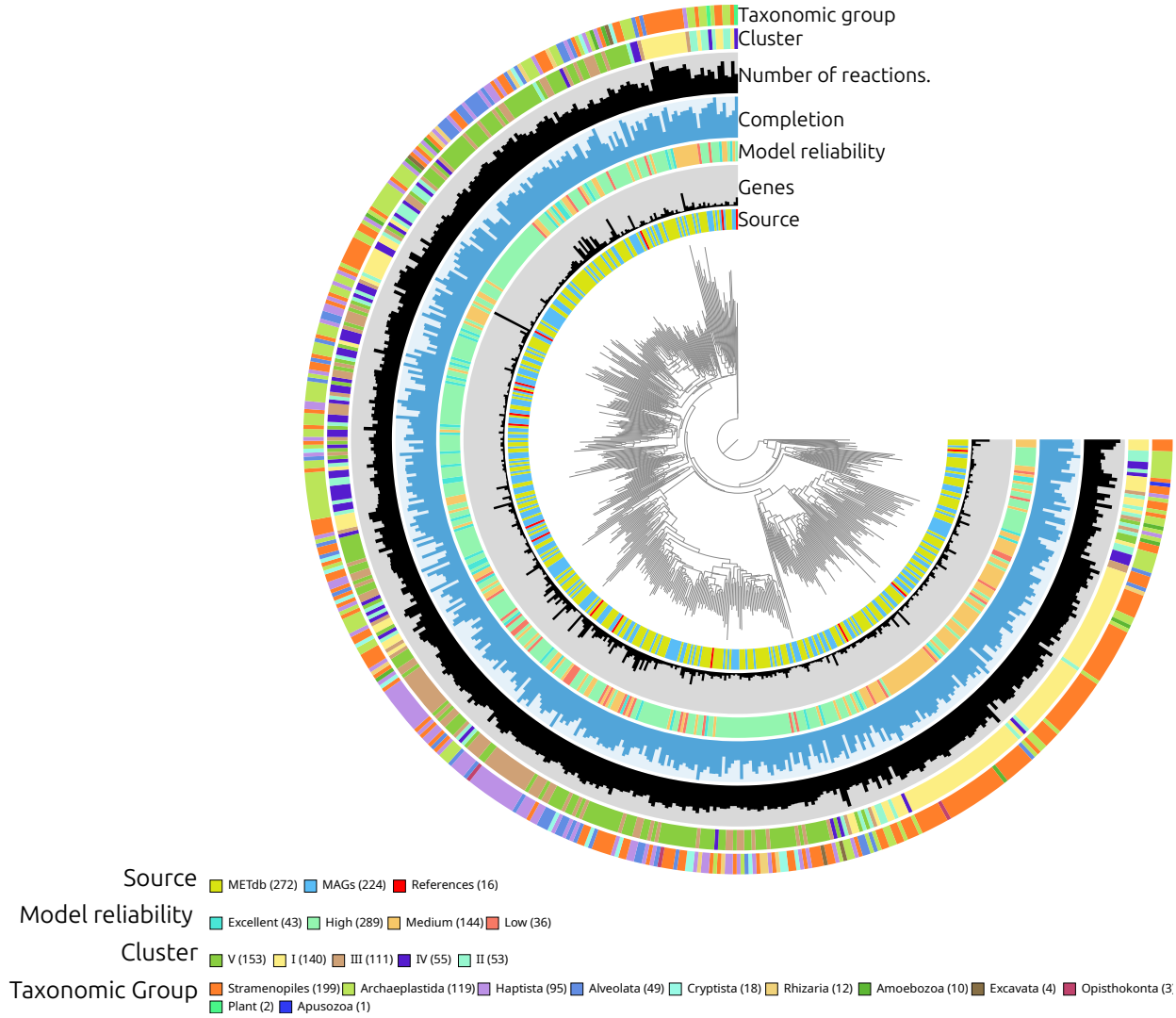


Figure 2: Main characteristics of PhotoEukStein derived GSMs. Central dendrogram represent Jaccard distance between GSMs reactions composition. Inner circle indicate the source of the sequence supporting the GSMs. Model reliability is defined by. Completion is the Busco-based evaluation of genome completion (Manni 2021) (see Supplementary Table S2 and Online Methods). The number of reactions indicates the extend of the GSMs (see Supplementary Table S2 and Online Methods). Clusters are defined following Supplementary Figure S4. Taxonomic groups are from (Delmont 2022) and reported in Supplementary Table S2. For the sake of readability, 10 Amoebozoa, 4 Excavata, 3 Opisthokonta and 2 Plants are displayed as “Others” in the Taxonomic supergroups ring.

To further scrutinise the internal consistency of PhotoEukStein-derived GSMs, we compared reaction fluxes as predicted by both models for *Phaeodactylum tricornutum* (for details, cf.

Materials and Methods section 3). We considered fluxes correlations between reactions within each model by sampling the whole metabolic space (Supplementary Figure S2). Resulting correlation maps were highly similar in reference and PhotoEukStein-derived GSMs, indicating that reactions in both models are very similarly interconnected. We confirmed this observation by plotting intra-model correlations values distribution (Supplementary Figure S3). When pairs of reactions were highly correlated within one GSM, they are as highly correlated within the other GSM, and low connected pairs of reactions had similar characteristics in both models. The automated top-down approach applied to *P. tricornutum* captured the same essentials as the expert-based model and represents the same biological features even when considering the distribution of metabolic fluxes within the GSM.

Unicellular eukaryote phototrophs MAGs data from *Tara* Oceans were extracted from previous study [8], while unicellular eukaryotic phototrophs transcriptomes were recovered from the METdb database [14] which extends the MMETSP resource [13]. In total, 259 MAGs and 274 transcriptomes, genomic data, along with the 16 organisms used as a reference to build PhotoEukStein (Supplementary Table S2), were used as input for the CarveMe method [25] to derive 549 dedicated GSMs from PhotoEukStein (for details, cf Materials and Method section 4).

The resulting GSMs contain a mean of 4154 reactions each (min. 1350, max. 7045), 72.7% of them (min. 41.6%, max 89.1%) being associated with a gene from the MAG/transcriptome input (as compared with the 67.67 % of PhotoEukStein reactions being associated with a reference sequence). As anticipated, the number of reactions in the resulting GSMs retained during the graph refinement (or carving) process decreased with the estimated level of genomes completion (Supplementary Table S2, Figure Extended Data). Nevertheless, when dealing with partially complete genomes, the intrinsic feature of CarveMe to keep reactions within the GSM even without supporting protein evidence allowed us to highlight mandatory reactions yet to be identified in a given genome.

The repertoire of metabolic reactions across species, as the global repertoire of genes of a given genome, is mainly influenced by the phylogeny. Acquisition of new metabolic functions resulting from horizontal gene transfers from viruses [37,38] or bacteria [39], can superimpose alternative connectivity within the metabolic pathways. In order to assess the reliability of the functional capabilities of marine unicellular phototroph eukaryotes across taxonomy, we analysed the reaction content of PhotoEukStein-derived GSMs. First, we computed the Jaccard distance between our 549 GSMs based on the presence/absence of metabolic reactions (Supplementary Figure S4). We observe groups of reactions associated explicitly with some taxonomical groups (for example, reactions linked with lipid metabolism are specific to the diatoms). We sketch a more global picture of the distribution of metabolic reactions across our collection of GSMs, by performing a Uniform Manifold Approximation and Projection (UMAP) analysis [40,41] of the presence/absence of reactions within each GSM (as listed in Supplementary Table S3). This showed that GSMs are not evenly distributed in the functional space and that there is strong imprinting of the taxonomic origin of the corresponding organisms in the functional proximity (Supplementary Figure S5). This observation agrees with the vertical transmission of most metabolic-associated genes (for a review, see [42]).

To further define the distribution of GSMs within the functional space, we performed a k-means clustering. Five clusters corresponding to metabolic groups were revealed (Supplementary Figure S5, Supplementary Table S2). When taxonomic information

associated with each GSM is projected on the clusters, we observe a substantial taxonomic-based composition effect: most of the Stramenopiles are concentrated in cluster I (131 out of 212) and V (66 out of 212), Archaeplastida in clusters II (52 out of 118) and III (52 out of 118), Hacrobia in cluster IV (114 out of 135), and Alveolata and Amoebozoa in cluster V (respectively 40 out of 51 and 8 out of 10). Cluster V was the more taxonomically diverse, with eight taxonomic supergroups represented out of 10.

Beyond the presence/absence of reaction analysis of metabolic models, studying their topology provides insights into how metabolites are intertwined within the model. We explored the diversity of our collection of 549 PhotoEukStein-derived GSMs using a diffusion map approach [43,44] to capture the non-linear combinations of metabolic capabilities variables represented in a 549 dimension space representing the internal metabolic reactions of GSMs. We then proceeded to a dimension reduction to visualise that diversity using the UMAP algorithm (Figure 3).

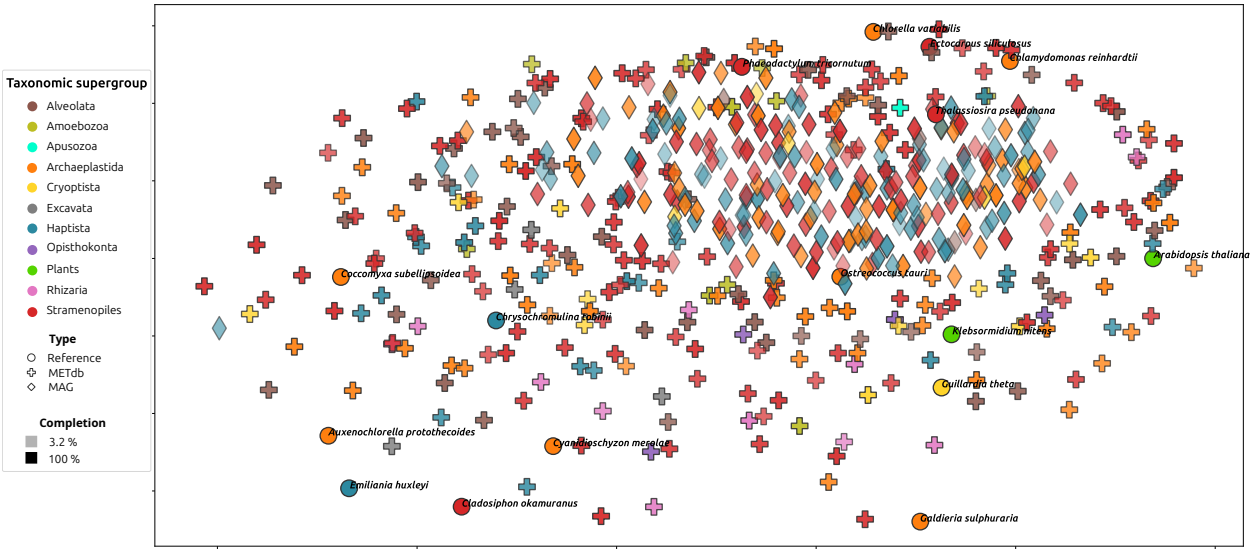


Figure 3: UMAP representation of diffusion map analysis of the 553 GSMs topology. MAGs are symbolised by diamonds, METdb by crosses, and reference genomes by circles. Colours indicate the taxonomic groups of each supporting genome, while transparency represents Busco-based genome completion estimation.

We showed a global spread of PhotoEukStein GSMs in the metabolic topological space, indicating that both these models are globally functionally diverse but also specific to each genome or transcriptome they are built upon. Moreover, despite the wide breadth of taxonomic diversity covered by these organisms, there is no evidence of structuration of the metabolic connectivity space based on the taxonomy. As diffusion map analysis captures the connectivity between reactions, this observation suggests that the taxonomy does not critically influence the metabolic circuitry of organisms. This result indicates how the various organisms mobilise their metabolic capabilities to produce biomass. These two visions of metabolic behaviour are similar to the genotype and phenotype (i.e., the difference between the potential and its realisation in a given set of conditions).

To further evaluate the ability of PhotoEukStein-based GSMs to respond to environmental changes and, therefore, to capture an organism' responses to environmental variability, we applied combinations of available metabolites fluxes (Supplementary Table S5) and evaluate GSMs predicted growth rate variations (Figure 4). In all cases, the predicted growth rate increased compared to the reference medium (Figure 4, Supplementary Table S6). As the

metabolites in the permuted pool can fuel a wide range of reactions, their addition increases the usability of possible metabolic routes to produce biomass. But the various models do not exhibit the same profile of response depending on the added metabolites combinations, and we can define groups of models sharing similar profiles, hence defining functional traits (Supplementary Figure S7). Interestingly, growth profiles correlate poorly with taxonomy or genome-wide gene content, similarly to what have already been described for Bacterias [45].

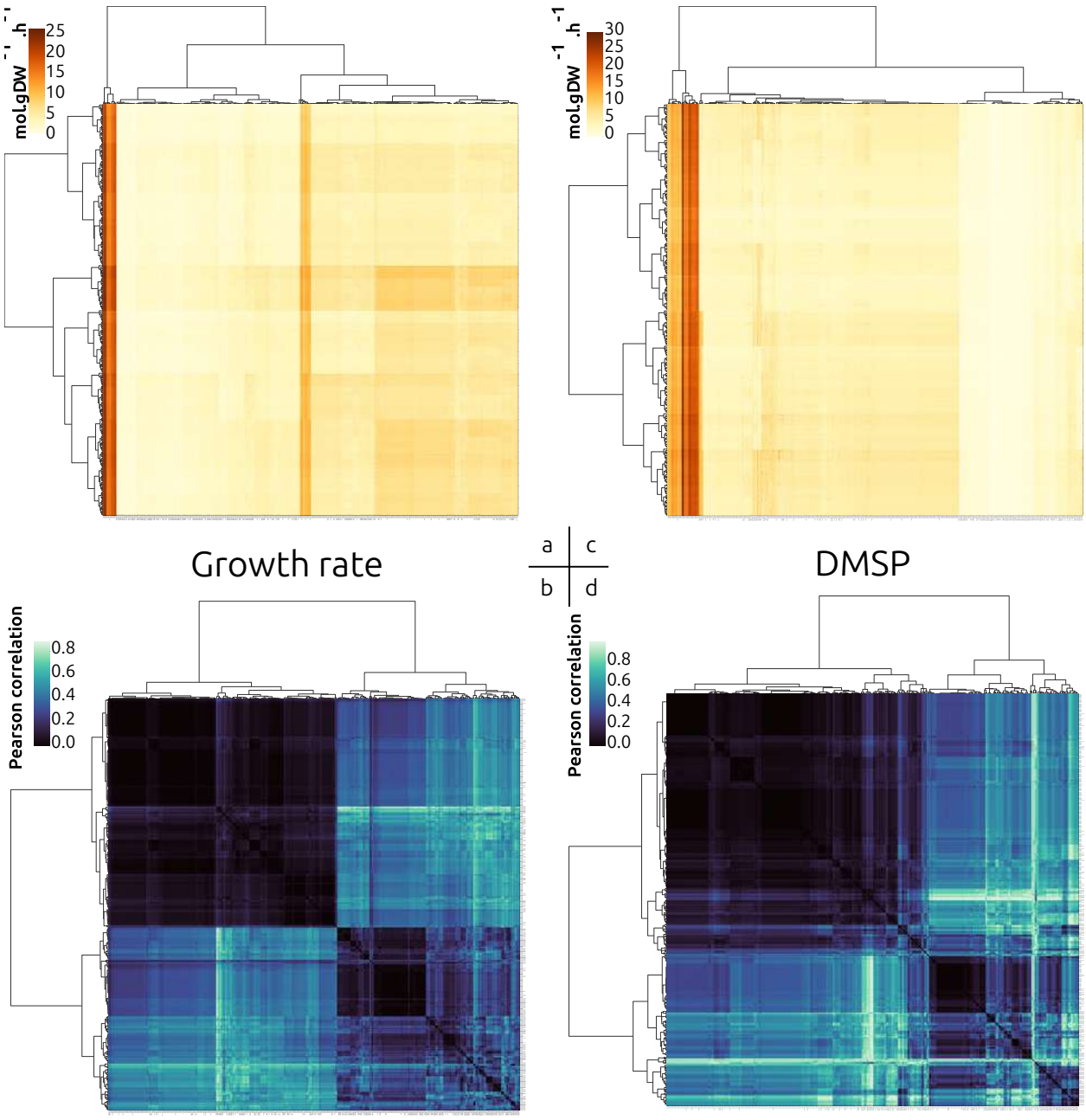


Figure 4: Metabolic niche exploration. The 553 GSM models are exposed to medium modification by systematic permutations of 1 up to 9 extra metabolites (listed in Supplementary Table S4 and S5), and growth rate is computed a). b) represents the clustered correlation matrix of growth rates modification profiles across the 1023 permutations. c) DMSP production rate as computed for the 337 GSMs producing that molecule for the same metabolic niche permutations as panel a, and d) shows the clustered correlation matrix of DMSP production variation under the 1023 permutations.

Outside of growth rate, GSMs can define functional traits. We considered the 387 GSMs

having the ability to produce dimethylsulfoniopropionate (DMSP), a molecule that plays many important roles for marine life, including use as an osmolyte, antioxidant, predator deterrent, and cryoprotectant for phytoplankton and as a reduced carbon and sulphur source for marine bacteria. It also produces the climatically active gas dimethyl sulphide (DMS), the primary natural source of sulphur to the atmosphere [46]. We computed the amount of DMSP produced in the various conditions tested (Figure 4). Similarly as for the growth rate variations, we observe that different GSMs respond to medium changes with various patterns, and when we consider these response profiles, we can define 2 very differently responsive groups of GSMs that, once again, do not follow taxonomy discrimination that we can consider as describing new functional trait.

Discussion

While bacterial and archaeal metabolisms have been extensively studied, much fewer efforts have been devoted to eukaryotes, and even more so for multicellular species. With the growing number of available environmental and isolate genome data, the repertoire of available MAGs and transcriptomes representing planktonic eukaryote species distantly related to reference organisms is franticly expanding. However, while they cover a broader diversity than the well-studied references, there is a lack of efficient ways to study their biology [47].

We propose PhotoEukstein as the first meta-metabolic model of unicellular phototroph eukaryotes for a fast and efficient top-down derivation of GSMs, applicable to genomes and transcriptomes. Its development and curation involved the integration of diverse experimental data and literature sources, as well as extensive manual curation and refinement, resulting in a model that captures a broad range of metabolic processes and interactions. We efficiently applied it to a collection of taxonomically diverse environmental genomes and transcriptomes covering a wide range of yet barely functionally described marine eukaryote unicellular planktons. We have shown that growth rates predicted from these GSMs are highly comparable for the three reference organisms and with the ones obtained from in-vitro measurements. Therefore, PhotoEukStein represents an important step towards understanding and modelling of metabolism, physiology, biogeochemistry and ecology of phototrophic eukaryotes. We propose a valuable resource of 549 new metabolic models for researchers, paving the way for an in-depth ecosystemic exploration of plankton communities from viruses to single-cell phototrophs. Moreover, PhotoEukStein can easily be extended to incorporate new metabolic knowledge to cope with the development of eukaryote phototrophs unicellular organisms studies, either through identifying or accumulating reference protein sequences associated with a given reaction or the description of new metabolic reactions and related protein sequences. Deriving PhotoEukStein-based GSMs from new genomes or transcriptomes does not requires heavy computational resources nor time-consuming expertise, allowing to cope with the rapidly growing repertoire of environmental genomes.

Metabolic models are well suited to represent the metabolic phenotype of microorganisms [21] and may provide a better specific delineation of functional traits distribution across species than solely considering taxonomy or presence/absence of particular genes. On one hand we have annotation based distance that follows phylogenetic distance [8], on the other hand we have laboratory experiments that assess phenotypes without finding correlation with

phylogenetic distances [45].

Similarly to the later, our results show a robust phylogenetic signal affects the metabolic reactions profiles composition (Figure 2 and Supplementary Figure S3) while no such signal is detected in functional/phenotypic GSMs clustering (Figure 3).It results that closely related organisms with similar repertoire of metabolic reactions may display dissimilar functional profiles, and (inversely) that distantly related organisms with a different set of metabolic reactions pools can mask metabolic similarities. Another one relates to the consideration that metabolic reactions act together to form biological functions. Profiling of organisms for each given functional trait will generate specific classifications that cannot a priori be reduced to taxonomy or presence/absence of a gene. Understanding the biological functions of organisms involves deciphering their metabolic capabilities, and using GSMs for this purpose could be the most effective even when only environmental genomic data are available.

Metabolic niches represent sets of environmental parameters (in the form of fluxes of available metabolites) for which a given metabolic model can generate biomass [20]. It's a formalisation of the organismal function as a space in which an organism can survive based on its ability to cope with the available resources through the set of metabolic reactions it holds. Being able to derive functional GSMs for unicellular phototrophs eukaryotes, even from environmental omics data, provides a unique way to assess their biological phenotype *per se* as it differs from the sole identification of functional genes. These features are new observations, or semantic traits that arise from genomics descriptions. For example, scrutinising the variability of metabolic fluxes and metabolites exchanges through the study of metabolic niches may allow differentiating allocation of cellular resources to resource acquisition, defence, signalling, and other survival needs [49], as well as community metabolic interactions as considered in the phycosphere [50] or the holobiont [51] concepts. Therefore, the ability to systematically derive GSMs for unicellular eukaryote phototrophs, as it is already the case for heterotroph prokaryotes, is an essential step toward a global description of phenotypic biodiversity and ecosystems modelling. In particular, we advocate for considering PhotoEukStein and derived GSMs as a resource for emphasizing better classes of omics-driven phenotypes that will considered as potential traits in future ocean system modellings.

Acknowledgements & Funding

Our survey was made possible by the sampling and sequencing efforts of the Tara Oceans Project. We are indebted to all who contributed to these efforts and other open-source bioinformatics tools for their commitment to transparency and openness. Tara Oceans (which includes the Tara Oceans and Tara Oceans Polar Circle expeditions) would not exist without the leadership of the Tara Oceans Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/). We also acknowledge the commitment of the CNRS and Genoscope/CEA. Some computations were performed thanks to the TGCC computing facility in France.. This study was supported in part by FRANCE GENOMIQUE (ANR-10-INBS-09). M.B. was funded by the Doctoral School "Structure and Dynamics of Living Systems" of Paris Saclay University. This article is contribution number XXX of Tara Oceans.

Data availability

PhotoEukStein is available at https://www.genoscope.cns.fr/PhotoEukStein/

Supplementary tables are available at https://www.genoscope.cns.fr/PhotoEukStein/

References

1. Alexander, H. *et al.* Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *PNAS* **112**, E5972–E5979 (2015).
2. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
3. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* **5**, 170203 (2018).
4. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
5. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533 (2017).
6. Olm, M. R. *et al.* Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* **7**, 26 (2019).
7. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569–580 (2018).
8. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **0**, (2022).
9. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
10. Duncan, A. *et al.* Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. *Microbiome* **10**, 1–21 (2022).
11. Alexander, H. *et al.* Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. 2021.07.25.453713 Preprint at https://doi.org/10.1101/2021.07.25.453713 (2022).
12. Delmont, T. O. *et al.* Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *ISME J* 1–10 (2021) doi:10.1038/s41396-021-01135-1.
13. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology* **12**, e1001889 (2014).
14. Niang, G. *et al.* METdb: A Genomic Reference Database For Marine Species. *F1000Research* **9**, (2020).
15. Abreu, A. *et al.* Priorities for ocean microbiome research. *Nat Microbiol* **7**, 937–947 (2022).
16. Palsson, B. Ø. *Systems Biology: Properties of Reconstructed Networks*. (2006).
17. Martini, S. *et al.* Functional trait-based approaches as a common framework for aquatic ecologists. *Limnology and Oceanography* **66**, 965–994 (2021).
18. Fang, X., Lloyd, C. J. & Palsson, B. Ø. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat Rev Microbiol* **18**, 731–743 (2020).
19. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature Biotechnology* **28**, 245–248 (2010).
20. Régimbeau, A. *et al.* Contribution of genome-scale metabolic modelling to niche theory. *Ecology Letters* **25**, 1352–1364 (2022).
21. Varma, A. & Palsson, B. Ø. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and Environmental Microbiology* **60**, 3724–3731 (1994).
22. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N. & Barabási, A.-L. Global organization of metabolic fluxes in the bacterium Escherichia coli. *Nature* **427**, 839–843 (2004).
23. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* **5**, 93–121 (2010).
24. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* **281**, 237–240 (1998).
25. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research* **46**, 7542–

7553 (2018).
26. Mendoza, S. N., Olivier, B. G., Molenaar, D. & Teusink, B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol* **20**, 1–20 (2019).
27. Magnúsdóttir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* **35**, 81–89 (2017).
28. Heinken, A. *et al.* Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nat Biotechnol* 1–12 (2023) doi:10.1038/s41587-022-01628-0.
29. Zorrilla, F., Buric, F., Patil, K. R. & Zelezniak, A. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Research* **49**, e126 (2021).
30. Bidkhori, G. *et al.* The Reactobiome Unravels a New Paradigm in Human Gut Microbiome Metabolism. 2021.02.01.428114 Preprint at https://doi.org/10.1101/2021.02.01.428114 (2021).
31. Zimmermann, J., Kaleta, C. & Waschina, S. gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* **22**, 1–35 (2021).
32. Chang, R. L. *et al.* Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Molecular Systems Biology* **7**, 518 (2011).
33. Juneja, A., Chaplen, F. W. R. & Murthy, G. S. Genome scale metabolic reconstruction of Chlorella variabilis for exploring its metabolic potential for biofuels. *Bioresource Technology* **213**, 103–110 (2016).
34. Broddrick, J. T. *et al.* Cross-compartment metabolic coupling enables flexible photoprotective mechanisms in the diatom Phaeodactylum tricornutum. *New Phytologist* **222**, 1364–1379 (2019).
35. van Tol, H. M. & Armbrust, E. V. Genome-scale metabolic model of the diatom Thalassiosira pseudonana highlights the importance of nitrogen and sulfur metabolism in redox balance. *PLOS ONE* **16**, e0241960 (2021).
36. Salguero, D. A. M. *et al.* Development of a Chlamydomonas reinhardtii metabolic network dynamic model to describe distinct phenotypes occurring at different CO2 levels. *PeerJ* **6**, e5528 (2018).
37. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472–482 (2015).
38. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336 (2022).
39. Husnik, F. & McCutcheon, J. P. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol* **16**, 67–79 (2018).
40. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
41. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37**, 38–44 (2019).
42. Caetano-Anollés, G. *et al.* The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology* **41**, 285–297 (2009).
43. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **102**, 7426–7431 (2005).
44. Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nat Commun* **11**, 4887 (2020).
45. Forchielli, E., Sher, D. & Segrè, D. Metabolic Phenotyping of Marine Heterotrophs on Refactored Media Reveals Diverse Metabolic Adaptations and Lifestyle Strategies. *mSystems* **7**, e00070-22 (2022).
46. Teng, Z.-J. *et al.* Biogeographic traits of dimethyl sulfide and dimethylsulfoniopropionate cycling in polar oceans. *Microbiome* **9**, 207 (2021).
47. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).
48. Strom, S. L. Microbial Ecology of Ocean Biogeochemistry: A Community Perspective. *Science* **320**, 1043–1045 (2008).
49. Seymour, J. R., Amin, S. A., Raina, J.-B. & Stocker, R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat Microbiol* **2**, 1–12 (2017).
50. Dittami, S. M. *et al. A community perspective on the concept of marine holobionts: current status, challenges, and future directions*. https://peerj.com/preprints/27519 (2019) doi:10.7287/peerj.preprints.27519v3.

Authors contributions

M.B., A.R., D.E. and E.P. designed the research, M.B. build, curated and applied PhotoEukStein, M.B. and A.R. validated PhotoEukStein, designed numerical experiments with inputs from D.E, and generated the data, M.B, A.R., D.E., and E.P. analysed the results, M.B. and E.P. wrote the manuscript with significant inputs from all authors.

Figure legends

Figure 1: a) PhotoEukStein construction. Taxonomic diversity of the 16 existing GSMs that were combined to generate PhotoEukStein. b) Taxonomic diversity of the 553 PhotoeukStein-derived GSMs applied on 259 MAGs from Tara Oceans data (Delmont 2022, in orange in the inner circle) and 274 transcriptomes from METdb (Niang 2018, in blue in the inner circle). The taxonomic distribution of the 16 reference GSMs is indicated in ref (outer circle). Main taxonomic groups are indicated in the medium circle. Center is a dendrogram representing the taxonomy.

Figure 2: Main characteristics of PhotoEukStein derived GSMs. Central dendrogram represent Jaccard distance between GSMs reactions composition. Inner circle indicate the source of the sequence supporting the GSMs. Model reliability is defined by. Completion is the Busco-based evaluation of genome completion (Manni 2021) (see Supplementary Table S2 and Online Methods). The number of reactions indicates the extend of the GSMs (see Supplementary Table S2 and Online Methods). Clusters are defined following Supplementary Figure S4. Taxonomic groups are from (Delmont 2022) and reported in Supplementary Table S2. For the sake of readability, 10 Amoebozoa, 4 Excavata, 3 Opisthokonta and 2 Plants are displayed as “Others” in the Taxonomic supergroups ring..

Figure 3: UMAP representation of diffusion map analysis of the 553 GSMs topology. MAGs are symbolised by diamonds, METdb by crosses, and reference genomes by circles. Colours indicate the taxonomic groups of each supporting genome, while transparency represents Busco-based genome completion estimation.

Figure 4: Metabolic niche exploration. The 553 GSM models are exposed to medium modification by systematic permutations of 1 up to 9 extra metabolites (listed in Supplementary Table S4 and S5), and growth rate is computed a). b) represents the clustered correlation matrix of growth rates modification profiles across the 1023 permutations. c) DMSP production rate as computed for the 337 GSMs producing that molecule for the same metabolic niche permutations as panel a, and d) shows the clustered correlation matrix of DMSP production variation under the 1023 permutations.

Supplementary Figure S1: Comparison of predicted growth rates from PhotoEukStein or reference GSMs for *Phaeodactylum tricornutum* (top left), *Chlorella variabilis* (middle left) and *Thalassiosira pseudonana* (bottom left). In each case, 10,000 iterations of random sampling within each GSMs’ niche space was performed and growth rates predicted for both models and reported. On the right are indicated the metabolites exchange reaction the most correlated with predicted growth rates differences between reference and PhotoEukStein GSMs.

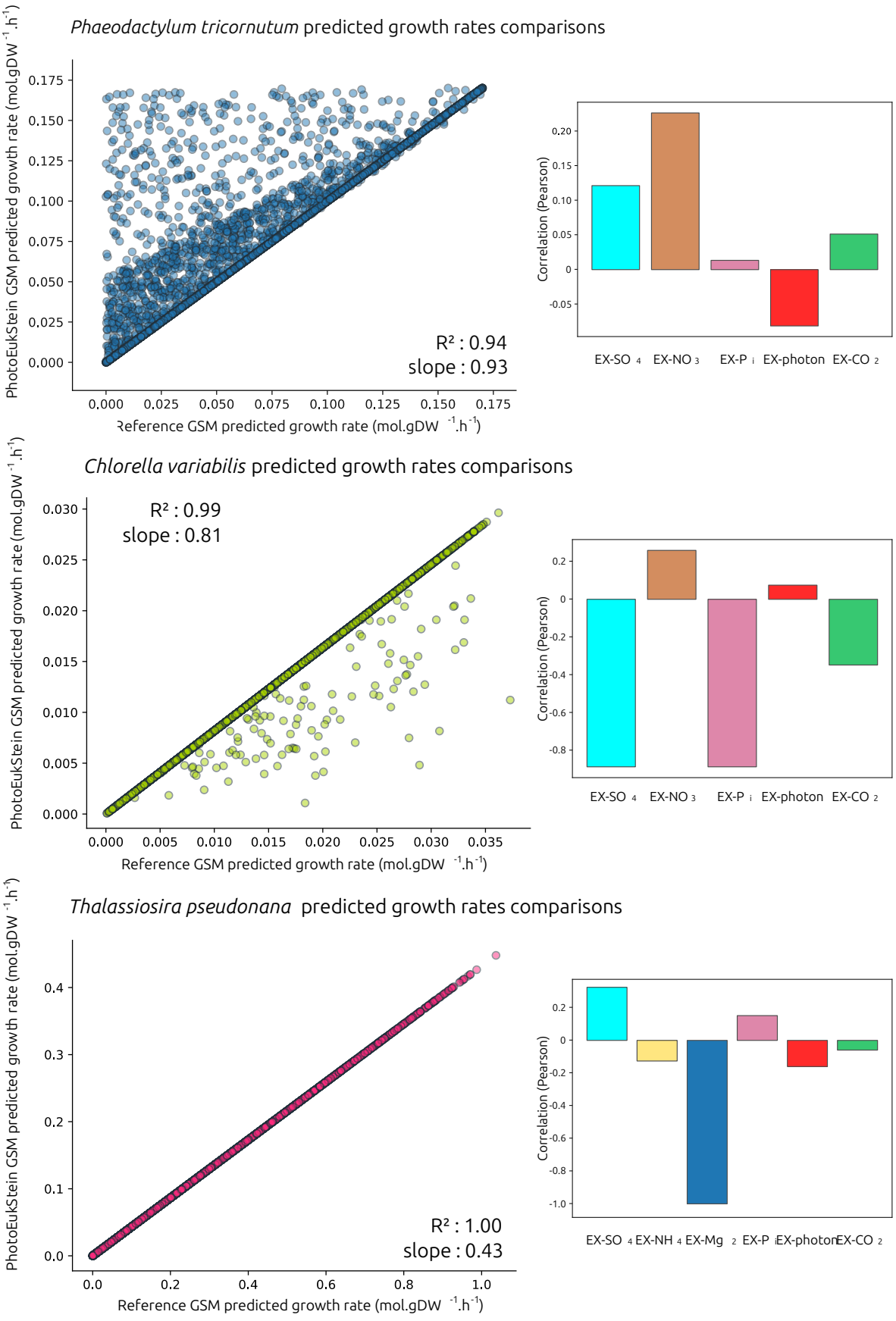
Supplementary Figure S2: Correlation matrix comparing original (iLB1034) and PhotoEukStein-derived GSMs reaction fluxes of *Phaeodactylum tricornutum*. The 434 shared reaction showing at least 1% or their correlations with other reactions fluxes greater to 0.2 when randomly sampling the whole metabolic space are displayed. Lower left rectangle:

iLB1034 reactions, upper right rectangle, PhotoEukStein reactions. Only correlation with absolute value greater than 0.5 are coloured.

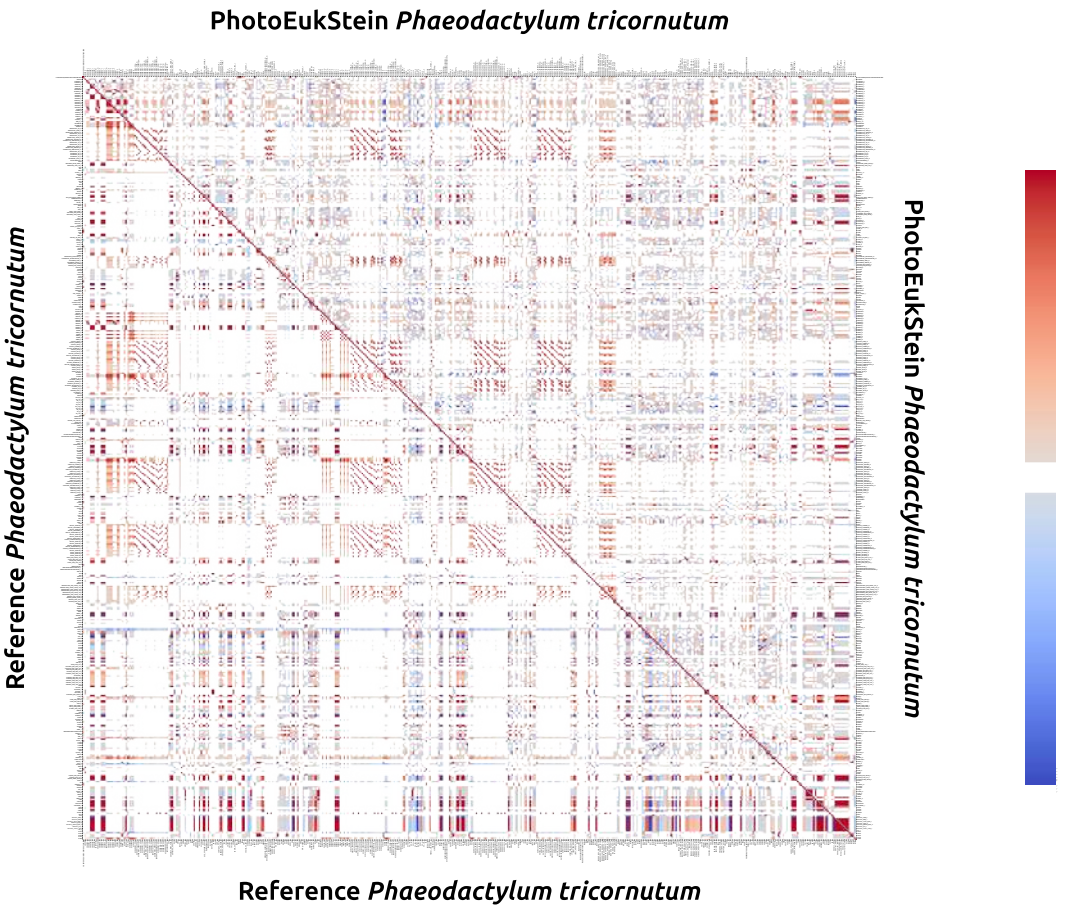
Supplementary Figure S3: HexBin representation of distribution of compared reaction fluxes correlations between PhotoEukStein and iLB1034 reference GSMs for *Phaeodactylum tricornutum*. Effective of correlations pairs of the 434 common reactions showing a possible correlation displayed in Figure S2 (see Materials and Methods section 3). Each cell indicates the number of reactions with corresponding correlations values pairs in iLB1034 (x-axis) and PhotoEukStein (y-axis) during random sampling of metabolic niche space.

Supplementary Figure S4: Compositional analysis of 512 PhotoEukStein-derived GSMs. Presence/absence of 9648 reactions were used to hierarchically cluster (euclidian distance, ward distance) both GSMs (lines) and reactions (columns). Genome source, Taxonomic group, model quality score (as defined in Extended Data), and metabolic cluster (as defined in Figure S5) are shown for each GSM. Frequency of reaction appearance among the 512 GSMs (bad quality models excluded) is indicated for each reaction (from white= 1 to dark red=512). Blue indicates significantly present reactions in a cluster, red if the absence of the reaction is signature reaction, grey if present but not significant.may nevertheless share similar biological features. Several implications arise from this apparent contradiction. The first relies to functional redundancy, notably the fact that several metabolic pathways can connect one metabolite to the other [48]. Thus distinct reaction

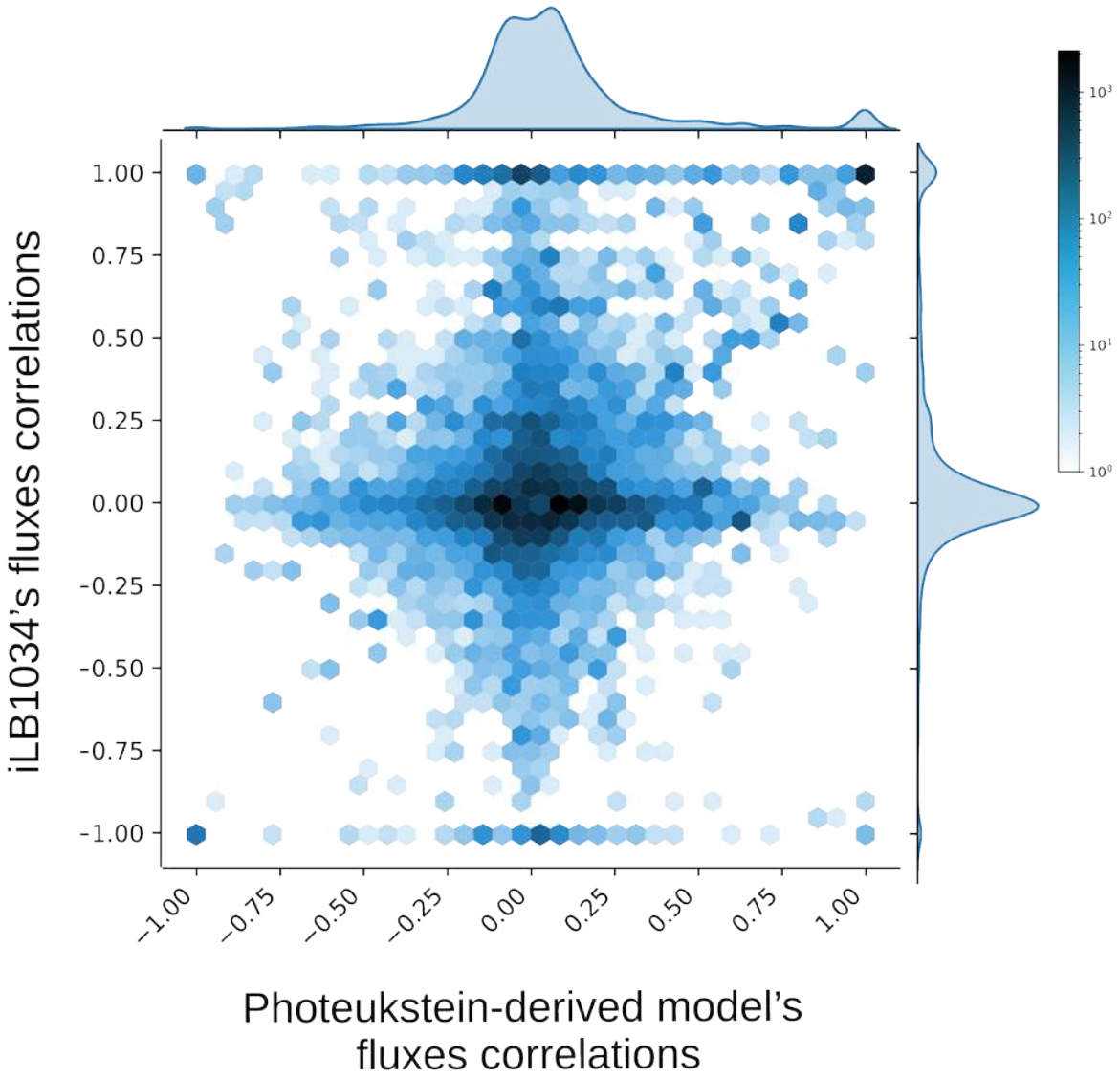
Supplementary Figure S5: Compositional analysis of the 549 PhotoEukStein derived GSMs. a) Diffusion map multiscale geometric analysis of GSMs topology followed by umap reduction of dimension processing have been performed to study the distribution of GSMs topologies in the functional space. Shapes indicate origin of the supporting genome (diamonds for MAGs, crosses for METdb and circles for references). Colours indicate main taxonomical groups, and transparency reflect genomes/transcriptomes completions estimations. Grey ellipses identify the 5 clusters supported by k-means signal deconvolution analysis (see Materials and Methods). b) Repartition of main taxonomic groups within each cluster. c) Distribution of each taxonomic group across the clusters.



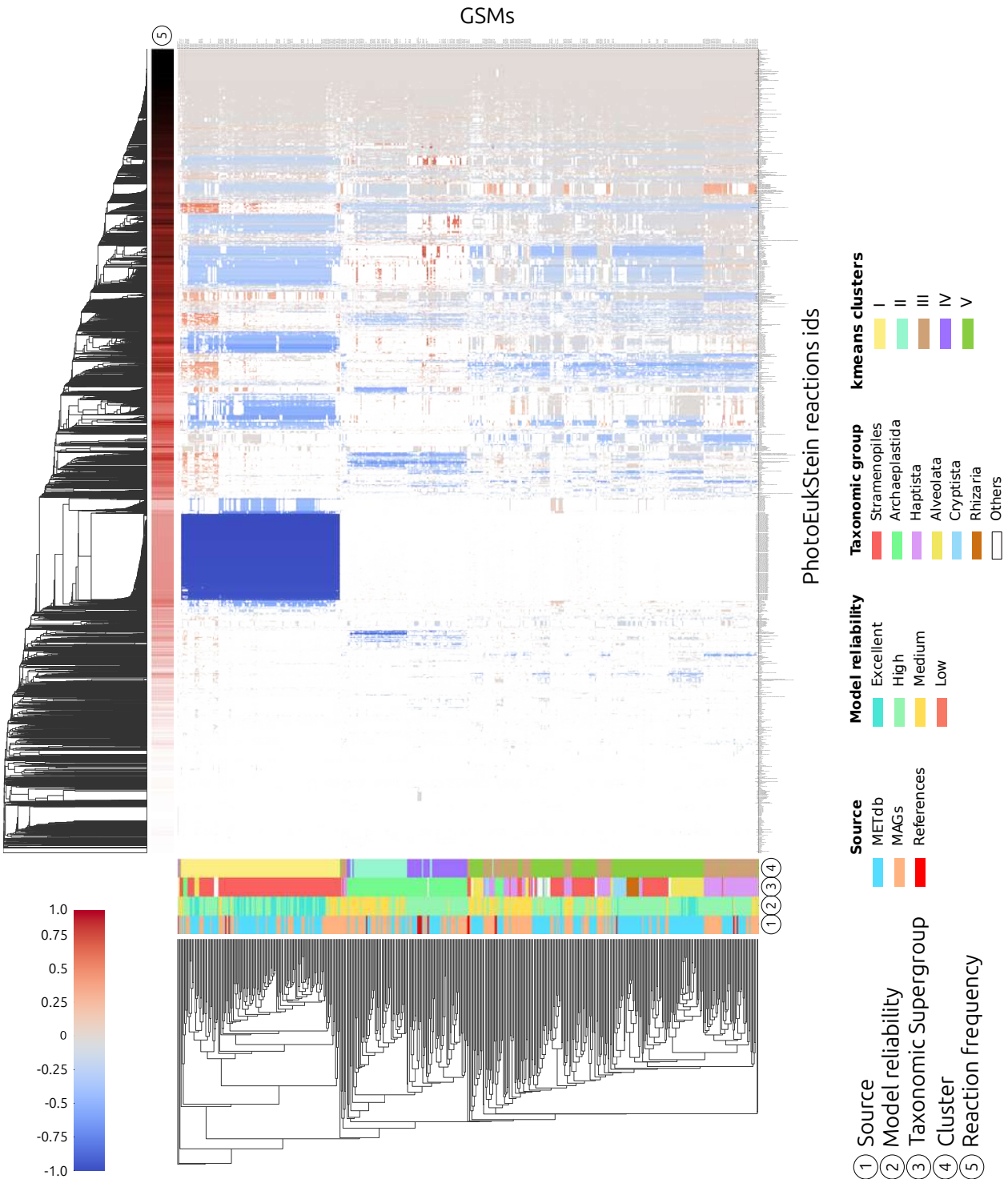
Supplementary Figure S1: Comparison of predicted growth rates from PhotoEukStein or reference GSMs for Phaeodactylum tricornutum (top left), Chlorella variabilis (middle left) and Thalassiosira pseudonana (bottom left). In each case, 10,000 iterations of random sampling within each GSMs' niche space was performed and growth rates predicted for both models and reported. On the right are indicated the metabolites exchange reaction the most correlated with predicted growth rates differences between reference and PhotoEukStein GSMs.



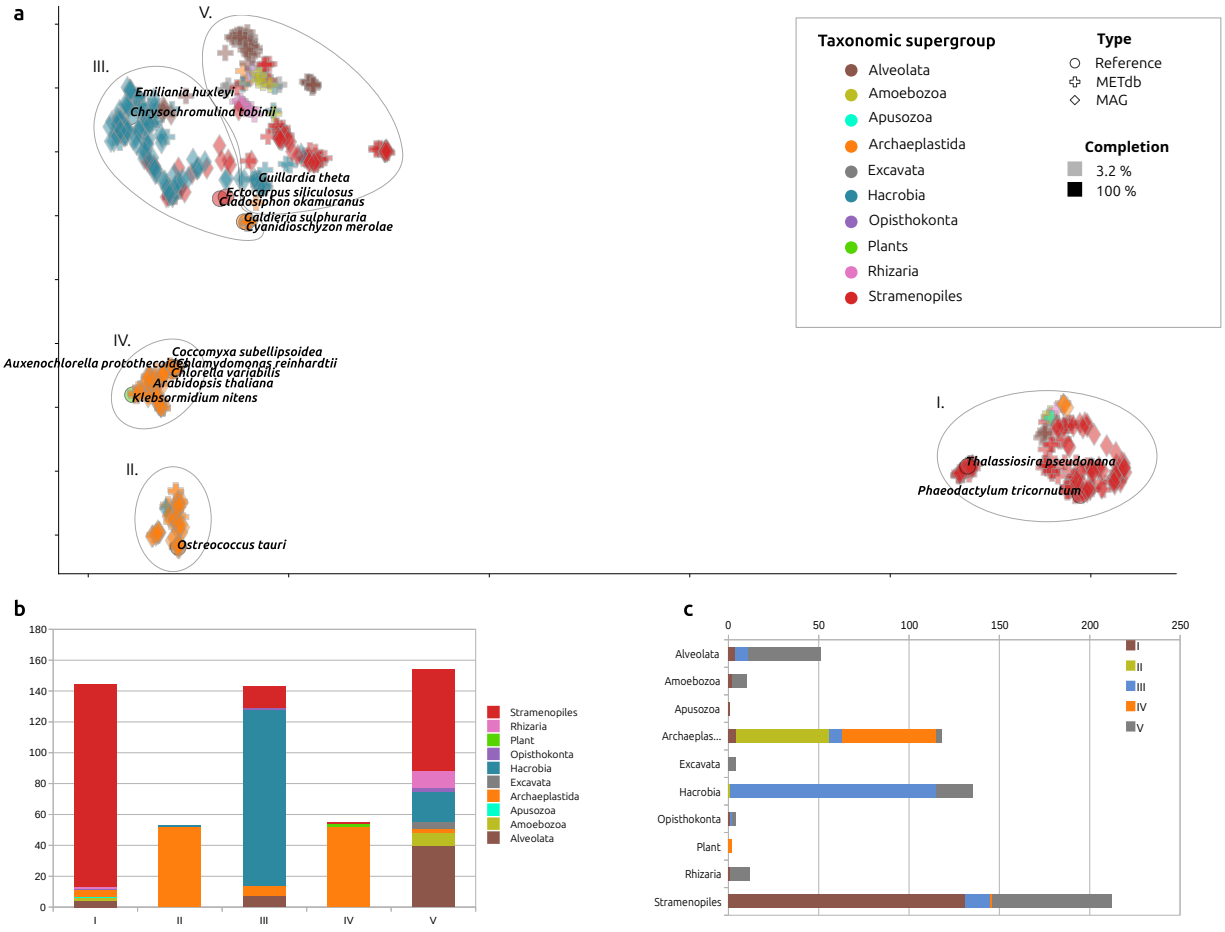
Supplementary Figure S2: Correlation matrix comparing original (iLB1034) and PhotoEukStein-derived GSMs reaction fluxes of Phaeodactylum tricornutum. The 434 shared reaction showing at least 1% or their correlations with other reactions fluxes greater to 0.2 when randomly sampling the whole metabolic space are displayed. Lower left rectangle: iLB1034 reactions, upper right rectangle, PhotoEukStein reactions. Only correlation with absolute value greater than 0.5 are coloured.



Supplementary Figure S3: HexBin representation of distribution of compared reaction fluxes correlations between PhotoEukStein and iLB1034 reference GSMs for Phaeodactylum tricornutum. Effective of correlations pairs of the 434 common reactions showing a possible correlation displayed in Figure S2 (see Materials and Methods section 3). Each cell indicates the number of reactions with corresponding correlations values pairs in iLB1034 (x-axis) and PhotoEukStein (y-axis) during random sampling of metabolic niche space.



Supplementary Figure S4: Compositional analysis of 512 PhotoEukStein-derived GSMs. Presence/absence of 9648 reactions were used to hierarchically cluster (euclidian distance, ward distance) both GSMs (lines) and reactions (columns). Genome source, Taxonomic group, model quality score (as defined in Extended Data), and metabolic cluster (as defined in Figure S5) are shown for each GSM. Frequency of reaction appearance among the 512 GSMs (bad quality models excluded) is indicated for each reaction (from white= 1 to dark red=512). Blue indicates significantly present reactions in a cluster, red if the absence of the reaction is signature reaction, grey if present but not significant may nevertheless share similar biological features



Supplementary Figure S5: Compositional analysis of the 549 PhotoEukStein derived GSMs. a) Diffusion map multiscale geometric analysis of GSMs topology followed by umap reduction of dimension processing have been performed to study the distribution of GSMs topologies in the functional space. Shapes indicate origin of the supporting genome (diamonds for MAGs, crosses for METdb and circles for references). Colours indicate main taxonomical groups, and transparency reflect genomes/transcriptomes completions estimations. Grey ellipses identify the 5 clusters supported by k-means signal deconvolution analysis (see Extended Data). b) Repartition of main taxonomic groups within each cluster. c) Distribution of each taxonomic group across the clusters.

Material and methods

1. Constraint-based metabolic modelling at genome-scale

Metabolic networks contain the metabolic capabilities encoded in the organism's genome. Indeed, from the genome of a specific organism, it is possible to predict the encoded genes and thus identify the corresponding enzymes and their associated metabolic reactions. Metabolic reactions are the set of life-sustaining chemical transformations in organisms. They allow organisms to grow and reproduce, maintain their structures, and respond to their environments. Because the products of some reactions are the substrates of others, the reactions are interconnected by what are called metabolites. Metabolic networks are modelled in order to study the physiology of the relevant microorganism. In particular, metabolic models are used to infer reaction rates, also known as fluxes, without using kinetic parameters. A metabolic model is formally described by its stoichiometric matrix S, where the rows correspond to the metabolites, the columns correspond to the reactions considered in the metabolic network. The entries are stoichiometric coefficients which are negative if the metabolite is a substrate, positive if the metabolite is a product and null if the metabolite is not implicated in the reaction. Assume there are m metabolites M_i to M_m and n reactions r_1 to r_n. We note respectively v_1 to v_n the fluxes of r_1 to r_n. Let to the stoichiometric coefficients of M_i in reactions r_1 to r_n. The change over time of the M_i concentration is given by the mass-balance equation:

$$\frac{d[M_i]}{dt} = v_1 + \ldots + v_n = \sum_{j=1\ldots n} v_j \text{, (Eq.1)}$$

Using a vector notation, the above equation can be written as:

$$\frac{dK}{dt} = Sv \text{, (Eq.2)}$$

where v stands for the fluxes vector, and K is the metabolites concentration vector. A metabolic network is formally described by its stoichiometric matrix $S \in R^{n,m}$ describing the relationship between the n metabolites and the m reactions. The entry is the stoichiometric coefficient of the metabolite M_i in the reaction R_j. By convention it is negative if the metabolite is a substrate, positive if the metabolite is a product and null if the metabolite is not implicated in the reaction.

In general, rate of reactions depends on metabolites concentrations and kinetic parameters, such as temperature, or pH. Determining these parameters and the function of reaction rate are complex experimental tasks. Moreover, these parameters are in general very sensitive to biochemical conditions, so *in vitro* determinations may not correspond with *in vivo* values (Edwards and Palsson, 2000). Thus solving Eq.2 is a daunting task for genome scale systems. When analysing metabolic networks using constraint-based approaches, we assume that organisms are homeostatic, keeping internal concentration as constant as possible by means

of regulation [1]. Thus the rate of formation of internal metabolites is equal to the rate of their consumption. The system is then considered in a quasi-stationary state [1], leading to:

$$Sv = 0. \quad \text{(Eq.3)}$$

In addition to this system of linear constraints, we also consider thermodynamic constraints on fluxes. Fluxes can be positive or negative. For inner reaction, a positive flux means that the reaction is occurring in its forward direction, whereas a negative flux means that it is occurring in the reverse direction. All fluxes must satisfy an inequality like:

$$lb_i \leq v_i \leq ub_i, \quad \text{(Eq.4)}$$

where lb_i represents the lower bound of the flux, and ub_i represents the upper bound of the flux. Fluxes are expressed in mole of product formed by gram of dry weight of the considered organism by hour (mol.gDW- 1.h- 1). Knowledge on the reaction direction and reversibility can also be encoded in those inequalities. For instance, if the reaction is known to be direct and irreversible, it means that the flux cannot be negative. Eq. 4 becomes:

$$0 \leq v_i \leq ub_i. \quad \text{(Eq.5)}$$

These equations result in a model described as a set of constraints. Altogether, Eq.3 and Eq.4 form a model called a constraint-based metabolic model (CBM) of the corresponding organism. A CBM at genome-scale is called a Genome-Scale Metabolic Model (GSM). It can be resumed in the system:

$$\begin{pmatrix} Sv = 0 \\ l_b \leq v \leq u_b. \end{pmatrix} \quad \text{(Eq.6)}$$

All solutions of v satisfy all constraints : 1) the steady state equation, and 2) the thermodynamic constraints, and thus define a steady- state flux space. This “flux space” may be further analysed through several state-of-the-art approaches. For a detailed review of these methods, the reader may wish to refer to [2-4]. A metabolic network and its associated GSM allows us to explore the metabolic phenotype of an organism [1].

The continuous supply of metabolites from and to the media is facilitated by exchange reactions. They are are responsible for uptake or secretion of metabolites. For boundary reaction, a positive flux means a secretion of the metabolite into the environment, whereas a negative flux means an uptake of the metabolite.

Finally, modellers developed fictive reactions to model the growth rate of organisms [5], and among them is the biomass reaction. This reaction encompasses the needs of the modelled system (nucleotides for DNA, RNA, amino acids for proteins, lipids, carbohydrates…), but also the energy cost of cellular division or cell maintenance.

2. PhotoEukStein : generic model reconstruction

Metabolic network of PhotoEukstein

PhotoEukStein was built from the merging of available biochemical and genomic information of 16 autotrophic eukaryotes (Figure 1.A). In the context of biological databases and data integration, using different identifiers for the same entity can create confusion and make it difficult to merge data from different sources. Thus, identifiers of reactions and metabolites are homogenized to the same namespace using MetaNetX [6] and manual curation. Duplicated entities are then removed. To enable seamless integration of the metabolic pathways of 16 different organisms into a single supraorganism, all enzymatic reactions were assumed to occur in a single compartment (exception see 2.B. below). All the reactions of the network are mass-balanced in order to predict reactions fluxes without relying on kinetics data (see Eq.6 above). Chemical formulae have been added to all metabolites using MetaNetX, manual curation and prediction algorithms. Duplicated reactions are deleted (those that propose similar metabolic transformations but have different identifiers). Some are modified based on the literature. When a reaction was not balanced, had no associated metadata, no associated genes either, or genes found only in *Arabidopsis or Okamuranus*, the reaction is most deleted.

Constraint-based metabolic model of PhotoEukStein

According to the general protocol of [7], the 'draft' model was then curated manually to generate a functional CBM of eukaryotic-algae metabolism.

In order to maintain the stationary state of the network, 674 sink reactions (SK) were added for metabolites consumed but never produced (with hard-constraint on the uptake flux : $-0.5 \leq v_{SK} \leq 0$), and 1033 demand reactions(DM) for those produced but never consumed ($0 \leq v_{DM} \leq 1000$¿. Sink and demand reactions are special reactions that allow us to maintain active metabolic pathways of which some knowledge is missing (they allow an active flux in 2,554 reactions. Indeed, if all SK andDM were blocked, they would disable 2,554 reactions of the network with them.The number of sink and demand reactions may be reduced in future versions of PhotoEukStein as new enzymatic reactions are discovered.

Directionality of some reactions have been corrected. For example, to avoid futile cycles or false proton gradients that could generate ATP out of nowhere, heuristic rules have been applied : reactions consuming ATP (except from respiration pathway), ABC transporters and proton pumps are irreversible. Blocked reactions and orphan metabolites were deleted to avoid false-negative analysis regarding gene deletion on flux redistribution.The photosynthetic system of PhotoEukStein is based on iLB1034 [8]. The addition of a pseudo-thylakoid and a chloroplast allows for a spatial organization that couples the photosynthetic apparatus, chloroplast ATPS, and carbon dioxide fixation by RuBisCo with light absorption, and powers the growth rate (see Extended Informations.

PhotoEukStein encompass 5,831 metabolites and 11,229 reactions.Two types of reactions are distinguished : 2,067 boundary reactions (including 360 exchanges reactions, 674 sink reactions, 1,033 demand reactions), and 9162 internal biochemical transformations. The meta-model has 15 biomass objective functions : 1 autotrophic BOF from *C. variabilis* [9] ; 3 (1

autotrophic, 1 hetereotrophic and 1 mixotrophic) from *C. reinhardtii* (iRC1080, BiGG, [10]) ; 11 (2 for biomass production during light or dark, and many for specific class of metabolites as DNA, RNA, lipids, carbohydrates production) from *P. tricornutum* (iLB1034 [8]). The BOF of iLB1034 is used for this manuscript.

Gene-Protein-Reaction rules

For each internal reaction in the curated universal model, we identify all those that are equivalent in the input models (i.e., duplicates) to recover the maximum number of logical gene conjunctions (monomeric, oligomeric, isoenzymes or multifunctional enzymes). Thus, 7,599 PhotoEukStein reactions (/9162) are associated to 20,468 protein sequences, from reference genomes[1],by their respective logical associations. Protein sequences are mostly retrieved from NCBI, UniProt, Diatomics and TAIR (*Arabidopsis thaliana* database).

17% of PhotoEukStein's reactions do not have associated genes either because the reaction is spontaneous, either no genes have been found yet to catalyse the reactions. PhotoEukStein can easily be extended to incorporate new metabolic knowledge to cope with the development of eukaryote phototrophs unicellular organisms studies, either through identifying new metabolic reactions, or accumulating reference protein sequences associated with a given reaction. For example, DMSP synthesis from methionine has been shown to take place via transamination pathway in some eukaryotic algae [11]. Although some of the models that make up PhotoEukStein (Figure 1.A) had the DMSP synthesis pathway (e.g. *Thalassiosira* [12], *Cladosiphon okamuranus* [13], or *Phaeodactylum* [8], none had a gene associated with the key enzyme of this pathway. However, two genes encoding for this enzyme in eukaryotic algae have been identified : (i) DSYB, and (ii) TpMT2 whose the function was confirmed in *T. pseudonana*[20]. We added 135 sequences to DSYB, and 6 for TpMT2 (from [11,14,15]) in the protein sequences database of PhotoEukStein. 337 models of the GSMs database can produce DMSP (Supplementary Table S7).

The genomic and biogeochemical information of *Thalassiosira pseudonana* included in PhotoEukStein comes from the PGDB of BioCyc (2012). The constraint-based model used to make the comparisons comes from a fairly recent publication [12]. Thus, unlike the other GSMs, the reference used to validate the PhotoEukStein-derive model of *Thalassiosira* is not included in PhotoEukStein (and therefore its BOF is also different). This may explain the greater differences in growth rates between the two models in Figure S1. However, the correlation being very high, we can see that the two models adapt to their environment in a rather similar way.

3. PhotoEukStein validation

Phototrophic phenotypes of PhotoEukStein

We ensure that PhotoEukStein can grow under photoautotrophic conditions with adapted physiological strategies like the ability to fix inorganic carbon. We expected a coupling between light uptake and CO_2 uptake from the environnement, and the underlying synchronization of photosystem reactions, ATP production by the chloroplastic ATP synthase,

as well as inorganic carbon assimilation by the ribulose-1,5-biphosphate carboxylase (RuBisCo).

Under photoautotrophic conditions (cf. Medium section below), we computed a projection of the allowable flux space on key reactions fluxes [16] of PhotoEukStein to scrutinise the fluxes variability and coupling of these key reactions, and thus assess some photoautrophic phenotypes for PhotoEukStein (see Extended Data).

Growth rates comparison

In order to validate PhotoEukStein, we compare our automatically reconstructed models with reference model of the same organism. The organisms were reconstructed with a medium that would allow the reference model to grow (see Medium section below).

When computing the niche space of models [16], we compare the flux through the biomass reaction for around 10^4 randomly generated environmental conditions. The environmental condition are composed of fixed fluxes of exchange reactions concerning the following metabolites: CO_2, photon, SO_4, NH_4, NO_3 or Phosphate. No other constraints were applied to the exchange reactions of the reference models. For the PhotoEukStein derived models, if the exchange reaction exist in the corresponding reference model, bounds are the same as the reference, else the lower bound is set to 0 except for the exchange reactions concerning H_2O, H, Mg^{2+}, Fe_2, Fe_3.

Sampling of the solution space

Reaction fluxes correlations

To compare further the models, we applied a sampling procedure of all the allowable solution space of 2 models of *Phaeodactylum tricornutum*. *Phaeodactylum tricornutum* original GSM (iLB31034, [8]) is composed of 2162 reactions, while PhotoEukStein-derived model (phaeo-photoeuk) is composed of 5366 reactions (Supplementary Table S2). The fluxes constraints of the exchange reactions of iLB1034 have been applied on phaeo-photoeuk. For each model, a sampling procedure have been applied (see https://cobrapy.readthedocs.io/en/latest/sampling.html, OptGPSampler, thinning=10,000, sample=10,000). Blocked reactions are removed, and the set of shared reactions are considered (1171 reactions). From those, fluxes correlations (pearson) for each pair of reactions were computed. Only the reactions having at least 1% of their absolute correlations being higher than 0.2 (and which are shared by the two models), were kept for the analysis (434 reactions, Figure S2). Python package Seaborn.heatmap have been used for the plot.

HexBins

From the 94,178 correlations (434x434/2 : upper or lower triangle of correlation matrix), we eliminated (1) the absolute correlations that were lower than 0.025 in both models, and (2) the 434 correlations of the diagonal, resulting in 69,442 remaining correlations. We compare values of these correlations with the HexBins.The x-axis corresponds to the correlation values

of phaeo_photoeuk, and the y-axis to iLB1034. For each correlation, we plot the result for each model. Python package Seaborn.jointplot have been used for the plot.

4. Exploration of PhotoEukStein-derived metabolic models DB

PhotoEukStein-derived models

Tara Oceans eukaryote MAGs resource

MAGs sequences (predicted CDS and their functional annotations) corresponding the (Delmont 2022) were downloaded from https://www.genoscope.cns.fr/tara/

The METdb database for eukaryote transcriptomes.

METdb is a curated database of transcriptomes from marine eukaryotic isolates that cover the MMETSP collection13 (new assemblies were performed, combining time points from the same culture in co-assemblies when available) as well as cultures from TARA Oceans [17]. The database is publicly available and can be accessed at http://metdb.sb-roscoff.fr/metdb/.

Identification of phototrophs MAGs or METdb

The subset of phototrophs MAGs and METdb was defined as those encoding proteins with the Chlorophyll A-B binding protein domain (InterPro entry IPR022796)

Deriving GSMs from PhotoEukStein

CarveMe can be easily installed using the pip package manager. Additionally, diamond package and IBM CPLEX Optimizer need to be installed manually (see https://carveme.readthedocs.io/en/latest/installation.html) .

To use PhotoEukStein with CarveMe, one need to download

- (1) the generic model,
- (2) the Gene-Protein-Reactions associations files,
- (3) the protein sequences database and
- (4) the media file. The sbml.py from cobra package need to be changed to support the reading of identifiers from BioCyc. Please read the associated
- (5) README file for more information on how to proceed.

The organisms where reconstructed with a medium that would allow the reference model to grow (see supp Mat medium). For the reconstruction of PhotoEukStein-derived models the following code have been used:

carve -d -v path/to/input/fasta.faa --universe photoeukstein --gapfill medium_name --output

path/to/output/model.xml

« Medium_name » is « phaeo » for the whole model reconstruction except for *the computation* of growth rate comparison with references for *Chlorella_variabilis* (medium « chlorella »), *and Thalassiosira_pseudonana* (medium « thalassio »)

Compositional analyses

Presence/absence of 9648 reactions (those mobilized by model resources) were used to hierarchically cluster (euclidian distance, ward distance) both GSMs (lines) and reactions (columns). Genome source, Taxonomic group, model quality score (as defined in Extended Data), and metabolic cluster (as defined in Figure S5) are shown for each GSM. Frequency of reaction appearance among the Figure S4 .We used the UMAP algorithm [18] to visualize how the presence/absence of the 9648 reactions (those mobilized by the model's resources) helps structure the models together. A k-means clustering was used to identify the clusters used in Supplementary Figures S4 and S5.

Topological analysis

We analyzed the dataset through the algorithm developed in [19] . We used the GSM reconstructed with their sink reaction, however only the internal reactions are considered. All the diffusion variables are then normalized, and we used the UMAP algorithm for better visualization [18].

Functional analysis

In order to assess functional phenotypes (towards new functional traits) of organism, we consider a basic medium (Supplementary Table S4), and we add a set of new nutrient (up to 9, Supplementary Tables S4 and S5). For each condition we maximize the flux through the biomass reaction. The set of new nutrient result from the use of *itertools.combinations* algorithm in the Python library, on the list of considered nutrient (Supplementary Table S5).

Anvio representations

All anvio-based representations (FIgures 1 and 2) were generated using anvio version 7.1 (http://anvio.org/) [20] from data available in Supplementary Tables S2 and S3.

Growth medium for GSMs

For each reference models (*Chlorella*, *Thalassiosira*, *Phaeodactylum*), we have retrieved their respective medium. To achieve this, FVA analysis have been performed. Then, for each exchange reaction, if the flux interval indicates values less than 0, then the metabolite can be imported into the system. It is therefore part of the medium of the organism considered.

References

1. Varma, A. & Palsson, B. Ø. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and Environmental Microbiology* **60**, 3724–3731 (1994).
2. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* **2**, 886–897 (2004).
3. Lewis, N. E., Nagarajan, H. & Palsson, B. Ø. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* **10**, 291–305 (2012).
4. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* **15**, 107–120 (2014).
5. Xavier, J. C., Patil, K. R. & Rocha, I. Metabolic models and gene essentiality data reveal essential and conserved metabolism in prokaryotes. *PLOS Computational Biology* **14**, e1006556 (2018).
6. Moretti, S., Tran, V. D. T., Mehl, F., Ibberson, M. & Pagni, M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Res* **49**, D570–D574 (2021).
7. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* **5**, 93–121 (2010).
8. Broddrick, J. T. *et al.* Cross-compartment metabolic coupling enables flexible photoprotective mechanisms in the diatom Phaeodactylum tricornutum. *New Phytologist* **222**, 1364–1379 (2019).
9. Juneja, A., Chaplen, F. W. R. & Murthy, G. S. Genome scale metabolic reconstruction of Chlorella variabilis for exploring its metabolic potential for biofuels. *Bioresource Technology* **213**, 103–110 (2016).
10. Chang, R. L. *et al.* Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Molecular Systems Biology* **7**, 518 (2011).
11. Curson, A. R. J. *et al.* DSYB catalyses the key step of dimethylsulfoniopropionate biosynthesis in many phytoplankton. *Nature Microbiology* **3**, 430 (2018).
12. van Tol, H. M. & Armbrust, E. V. Genome-scale metabolic model of the diatom Thalassiosira pseudonana highlights the importance of nitrogen and sulfur metabolism in redox balance. *PLOS ONE* **16**, e0241960 (2021).
13. Nègre, D. *et al.* Genome–Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae Saccharina japonica and Cladosiphon okamuranus. *Antioxidants* **8**, 564 (2019).
14. Kageyama, H., Tanaka, Y., Shibata, A., Waditee-Sirisattha, R. & Takabe, T. Dimethylsulfoniopropionate biosynthesis in a diatom Thalassiosira pseudonana: Identification of a gene encoding MTHB-methyltransferase. *Archives of Biochemistry and Biophysics* **645**, 100–106 (2018).
15. O'Brien, J. *et al.* Biogeographical and seasonal dynamics of the marine Roseobacter community and ecological links to DMSP-producing phytoplankton. *ISME Commun* **2**, 16 (2022).
16. Régimbeau, A. *et al.* Contribution of genome-scale metabolic modelling to niche theory. *Ecology Letters* **25**, 1352–1364 (2022).
17. Niang, G. *et al.* METdb: A Genomic Reference Database For Marine Species. *F1000Research* **9**, (2020).

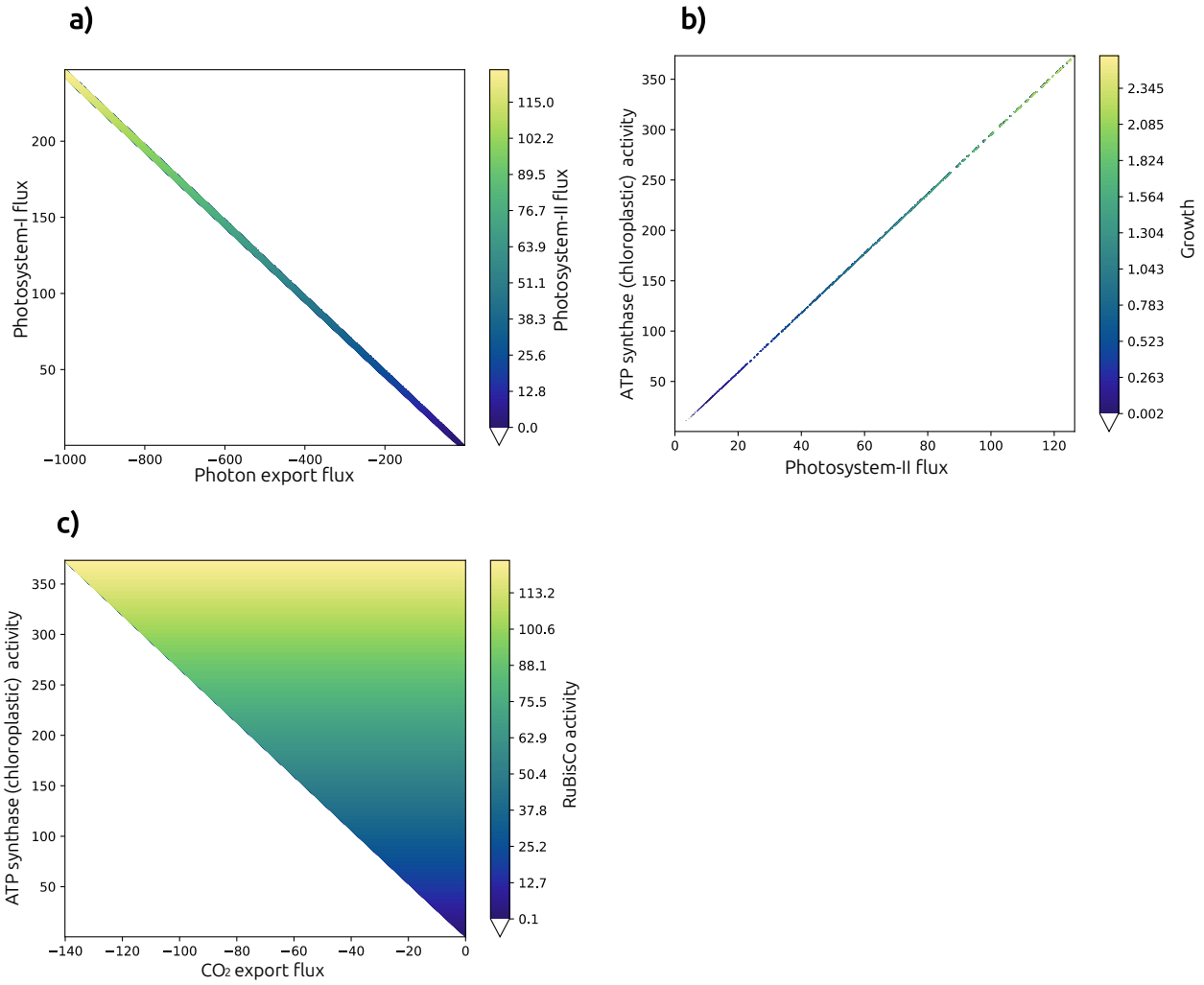
18. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
19. Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nat Commun* **11**, 4887 (2020).
20. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**, 3–6 (2021).

Extended Data

1 - PhotoEukStein validation

In an epistemological context, model validation and sensitivity analyses are critical steps to ensure the robustness and reliability of the model's predictions. The validation process generally consists of comparing the model's predictions with observed data, the literature or other reliable models, and thus assessing the model's ability to reproduce known phenomena. Then, one can use this model to test new hypothesis and predict future outcomes for which one does not yet have empirical values. When a generic model is converted to ready-to-use organism-specific models using CarveMe, the whole manual curation and relevant structural properties are preserved (Machado 2018).

Therefore, we ensure that PhotoEukStein can grow under photoautotrophic conditions (see Medium M&M) with adapted physiological strategies. Indeed, the ultimate goal of photoautotrophic organisms is to use light energy to convert water and carbon dioxide into oxygen and energy-rich organic molecules such as glucose or starch. Therefore, we expected a coupling between light uptake and CO_2 uptake from the environment, and the underlying synchronization of photosystem reactions, ATP production by the chloroplastic ATP synthase, as well as inorganic carbon assimilation by the ribulose-1,5-biphosphate carboxylase (RuBisCo being the key enzyme of Calvin cycle). In order to validate the basic internals of PhotoEukStein, and more specifically the phototrophy-associated reactions, we computed a projection of the allowable solution space (Régimbeau et al., 2022) on these key photoautotrophic reaction fluxes to determine their distribution and couplings. The more photons enter the system, the more the photosystems are stimulated with a synchronization of the two photosystems (figure A). We also see that the ATP production by chloroplastic ATPS is coupled to the photosynthetic activity and fuels the growth reaction (figure B). This ATP production is also coupled to CO_2 uptake and the activity of RuBisCo (figure C). Overall, the uptake of photon into the system stimulates the photosystem apparatus (PSII, PSI) and empowers ATP production. The ATP allows the CO_2 fixation by RuBisCo and fuels the biomass production.

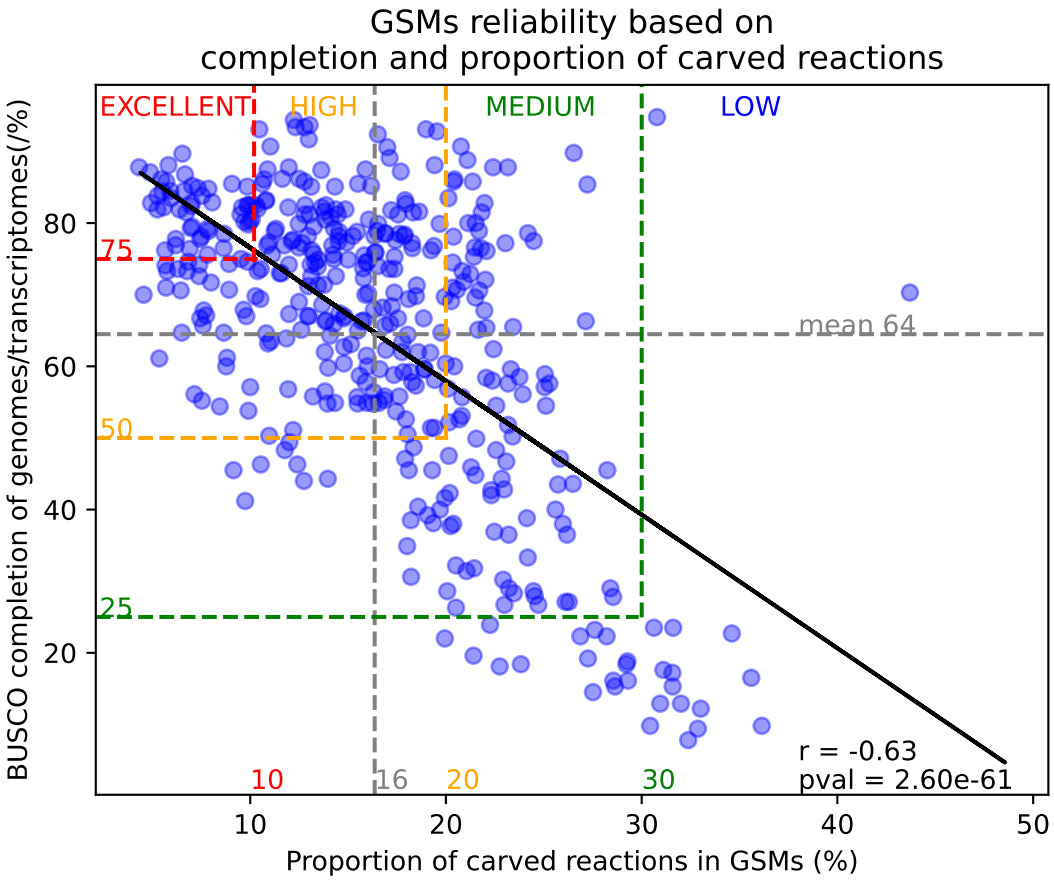


2 - Definition of PhotoEukStein-derived GSMs quality

To ensure the biological reliability of the generated models produced by PhotoEukStein, we made sure to classify them into 4 categories. We based our classification on the genome/transcriptome completion, and the frequency of carved reactions (reactions imported without direct genetic evidence). Quality thresholds for PhotoEukStein-derived GSMs is based on the following criteria (c.f. Supplementary Table S2).

- Excellent: ≥ 75% completion AND 10% carved reactions
- High: ≥ 50% completion AND < 20% carved reactions
- Medium: ≥ 25% completion AND < 30% carved reactions
- Low: the rest

The following figure shows the distribution of PhotoEukStein-based GSMs quality



4 PHOTOEUKSTEIN PAVES THE WAY FOR MECHANISTIC MODELLING OF PHOTOTROPHIC MICROEUKAYOTE METABOLISM

4.1 A BALANCE BETWEEN OVERBURDENED AND OVERSIMPLIFIED MODELLING OF BIOLOGICAL SYSTEMS

« We sometimes seem to have forgotten that the original question in genetics was not what makes a protein but rather what makes a dog a dog, a man a man. », Noble [195]

4.1.1 Gene ontology will fail without higher-level insight

4.1.1.1 A mechanistic causal chain perpetuating shortcuts

In order to understand biological systems, it is necessary to decipher the relationship between the genome and the phenotype[***]. A habit of biologists is the characterisation of phenotypes since this applies to human health[196] (such as endophenotypes), to crop productivity[254], or to ecosystem monitoring (biomarkers)[100], for example. The relationship was initially thought to be simple. For each inheritable « phenotypic character », there was postulated to be a discrete genetic element (a gene) transmitting it through the generations. This reducing approach of molecular biology and genomics is often understood as a mechanistic causal chain perpetuating shortcuts such as "the gene(s) X for trait Y"[†††]. Thus, it does not really matter which way one looks at it, genotype and phenotype are effectively equivalent from this view at least[197,198]. Indeed, the phenotype is often an imperfect indicator of the genotype : the same genotype may give rise to a wide range of phenotypes, and the same phenotype may have arisen from different genotypes[‡‡‡].

4.1.1.2 Ambiguous functional labels to genes based solely on the proteins they encode

It's crucial to recognize that high-level biological functions often involve the coordinated activity of numerous genes, up to hundreds or more (a phenomenon coined as polygeny)[179]. Similarly, individual genes can participate in multiple functions (pleiotropy). This complexity makes it difficult to assign unambiguous functional labels to genes solely considering proteins they encode. This conventional approach of gene labeling is limited because it does not directly address the higher-level phenotype characteristics that researchers are often interested in understanding. Therefore, assuming that a

[***] For example : A character/trait being «CO_2 fixation », the function being « to produce organic molecules », and phenotypes showing differences within this function-valued trait (2.2.2.2).

[†††] This does not exclude the need to search for functional trait marker genes.

[‡‡‡] Wilhelm Johanssen introduced the word “gene.” His research on self-fertilised lines of beans revealed that quantitative variability in the phenotype confounded thinking about separable contributions of heredity and environment. He introduced this non-linearity between genotype and phenotype.

gene represents its sole or primary function would be an oversimplification (4.1.1.1). Biological function emerges from complex interactions between proteins and other cellular components (4.1.2). Higher-level insight and a systems-level perspective are essential to comprehensively unravel the complexities of gene function and their contributions to phenotypic characteristics. It requires considering the logic and principles operating at various levels, not solely focusing on the lower levels. Moreover, much of the logic of living systems is found at higher levels, since it is often at these levels that selection takes place[199,200], and determines whether organisms live or die (based on their fitness for their environment). Each level has its own integration of functions, and it is the task of biologists to determine at which level a specific function is integrated.

4.1.1.3 Existence of feedforward and feedback loops between different levels of biological organisation

Multilevel modelling in biology recognises that causation operates in both upward and downward directions, meaning that genetic factors influence higher-level biological processes, and *vice versa*[195]. This understanding highlights the existence of feedforward and feedback loops between different levels of biological organisation. According to the central dogma in molecular biology[201], information flows from DNA to RNA, then to proteins, and subsequently to higher levels of biological organisation. This view excludes the possibility of information flowing in the opposite direction, which is associated with Lamarckian inheritance (the inheritance of acquired characteristics). This dogma of the unidirectional transmission of information is considered incomplete in at least two respects. Firstly, it defines relevant information solely in terms of the DNA code, neglecting other factors that influence biological processes. For example, the DNA sequence determines which protein will be synthesised, but it does not specify the quantity of each protein produced. Secondly, the dogma assumes that knowing enough about genes and proteins would be sufficient to reconstruct all other levels of biological organisation, implying a bottom-up approach in systems biology (a reductionist causal chain). Thus, this view overlooks the existence of complex control mechanisms and the robustness of biological systems, meaning they can maintain stability and functionality despite perturbations or variations. This robustness suggests the presence of control mechanisms beyond the simplistic linear flow of information. While the exact nature of these control mechanisms is still not fully understood, their existence is apparent in the ability of biological systems to adapt, respond, and maintain stability. It emphasises the need to consider additional factors and control mechanisms that contribute to the robustness and complexity of biological systems.

One of the challenges in multilevel modelling is developing appropriate mathematical and computational tools to handle these complex causation loops. Each level of biological organisation may require different mathematical approaches, and connecting these levels is not a straightforward task. It requires careful consideration of the biological insights to determine the relevant level of detail at one level that influences functionality at other levels. Achieving a comprehensive understanding of the complex relationships between biological levels remains a significant challenge in the field.

In fact, no single level in biology can be considered privileged, and identifying the level at which functions are integrated is an important aspect of biological research.

4.1.1.4 *The genome as the "book of life" is only a metaphor*

The concept of a genetic program, as originally proposed by Monod and Jacob[202,203], drew an analogy between the digital code of DNA and the sequences of instructions in a computer program. It is essential to recognize that these metaphors should not be taken too literally, as they can fuel misconceptions of genetic determinism[195]. While it is a useful metaphor, it implies that only coded information is important, as seen in the notion of the genome as the "book of life", and may lead to gene-determinism: « They [genes] created us body and mind »[198]. Instead, genomes serve as a database of information used by the biological system as a whole. Modern molecular biology, starting with Watson and Crick's work, has made significant progress in mapping DNA sequences to amino acid sequences in proteins. However, protein-coding DNA accounts for only a small portion of the genome, and there are various mechanisms involving non-coding DNA that influence gene expression and phenotype. Not to mention epigenetics, we also need to extend our search beyond the genome, because this « gene-centric view » is limited in explaining the complexity of phenotypes. The relationship between genomes and phenotypes is far more intricate and influenced by non-genetic factors than previously thought.

4.1.2 Genotypes-environments (GxE) - phenotypes relationships

4.1.2.1 *The environment outside the system*

All levels of biological organisation are influenced by the external environment. It has been reported various cases where a genetic difference is not visible at the phenotypic level due to environmental influences[204]. To give a well-known example, the red-white genetic difference in the color of Primrose flowers is no longer visible when plants are grown at 30°C-35°C because at high temperatures all flowers are white. As another popular example, Waddington knew from his developmental studies that fruit flies embryo could display different thorax and wing structures, simply by changing the environmental temperature or by a chemical stimulus (Figure 36 from[205–207]). To come back to plankton, diazotrophs have the genetic ability to fix the diatomic gas N_2 as nitrogen source through nitrogenase enzyme. It may thus be tempting to automatically assume that they always fix N_2, however the expression of nitrogenase occurs only when the diazotrophs cannot attain sufficient nitrogen from other inorganic sources such as NH_4. This functionnal trait depends on the environmental conditions and is an acclimative event[208]. We will also discuss the production of DMSP under nitrogen stress in *Phaeodactylum tricornutum* (4.2.2.1). In addition to these abiotic factors, metabolic interdependencies with other organisms (such as cooperative relationships) also allow particular phenotypes to emerge (4.2.2.2 and 4.2.3).

4.1.2.2 *The immediate surroundings*

Immediate environment of the system, such as structural information and evolutionary history, are critical components for comprehending the complexity of biological systems and their behaviours. These constraints are encoded in both DNA sequences and the inherited cellular architecture. The strong version of the « gene-centric view » (as introduced here 4.1.1.4) suggests that the complete structure of an organism is somehow encoded in the genetic information. However, this view is deemed implausible and unsupported by current understanding. Indeed, DNA is not the sole carrier of heredity[195]. While DNA sequences determine the amino acid sequences in proteins, the cellular

architecture influences their locations, movements, and interactions[209]. Cellular machinery, including mitochondria, endoplasmic reticulum, microtubules, membranes, and specific chemical arrangements within compartments, also determine protein behaviour[210] (as we have seen in 2.2.2). These inherited components are not primarily dictated by DNA sequences. Genes do not need to encode every aspect of cellular function. For instance, lipids, essential for cell structure, are not encoded by DNA and are part of what we call the « membranome »[211].

Eukaryotic cells, in particular, are highly structured, with membranous organelles and other compartments that contribute to their complexity. It is not simply a bag formed by a cell membrane enclosing a protein soup. Even prokaryotes, once thought to lack structure, have been found to possess organisation[212] and compartmentalisation[213]. The biophysical properties and self-organisation processes of molecules and structures play significant roles in phenotypic development.

The question of cytoplasm inheritance, which refers to the influence of non-DNA components on inheritance, has a long history in biology[195]. While early theories of cytoplasm inheritance were largely disproven, it is now acknowledged that the cellular machinery play important roles in inheritance. The limited success of cross-species clones (nuclear transplantation in egg) in developing to the adult stage suggests that the complex architecture of the cytoplasm may have a greater impact on development than previously realised[214–216]. To illustrate my point, let's look at the study conducted by Sun *et al*.[214] focusing on cross-species cloning involving goldfish eggs and carp nuclei. The process began with the enucleation of fertilised goldfish eggs, which means the removal of the nucleus from the egg. Subsequently, a nucleus from a carp was inserted into the enucleated goldfish egg. The outcome of this cloning procedure resulted in the development of adult fish with an overall body structure that exhibited intermediate characteristics. If a carp were generated, it might suggest that DNA is the primary and privileged information (what most « genetic determinists » would expect). Conversely, if a goldfish were generated, it would challenge the notion of DNA primacy[217]. However, the outcome is an hybrid (or non-viable organisms if the species are too far apart phylogenetically). Thus, the non-genetic structural information inherited by cells plays a crucial role in development and the realisation of phenotypes. This structural information is not solely determined by genes but interacts with genetic information to shape phenotypic outcomes.

Much of the evolution of cellular structures may have occurred independently of the cell's own DNA, particularly during the early evolution of eukaryotic cells, which involved various forms of endosymbiosis. A well-known example is that of chloroplasts originating from the engulfment of free-living cyanobacteria by a eukaryotic host cell in a process called endosymbiosis[255]. This event likely occurred more than two billion years ago, giving rise to the first photosynthetic eukaryotic cells (as discussed here 1.1.3). Over time, the engulfed cyanobacteria evolved into specialised organelles within the host cell[218,219] . The organelles retain some of their original prokaryotic DNA, although some genes have migrated to the nucleus (an evolutionary process called endogenosymbiosis[§§§]).

4.1.2.3 The genetic determinism is obviously fragmentary

The concept of genetic determinism, which assumes that genetic information alone determines the behaviour of a biological system, is fragmentary because it overlooks the complex relationship between genotype and phenotype. Understanding phenotypic traits sometimes requires taking into

[§§§] It is considered a specific form of gene transfer, but it differs slightly from horizontal gene transfer as it is generally understood.

account complex interactions between a wide variety of components, including proteins, but also non-genetic factors such as cell architecture, biotic and abiotic environmental factors. All of these parameters play significant roles in shaping phenotypic outcomes, highlighting the limitations of a « gene-centric » perspective.

We can come close to completely characterising a genome but not a phenome[****], because the information content of phenomes dwarves those of genomes : phenotypes vary from biological scale to scale, from cell to cell, and from moment to moment, and therefore can never be completely characterised[220]. The concept of a genotype-phenotype (G-P) map is a widely used metaphor for the multiple ways in which genotypic information influences the phenotype of an organism. Indeed, phenotypic variation arises from intricate interactions between genotypes and environments. An early version of the G-P map concept are the epigenetic landscape of Conrad Hal Waddington (Figure 36 from[205–207]) which inspired me a lot but also and especially the biologists working on cell fate specification and the possible use of stem cells for biotherapy[221,222]. The continuous and multivariate nature of most phenotypes suggests that categorical phenotyping loses information[223]. They are often best thought of as a function-valued trait, rather than as discrete measurements that can be used to capture the shape of the function[224].

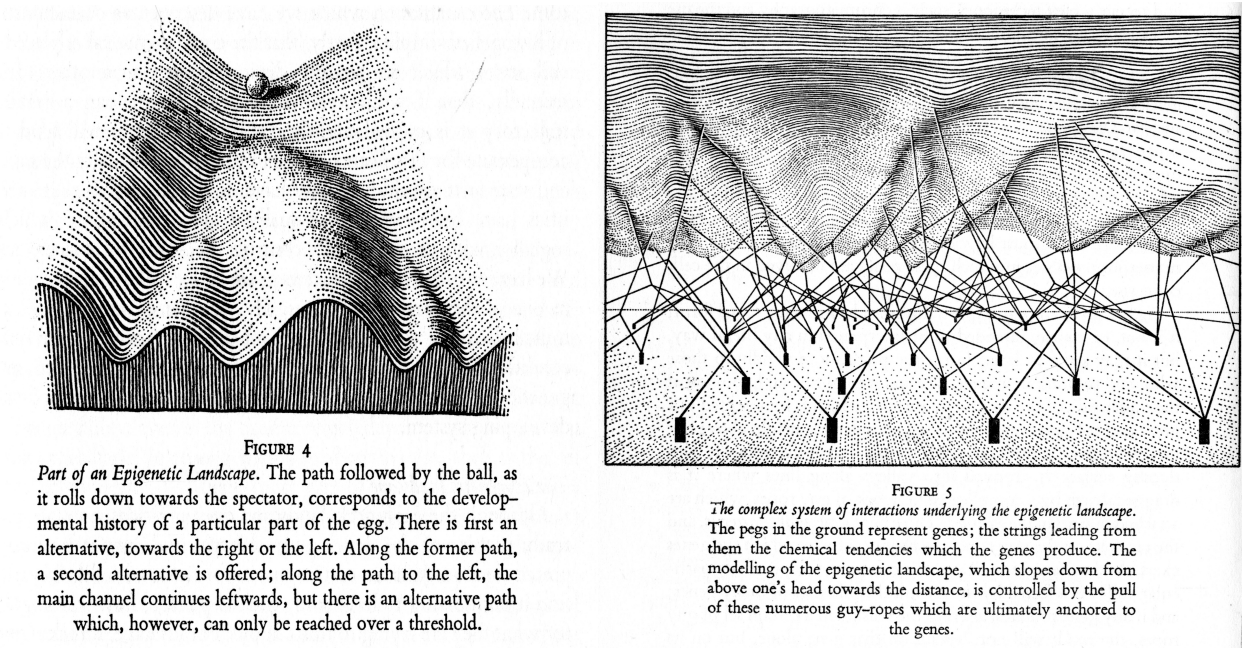


FIGURE 4

Part of an Epigenetic Landscape. The path followed by the ball, as it rolls down towards the spectator, corresponds to the develop–mental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold.

FIGURE 5

The complex system of interactions underlying the epigenetic landscape. The pegs in the ground represent genes; the strings leading from them the chemical tendencies which the genes produce. The modelling of the epigenetic landscape, which slopes down from above one's head towards the distance, is controlled by the pull of these numerous guy–ropes which are ultimately anchored to the genes.

Figure 36: The epigenetic landscape in the course of time by Conrad Hal Waddington (The Strategy of the Genes, Waddington, 1957)

A. Waddington knew from his developmental studies that embryo fruit flies could show different thorax and wing structures, simply by changing the environmental temperature or by a chemical stimulus. In his landscape diagram, this could be represented as a small manipulation in slope that would lead to one channel in the landscape being favoured over another, so that the adult could show a different phenotype starting from the same genotype.

B. Genes (solid pegs at the bottom) are viewed as parts of complex functional networks so that many gene products interact between themselves and with the environment to produce the phenotypic landscape (top) through which development occurs. Waddington's insight was that new forms could arise through new combinations to produce new landscapes in response to environmental pressure, and that these could then be assimilated into the genome.

[****] A phenome would encompass all observable traits and characteristics of an individual or organism.

4.1.3 Overburdened modelling can trigger the « error cascade »

4.1.3.1 Models are simplifications of reality

This following part echoes this one 2.2.1.

As we strive to model and understand biological systems, we encounter the inherent complexity that they possess. However, it is crucial to acknowledge that the more detailed and intricate our models become, the greater the potential for introducing inaccuracies and uncertainties, leading to a « propagation of errors » that can have cascading effects. This phenomenon highlights the interconnected nature of biological systems. Even small errors or uncertainties in one component of the model can have amplified effects on subsequent calculations, ultimately leading to erroneous outcomes (because our models are not as robust as biological systems[195]). To mitigate the risks associated with error propagation, scientists employ various strategies. Rigorous data collection (2.1.1.1) and validation (2.2) are essential to ensure the accuracy of the inputs. Sensitivity analyses help identify key parameters or variables that significantly influence the model outcomes (2.2). Model validation against empirical observations provides a critical checkpoint for assessing the model's reliability (what we have started to do here 4.2.2.1). But above, it is necessary to simplify biological systems when modelling them.

All models, whether conceptual (simply to think for example), mathematical / computational, or experimental / clinical, are simplified representations of reality, offering valuable insights but unable to fully capture all the intricacies of natural systems. Even within species, there is considerable individual variability. A model based on an individual cannot perfectly represent the entire species. The complexity of biological systems makes it challenging to identify causal interrelations accurately. Models may struggle to capture all the intricate interactions and feedback loops present in biological processes, limiting their ability to predict outcomes accurately. But, an overly comprehensive model would lose its advantages. It would become overburdened with excessive complexity, making it difficult to simplify or explain specific phenomena (in contrast to this conceptual model 2.2.1.1). Such an overburdened model would not offer targeted assessment of hypotheses or provide practical utility. Like tools in a toolbox, each model has inherent limitations and specific utility. Different models serve different purposes and are designed to address specific aspects of the biological system under investigation.

Striking a balance between necessary details and excessive complexity is key. Scientists must approach model interpretation with caution, acknowledging the complexities of biological systems and the limitations of our current understanding. Although, GSMs do not take into account kinetics parameters, all the layers of regulation of protein activity such as epigenetics, protein-protein interactions, microRNA, or their catalytic properties, they have indeed proven to be powerful tools in systems biology, providing valuable insights into genotype-phenotype mapping and addressing various biomedical and environmental challenges.

4.1.3.2 *Irrefutability is not a virtue of a theory*

"A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often thinks) but a vice", Karl Popper[225]

When a hypothesis is validated and found to be in agreement with the anticipated outcome, it does not necessarily generate new insights. It merely confirms what was already expected or hypothesised. The rejection of a hypothesis, often viewed as a less desirable outcome, actually leads to new insights and progress[226]. When our best-conceived predictions are proven wrong, it highlights shortcomings in input data, their interpretation, and the hypothesis itself. This stage of the scientific process is where new insights are generated, limitations are identified, and future research directions are determined.

4.2 FROM INDIVIDUAL-BASED TO TRAIT-BASED MODELS

This part suggests some ideas for future modelling which becomes accessible thanks to PhotoEukStein. The results presented are still preliminary, but give good illustrations of the potentials of GSMs to better understand biology. I have organised them according to the type of modelling : individual-based or trait-based approaches.

4.2.1 Very short introduction on these two modelling approaches

Individual-based models explicitly represent individual organisms as objects with specific characteristics or traits. These traits influence interactions with other individuals and the environment[256] (Figure 37). These models are closely connected to trait-based approaches (see below for definition), as traits play a mediating role in interactions within individual-based models. For example, studying DMSP biosynthesis and its regulatory processes at the scale of an individual (as developed here 4.2.2.1) falls between two stools (between trait-based and individual-based modelling) : the modelling explicitly represents both an organism and its specific traits (individual-based), and both allow to study the trait in question and try to understand the combination of response traits that could influence this effect trait (trait-based). The bottom-up approach of individual-based models allows population-level behaviour to emerge from these individual interactions[256] (for example 4.2.2.2).

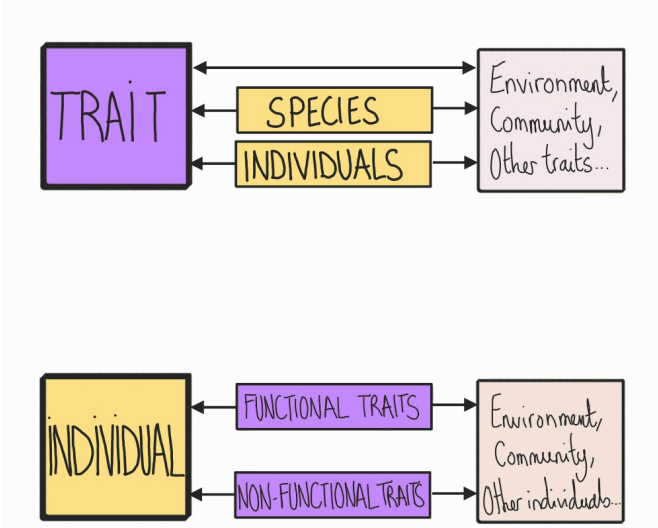


Figure 37: Trait-based modelling (upper part) differs from individual-based modelling (lower part) in the main entities of the models (traits or individuals, respectively) and in the ways interactions are represented (arrows). Figure based on [256].

Trait-based modelling primarily examines the effects and responses of traits themselves, potentially involving trade-offs (we have mentioned this kind of modelling here 1.2.3.3). Trait-based modelling can include species as carriers of traits, but they can also function without explicitly modelling species (for instance 4.2.3). In essence, trait-based models consist of combinations of functional traits that respond to environmental changes (response traits) and affect the properties of communities and ecosystems (effect traits)[256].

Implementing trait-based approaches in modelling can help overcome the data demand of individual-based models and has the potential to reduce computing times. Additionally, the incorporation of traits in modelling facilitates the scaling of physiological processes to global scales since traits can serve as a universal currency across different scales in these models.

4.2.2 Individual-based modelling

4.2.2.1 DMSP study at molecular and physiological scales

The main distinction between a biologist who utilises mathematical modelling and one who does not is that the former quantitatively explores the implications of their ideas, including conducting computational experiments to assess their plausibility. The potential benefits of such an approach are evident, as quantitatively plausible predictions enhance subsequent hypothesis-driven experimental research.

DMSP biosynthesis would contribute to the regulation of C:N:S ratios

When Phaeodactylum cannot export DMSP, its growth drops

To assess the validity of PhotoEukStein-derived GSMs, we reconstructed models in order to compare them with those of expert-based GSMs (see 2.2.3.1 for details). At our first attempt, we obtained this result for *Phaeodactylum* models (Figure 38).

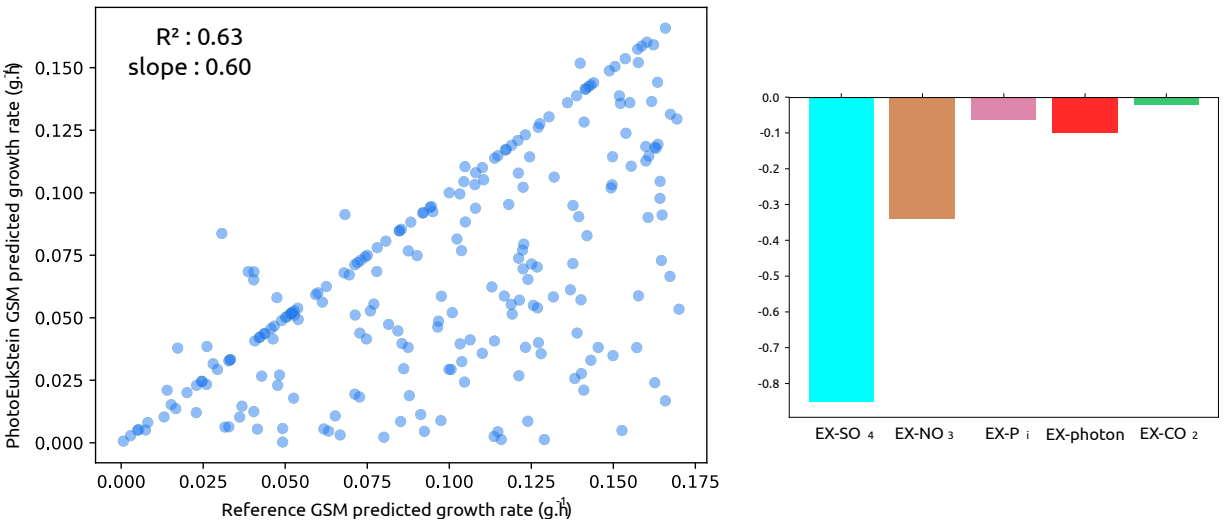


Figure 38: Comparison of predicted growth rates from PhotoEukStein or reference GSMs for Phaeodactylum tricornutum. In each case, 10,000 iterations of random sampling within each GSMs' niche space was performed and growth rates predicted for both models and reported. On the right are indicated the metabolites exchange reaction the most correlated with predicted growth rates differences between PhotoEukStein and reference GSMs. In this PhotoEukStein-derived model, the DMSP anabolism pathway is missing, highlighting the potential importance of DMSP in Phaeodactylum for the removal of excess sulphur and energy.

The R^2 between the two models growth rates (left) was weak (=0,63), showing that the PhotoEukStein-derived GSM may not capture all fundamental biological knowledge of the reference model. Moreover, exploring metabolic niches allows to assess the metabolic exchange fluxes differentiating growth rates between PhotoEukStein-based and reference GSMs (right). We see that SO_4^- uptake favours growth of the reference GSM (cor = -0,8). In other words, the sulphur seems to be a poison for our model, pointing out potential missing metabolic reactions leading to the emergence of somewhat different functional strategies.

Our model initially did not include the anabolism pathway for DMSP due to missing genes in the PhotoEukStein dataset. However, this omission turned out to be an interesting finding. It's worth noting that some models within PhotoEukStein do include the DMSP synthesis pathway, but none of them have a gene associated with MHM, as detailed in section 2.1.3.2. To address this, we added 135 sequences for DSYB and 6 for TpMT2 from litterature to the protein sequences database of PhotoEukStein. The TpMT2 sequence from *Thalassiosira pseudonana* showed high similarity to PtMT of *Phaeodactylum tricornutum* CCAP 1055/1, suggesting its potential role as a putative MHM[21]. By incorporating the protein sequences associated with DMSP synthesis reactions into PhotoEukStein, we successfully imported the pathway during reconstruction. As a result, after running the digital experiment again, we obtained the new result presented in our paper (3.2 Supplementary Figure S1). This experiment highlighted the potential importance of DMSP in *Phaeodactylum* for the removal of excess sulphur and energy (see below).

DMSP production under nitrogen stress ?

Many factors can affect DMSP biosynthesis like light, salinity, or temperature, depending on its physiological functions (1.1.4.2). Besides them, other factors also appear to affect cellular DMSP quotas, but the exact regulatory mechanisms are still unclear. A hypothesis is presented in which DMSP production is described as an overflow mechanism for excess reduced-carbon and -sulphur compounds[246]. In higher plants, there is a reciprocal regulatory coupling between the pathways of assimilatory sulphate and nitrate reduction to maintain appropriate proportions of amino acids for protein synthesis[257]. However, it has been observed that N-limitation can lead to increased DMSP production in many DMSP-producing algae and plants, resulting in higher sulphur incorporation relative to nitrogen incorporation[258]. Interestingly, DMSP does not contain nitrogen (see Figure 8 for reminder). The overflow mechanism can be seen as a response of the cell under conditions of unbalanced growth, producing and discarding compounds to ensure the continuation of other metabolic pathways (see hypothesis (1) in Figure 43). This mechanism allows continued sulphate assimilation even under nitrogen-limited conditions. Thus, increased excretion into the medium may serve as a way to dissipate excess sulphur and carbon.

Therefore, we compared the ability of *Phaeodacylum tricornutum* to produce DMSP[131] under nitrogen stress both *in silico* (Figure 39) with the PhotoEukStein-derived model, as well as *in vivo* (Figure 40) with the alga-culture performed in the Genoscope. The results below are preliminary and further exploration is necessary before drawing any conclusions. For the digital experiment, only NO_3^- is available in the environment as a nitrogen source. Then a projection of the space of possible solutions on the following three axes was performed : NO_3^- uptake flux, DMSP secretion, and growth rate (Figure 39).

A quasi-linear relationship is observed between the import of NO_3^- uptake flux and the DMSP secretion. The biomass flux is maximal (about 13 $mmol.gDW^{-1}.h^{-1}$) when the system uptake NO_3^- with a rate of 150 $mmol.gDW^{-1}.h^{-1}$. Between -150 and 0, we see that the production of DMSP is maximal and allows to maintain the growth. The more nitrogen is imported, the less growth can be maintained and the DMSP is exported less and less in order to conserve carbon and sulphur in the system. These observations seem to confirm our hypothesis. A potential issue, particularly in formalised (mathematical) modelling, is what can be referred to as the "plausibility trap". It is important to be cautious and not assume that just because a model replicates an observed behaviour, the underlying mechanisms are significant contributors or even involved at all. Since the system is in QSSA (1.3.3.4), mainly stoichiometry regulates the flux of metabolites within the system. Therefore, what enters the system must leave in the same stoichiometric ratio. Thus, it is possible that the observed balance of C:N:S ratios is a modelling bias rather than a biological phenomenon.

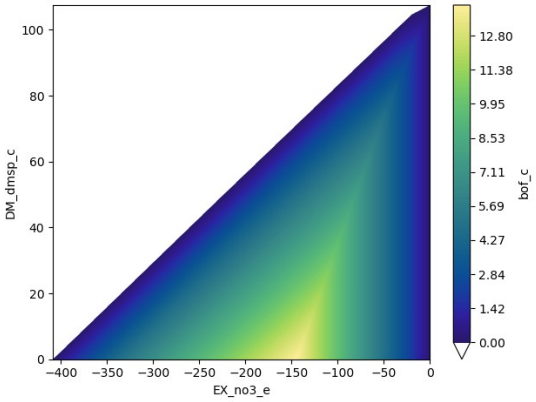


Figure 39: Projection of the allowable solution space on nitrogen uptake, DMSP secretion and growth fluxes ($mmol.gDW^{-1}.h^{-1}$)

To evaluate this hypothesis, we conducted an *in vivo* experiment (Table 5). Using a standard culture of *Phaeodactylum*, we divided the cells into two groups : one placed in fresh medium lacking nitrate, and the other in fresh medium containing nitrate. This experiment was repeated four times. We measured the intracellular concentration of DMSP, with the dashed line representing the concentration in nitrate-free cultures and the solid line representing the concentration in cultures with nitrate. The concentrations were normalised so that they all initially started at 100 µM. In the nitrate-free cultures, the concentration of DMSP increases between 3.5 and 8.5 times.

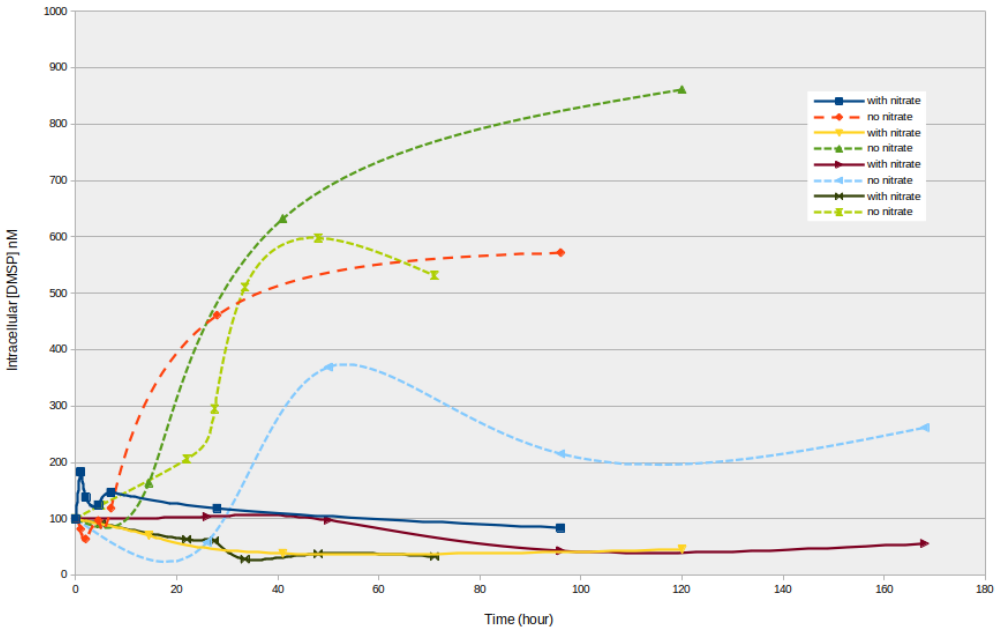


Figure 40: Measurement of intracellular DMSP concentration in Phaeodactylum cultures under nitrate and nitrate-free conditions

In the repeated experiment (Figure 41), we specifically tested the reversibility of the molecular process. During the peak of DMSP production, we observed that the concentration was 6 times higher in the nitrate-free culture compared to the culture with nitrate. Following a 4-days culture deficient in nitrate, the cells were divided into two groups : one placed in fresh medium without nitrate and the other in fresh medium with nitrate. Notably, the nitrogen-deficient cells that had been producing higher levels of DMSP showed a reduction in production when placed back in a medium containing nitrogen. On the other hand, when the cells were transferred to fresh nitrate-free medium, an increase in intracellular DMSP concentration was once again observed.

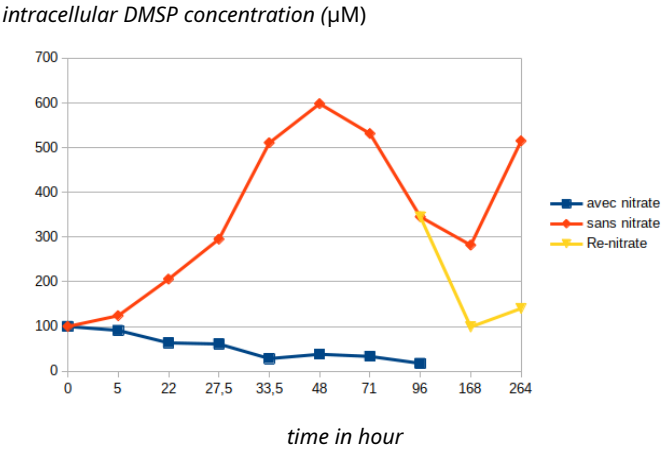


Figure 41: Measurement of intracellular DMSP concentration in Phaeodactylum cultures under nitrate and nitrate-free conditions.

Blue : with nitrate ; Orange ; without nitrate ; Yellow : back in medium with nitrate.

Finally, we repeated the experiment by measuring the extracellular DMSP. The experiment was done only once, and one point had to be removed for technical handling reasons. We still observe a higher concentration of DMSP in the culture without nitrate up to 15 times more. In the nitrate-free culture, there is a notable and unexplained decrease in DMSP concentration after 48 hours. It raises the question of whether this phenomenon is a result of experimental manipulation bias. Therefore, it is recommended to repeat the experiment to validate these findings. If the same phenomenon is observed again, it would be worth investigating the products of DMSP degradation. The degradation pathways of DMSP is well-described for prokaryotes, but in the case of eukaryotic algae, only the Alma1 gene has been yet discovered, and no homolog appears to be present in *Phaeodactylum*. An alternative approach could be incorporating labeled DMSP to track its trajectory and determine if it can be reincorporateby the algae.

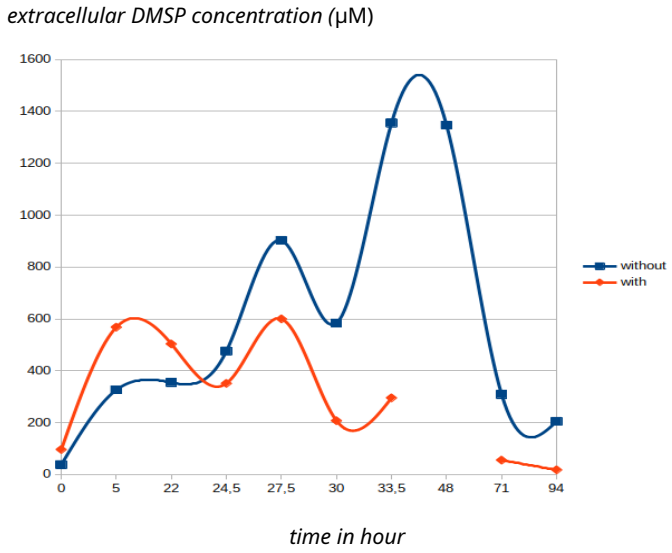


Figure 42: Measurement of extracellular DMSP concentration in Phaeodactylum cultures under nitrate and nitrate-free conditions.

Blue : whitout nitrate ; Orange ; with nitrate

The increase in DMSP production by *Phaeodactylum* during nitrogen stress is prominently observed. However, the experiments must be deepened to determine further the mechanisms.

Underlying metabolic mechanisms ?

By combining *in silico* modelling (example below « Transcriptomics to refine biological networks ») and quantification of key metabolites or transcripts *in vivo,* under different culture conditions, it would be possible to describe the mechanisms that explain the increase in DMSP production under nitrogen stress. Here are some leads I found in the bibliography (see Figure 43).

Hypothesis (1) being the main hypothesis : DMSP production as an overflow mechanism for excess reduced-carbon and -sulphur compounds.

Hypothesis (2). This overflow mechanism may also play a role in *protein turnover*. Protein turnover is an essential process, allowing plants to re-utilise amino acids, to change protein content during development and to adapt their enzyme system to new environmental conditions, especially under stress[259]. When methionine (precursor of DMSP) is produced from the degradation of proteins by proteases, the function of DMSP production would be to redistribute nitrogen into new amino acids through transamination reaction[165]. Does nitrogen deprivation increase the flux through the transaminase of the DMSP pathway? Can we show a protein turnover?

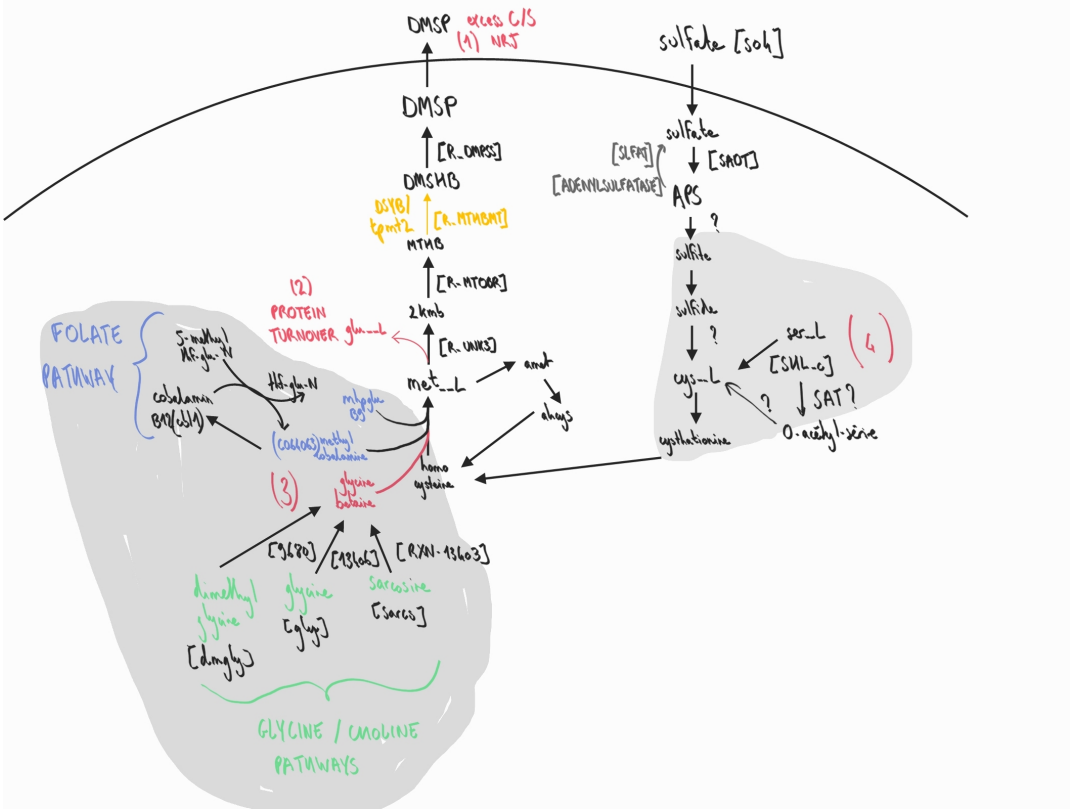


Figure 43: Draft representation of the underlying metabolic mechanisms of DMSP production in PhotoEukStein. Four hypothesis are formulated. (1) DMSP production is described as an overflow mechanism for excess reduced-carbon and -sulphur compounds ; (2) This overflow mechanism may also play a role in protein turnover ; (3) DMSP could act as an osmolyte in algal cells and replace glycine-betaine under nitrogen limitation ; (4) Availability of carbon and nitrogen substrates may be important in the regulation of this pathway rather than sulphur.

Hypothesis (3). Methionine is produced from homocysteine and three different compounds: (a) glycine-betaine (GBT) (from glycine pathway or choline pathway), (b) B12 vitamin, (c) B9 vitamin (both from folate pathway). Similarity both structure and properties between DMSP and its nitrogen analogue GBT was noted[164]. Possibly due to the different bioavailabilities of sulphur and nitrogen in marine and terrestrial environments DMSP is the preferred compatible solute for marine organisms, while terrestrial species use nitrogen compounds such as GBT, choline, carnitine, or ectoin (instead). This led to the suggestion that DMSP could act as an osmolyte in algal cells and even replace GBT under conditions of nitrogen limitation. It is hypothesised that metabolite concentration of the DMSP pathway would increase and those in the GBT pathway would decrease due to nitrogen limitation. Therefore, during nitrogen starvation, should pathway (a) be down-regulated while (b) and (c) be upregulated?

Hypothesis (4). Homeocysteine can be produced from cysteine. Cystein is a sulphur-amino acid. It can be produced from the uptake of sulphate or from serine amino acid. Kettles et al.,[264] findings suggest that increased sulphur assimilation might not be required for increased DMSP synthesis. During the nitrogen starvation, analysis of transcript and protein responses reveals certain patterns that could indicate potential regulatory points beyond sulphate assimilation. Interestingly, some of these changes occur in the branches responsible for supplying carbon and nitrogen skeletons to the central pathway of sulphur assimilation. For instance, there is an increase in transcript levels of SAT (serine pathway) during nitrogen starvation. This observation suggests that the availability of carbon and nitrogen substrates may play a crucial role in regulating this pathway, rather than sulphur itself. This is in contrast to the regulation of sulphur metabolism in higher plants, where upregulation of multiple sulphur assimilatory enzymes is commonly observed.

Transcriptomics to refine biological networks

The models we generally reconstruct are based on the comprehensive potential of the organism, assuming that all proteins encoded in the genome can be utilised by the model. However, this approach may lead to an overestimation of the organism's metabolic capabilities. In reality, not all proteins are expressed simultaneously under different conditions. For a more detailed explanation, please refer to section 4.1.1.3. By taking transcriptomes into account, we obtain a subset of the metabolic network that more accurately reflects the strategies employed by the organism under specific conditions. This approach is extremely innovative.

1) Use PhotoEukStein as generic model to reconstruct an organism-specific model with all the proteins encoded in its genome (which is currently done) ;

2) Use the PhotoEukStein-derived model as a generic model, and reconstruct a sub-model using the fasta containing only the proteins being expressed in environmental conditions X (based on the transcriptome) ;

3) Although it still needs to be thought about, it would be very interesting to adapt the reactions fluxes according to the expression levels of each transcript ;

4) And finally, apply accurately the conditions X to the boundary reactions. And for this fourth point, we must transform the metabolite concentration measured in the medium into a reaction uptake rate (this is already elucidated, especially in biotechnology field).

4.2.2.2 Modelling at the scale of small communities

Competition for metabolic resources can affect community composition through competitive exclusion or by facilitating niche differentiation[261]. Cooperative and syntrophic interactions, such as beneficial metabolic exchanges, are also likely to play an important role, as they can significantly alter the nutritional quality of the habitat[262]. One fascinating aspect of these interactions lies in the mutual exchange of nutrients, such as vitamins, between different organisms. Vitamins are essential organic compounds required for various biological processes[263]. For example, Croft *et al*.[264] showed that 50% of algae surveyed require vitamin B12 for growth, but cannot synthesise it *de novo* (auxotrophy).

Many enzymes that have a B12 coenzyme are known in eukaryotes, including the B12-dependent methionine synthase (see Figure 43, hypothesis 3). This means that these algae rely on external sources for their supply. Thus, prokaryotes often form partnerships with microalgae, providing them with the needed vitamins (Figure 44). In return, microalgae offer prokaryotes a stable environment and nutrients they can synthesise. This cooperative exchange of resources illustrates the power of symbiosis in maintaining ecological balance.

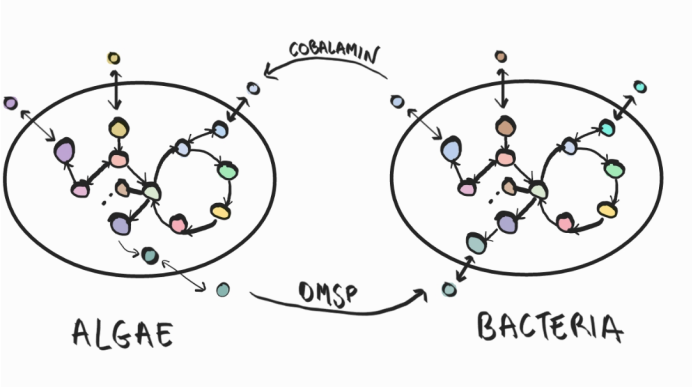


Figure 44: Illustration of a symbiotic interaction between a eukaryotic microalga and a bacterium. Hypothetically, the alga could provide DMSP as a source of organic sulphur for a bacterium that lacks the assimilative pathway of sulphate reduction, and the bacterium could provide cobalamin (B12) to the microalgae so that it could produce methionine in the event that glycine betaine is lacking (nitrogen stress)

Having more than 549 algal models at our disposal will allow us to dig much further in the study of these interactions, whether they are viral, parasitic or synergistic. By delving into the study of microorganisms' interactions, we gain insights into the interdependence of different species, their roles in nutrient cycling, how it shapes community composition through competitive exclusion or by facilitating niche differentiation, how it alters the nutritional quality of the habitat and so on. It is possible to use co-occurrence techniques to capture modules of species likely to interact[265] and to highlight their metabolic interdependencies with tools adapted to GSMs[8,103]. However, going to population-level behaviours that emerge from these individual interactions from this bottom-up approaches requires significant computational time and is therefore limited to small communities (up to 4 species to my knowledge). In order to move to ocean-scale modelling, implementing trait-based approaches may overcome the computational demand of individual-based models.

4.2.3 Trait-based modelling at ocean-scale

Ocean System Models (OSMs) have become more sophisticated allowing for detailed simulations that provide valuable insights into the ocean, its resources, and its future (1.2.3). These models, like NEMO-PISCES, use differential equations to depict the growth of emblematic organisms by linking nutrient availability with growth rate at ocean-scale. They incorporate physical processes to predict nutrient availability globally and over time. However, these equations require numerous parameter values that are often challenging to obtain experimentally. Furthermore, while OSMs are computationally efficient, they do not fully consider recent omics data, such as genes and associated functions, limiting their ability to capture all intra-individual variability and molecular processes. Additionally, these models oversimplify the association of functional traits with phylogeny, which is known to be a reducing approach.

A significant advancement in this field is the integration of Genome-Scale Models (GSMs) into OSMs, as proposed in the forthcoming paper "modelling genome-scale knowledge in the global ocean" by Regimbeau *et al*. This integration addresses the challenge of estimating growth rates while holistically considering the metabolism of the organism. They also leverage the OSM's environmental conditions to explore the niche space, revealing the physiological properties of modelled organisms. This is the first time that omics knowledge is applied into OSMs, opening doors to evolutionary theory.

To facilitate the integration of GSMs with OSMs, it becomes crucial to enhance the linkage between these two types of models. While OSMs may have a smaller set of metabolites compared to GSMs, there are still certain metabolites that are absent from GSMs. For example, in the case of PhotoEukStein, iron and silicate are not included in the model. Currently, the connections between GSMs and OSMs primarily involve the consideration of three key factors : nitrogen, phosphorus, and light. These 3 connection points are not at all sufficient to correctly predict the growth of some organisms. For example, in most open ocean ecosystems, there is typically a positive correlation between macronutrient concentrations and phytoplankton biomass, especially when sunlight is sufficient. However, this conventional understanding does not hold true in certain regions of the world ocean, namely the subarctic Pacific, the eastern and central equatorial Pacific, and the Southern Ocean. These regions, referred to as high nutrient-low chlorophyll areas, exhibit elevated nitrate and phosphate concentrations throughout the year but relatively low phytoplankton levels[266]. Indeed, the growth of large phytoplankton cells, particularly diatoms, is limited not only by phosphate but also by the availability of iron or silicate[267] and explain the limiting autotrophic activity in these regions. Expanding the range of metabolites considered in GSMs and aligning them with the relevant components in OSMs will be an important step in achieving a more comprehensive and accurate representation of ecosystem dynamics.

In addition to integrating the metabolites of the OSM into PhotoEukStein, we can also inversely propose new key metabolites to integrate in the OSM. This will require a thorough review of all SKs, DMs and EXs of PhotoEukStein. Going beyond the traditional PFT concepts[97], GSMs can serve as valuable tools for defining functional traits that are specific to certain environmental conditions, independently of taxonomic or phylogenetic considerations, as proposed in our paper (3.2). This approach allows us to explore the functional characteristics of organisms in a more nuanced and context-dependent manner.

Today, this feat of integrating GSMs with OSMs is done for one organism at a time. However, microorganisms rarely exist in isolation and often rely on synergistic interactions with other organisms[268]. The intricate associations within these communities contribute to their stability across diverse and variable environments[269]. In this regard, the next frontier in metabolic modelling lies in utilising metabolic networks with a focus on modelling multi-organism systems. However, the complexity of metabolic networks as data structures[270] poses challenges. Efforts to adapt metabolic networks to ecosystem modelling are essential. An interesting strategy would be to change the biological scale and not to consider one compartment (a GSM) per organism, but rather a model that contains the functional diversity of several organisms sharing the same trait. The idea would be to reconstruct a meta-model containing the whole set of reaction of the models clustered in a the same functional group. The objective with this approach is once again to go beyond the taxonomic classification.

Another current limitation in metabolic modelling is the reliance on a biomass objective function that is typically parameterised for a specific algal species cultured in laboratory, making it less applicable to a wide range of algae in their natural environments. The optimisation for growth assumes that microbial cells maximize their growth, which may be suitable for bioengineering purposes but not necessarily for ecological applications, where nutrient stresses are common. Furthermore, in a study using multi-objective modelling of a small ecosystem[103], it has been demonstrated that when each species grows at its maximum rate, other guilds fail to produce biomass. And to maintain ecosystem stability, species need to grow at suboptimal rates. If it is

necessary to consider a growth function, it would be more appropriate to generate a generic biomass reaction that considers the minimum requirements while minimising energy consumption, for example. This approach would better reflect the realistic metabolic behaviour of organisms in their natural environments.

In summary, it is crucial to emphasise the importance of integrating diverse organisms and community traits, addressing challenges related to the integration of omics data, and comprehending the variability and biogeographical structure of planktonic communities in ecosystem modelling. Despite the obstacles, the incorporation of omics data into ecosystem models has the potential to enhance our understanding of planktonic ecosystems and their responses to environmental changes at ocean-scale. To further advance our understanding of plankton diversity and its contributions to Earth system functioning, collaborative efforts across multiple research fields and the development of innovative approaches and technologies are essential.

CONCLUSION

This thesis focuses on connecting omics data to marine ecosystems through metabolic modelling.

Marine plankton, encompassing a wide range of organisms from viruses to meter-sized cnidarians, including archaea, bacteria, and single-celled eukaryotes, dominate the ocean and engage in dynamic interactions. These organisms actively contribute to Earth's functioning by conducting nearly half of the planet's net primary production and transporting photosynthetically fixed carbon to the deep oceans. However, a significant portion of planktonic life remains understudied.

Advancements in sequencing technology and bioinformatics have enabled the generation of vast amounts of sequencing data from environmental samples at increasingly affordable costs. These developments have facilitated the reconstruction of numerous Metagenome-Assembled Genomes (MAGs) for viruses, bacteria, archaea, and eukaryotes. These MAGs cover a significant portion of biological diversity in various environments, providing valuable genomic and transcriptomic insights beyond what can be obtained from cultured organisms. By leveraging omics-based approaches, we can substantially enhance our understanding of the biology of these uncultured organisms and their contributions to ecosystem functioning.

Genome-scale metabolic models (GSMs) provide quantitative and computable relationships between genotypes and phenotypes of target organisms, despite not incorporating various biological regulations that affect enzymatic activities within cells. Originally used for modelling cellular physiology and growth in model organisms across fields like biotechnology and synthetic biology, these constraint-based approaches are now being extended to predict and understand microbial communities. GSMs encompass all the metabolic reactions encoded in a genome or transcriptome and their interconnectedness. By exploring the solution space (which represents all possible solutions in the multidimensional space defined by metabolic fluxes subject to thermodynamic constraints), we can compute and predict metabolic phenotypes. Typically, this involves optimising an objective function of interest, often the growth rate, to determine the most favorable metabolic state within the model.

Currently, there are several ecologically relevant genome-scale metabolic models (GSMs) available for prokaryotes (BiGG[130], EcoCyc[252], or CyanoCyc[not published]). However, the development of models for marine eukaryotic microbes lags behind. This is primarily due to the limited number of model organisms with sequenced genomes and the time-consuming process of manual curation required to construct effective models. In traditional bottom-up approaches, the manual curation steps for each new model reconstruction are labor-intensive. To address this challenge, the top-down approach introduces a generic meta-model that undergoes manual curation and captures relevant structural properties. This meta-model serves as a template, which can be converted into organism-specific models without requiring repeated curation efforts. This approach has been successfully applied to prokaryotes. However, its application to marine eukaryotic microbes is yet to be explored.

In this thesis, we present PhotoEukStein, a generic meta-model designed to facilitate the automated reconstruction of metabolic models for eukaryotic algae. PhotoEukStein integrates biochemical and genomic information from 16 eukaryotic algae species, capturing the key features of photosynthetic eukaryotic cells that utilise light energy to convert carbon dioxide into organic

compounds. Through extensive manual curation, we have prepared PhotoEukStein for simulation purposes. To evaluate the performance of PhotoEukStein-derived models, we sampled the computed metabolic niches and compared the predicted growth rates with those obtained from expert-based GSMs. In all cases, the predicted growth rates from both approaches showed a high correlation, indicating that the PhotoEukStein-derived models accurately capture relevant metabolic properties comparable to manually reconstructed and experimentally validated models of specific algae species (e.g., *Phaeodactylum, Chlorella, Thalassiosira*). Furthermore, by examining the correlations between reaction fluxes within each model, we observed highly similar correlation maps between the reference models and the PhotoEukStein-derived GSMs. This finding suggests that the interconnectedness of reactions in both models is closely aligned. Overall, PhotoEukStein represents a significant advancement in our understanding and modelling of the metabolism, physiology, biogeochemistry, and ecology of phototrophic eukaryotes.

The application of this new method to *Tara Oceans* environmental genomes (MAGs) and transcriptomes (MetDB) of phototrophic marine unicellular eukaryotes has resulted in the derivation of 549 models from PhotoEukStein. This expanded database serves as a valuable resource, providing opportunities for comprehensive ecosystemic exploration of plankton communities spanning from viruses to single-cell phototrophs. Importantly, the process of deriving PhotoEukStein-based GSMs from new genomes or transcriptomes does not demand extensive computational resources or time-consuming expertise. This enables researchers to efficiently handle the ever-increasing collection of environmental genomes and conduct further investigations in this field.

Bridging plankton ecosystems and biogeochemistry poses a significant challenge in modelling. Traits-based models, such as Planktonic Functional Types (PFT), rely on taxonomic data and do not capture the mechanistic aspects of planktonic ecosystems. In contrast, GSMs offer a valuable approach to capture the metabolic phenotype of microorganisms, considering environmental conditions, and can provide a more accurate depiction of functional trait distribution across species compared to relying solely on taxonomy or gene presence/absence.

GSMs can quantitatively link the biological functions of planktonic ecosystems to biogeochemical processes. Indeed, metabolic niches represent the environmental parameters (as fluxes of available metabolites) in which a specific metabolic model can generate biomass. They formalise the organism's function by defining the space in which it can survive based on its ability to utilise available resources through metabolic reactions. The derivation of functional GSMs for unicellular phototrophic eukaryotes, even from environmental omics data, provides a unique approach to assess their biological phenotype beyond the identification of functional genes. These features represent new observations or semantic traits that emerge from genomic descriptions. Analysing the variability of metabolic fluxes and metabolite exchanges through the study of metabolic niches enables the differentiation of cellular resource allocation to various survival needs, including resource acquisition, defense, signaling, and community metabolic interactions (as seen in the concepts of phycosphere or holobiont). This approach offers insights into the complex dynamics of metabolic interactions within planktonic ecosystems and their ecological significance.

The integration of the eukaryotic layer in ecological studies represents a significant advancement, enabling a comprehensive exploration of plankton communities from viruses to single-cell phototrophs. This breakthrough opens up new possibilities for understanding the intricate dynamics of these ecosystems and contributes to a holistic description of phenotypic biodiversity and ecosystem modelling. By utilising PhotoEukStein and its derived GSMs, we can highlight and emphasise omics-

driven phenotypes that serve as essential traits in future ocean system models. These models provide a valuable resource for capturing the diverse functional characteristics of unicellular eukaryotes, greatly enhancing our understanding of their ecological roles. Through this integration, we can advance towards a more comprehensive and accurate representation of phenotypic biodiversity, leading to improved ecosystem modelling and a deeper understanding of marine ecosystems as a whole. Therefore, the systematic derivation of GSMs for unicellular eukaryote phototrophs, similar to what has been achieved for prokaryotes, represents a crucial step towards developing omics-trait-based models that can enhance our global ecological understanding.

TABLES

Table 1: Microbes rule the world

Microbes, short for microorganisms, are microscopic living organisms that are too small to be seen by the naked eye. They are found virtually everywhere on Earth, including in the air, clouds, soil, water (plankton), and even in very close collaboration with other organisms. Microbes can be classified into several taxonomic groups, including bacteria, archae, viruses, fungi, protozoa and algae, and cover a wide range of functions through biological processes (nutrient cycling, decomposition, disease, free-living symbiosis or holobiont...). They have adapted to thrive in a wide range of environments, making them an essential part of Earth's ecosystems. Prokaryotic microorganisms were the first life forms on Earth and have participated in the formation of other more complex life forms. Scientists have come to appreciate the importance of plankton in mediating major biogeochemical cycles of the Earth. Research in the late 1980s by geochemists and biologists contributed to a better understanding of their role in maintaining the balance of the Earth's systems[8,9]. These organisms not only help to maintain the steady-state gas composition of the atmosphere but also respond to climate feedbacks, contribute to the regulation of the Earth's climate and weather patterns[10,11]. Although understanding microbial responses to climate change is an active area of research in microbial ecology and climate science, it is still very difficult to accurately capture and predict the dynamics of such populations. However, in my modest opinion, even in the event of a collapse of life, it is very likely that some microbes could survive. Some of them are able to live in extreme environments (hot water source, volcano, ice...). In addition to their essential role in sustaining life on Earth, they are also used extensively in industrial applications such as food production, biotechnology, and environmental cleanup. Moreover, they feed petroleum and natural gas reservoirs that fuel contemporary civilisation. My purpose in this box is to emphasise the essentiality of microbes to sustain life, and highlight their enormous range of taxonomic and functional diversities. Microbes rules the world.

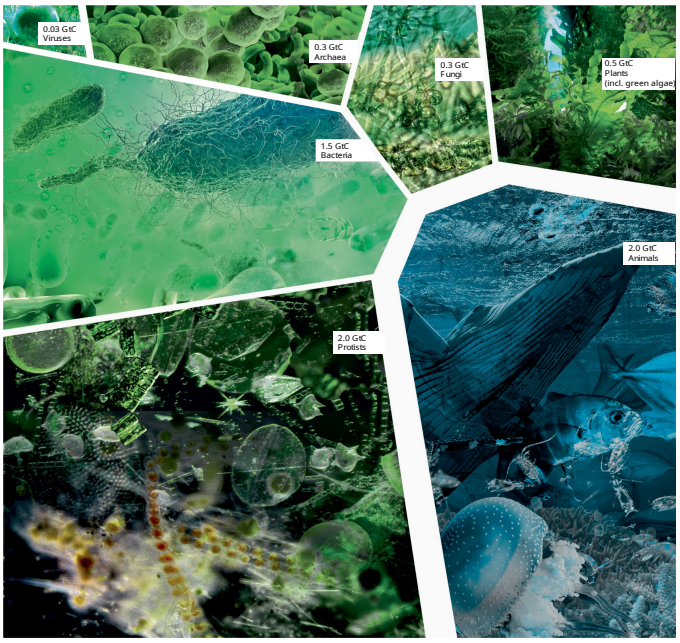


Figure 45: Living organisms carbon biomass (Giga-ton). Microbes biomass is higher than animal biomass.

Table 2: Heuristic constraints applied to prevent the generation of ATP out of nowhere.

- ATP citrate lyase is an important acyltransferase in fatty acid biosynthesis that cleaves citrate to oxaloacetate and acetyl-CoA (coenzyme A) with concomitant hydrolysis of one molecule of ATP to ADP and Pi [235,236]. « ATPCS_c » reaction has become irreversible to avoid ATP production.
- But this reaction is the reciprocal of the first step of the Krebs cycle [237]. So I manually added « CISY_m : acetyl-CoA + H2O + oxaloacetate → citrate + CoA », producing citrate without creating ATP.
- There are two ATP-generating enzymes in glycolysis : PYK (pyruvate kinase) et PGK (Phosphoglycerate kinase). PKG is reversible to either function in glycolysis or gluconeogenesis. Thus, « PGK_m » would catalyse reversibly both reactions : under biochemical standard conditions, the one direction is favored [39]. But the enzyme is also used in the Calvin cycle (chloroplast) consuming ATP to catalyse the phosphorylation of 3-phosphoglycerate (Figure 31). I created then the irreversible reaction « PGK_h » . PGA enzyme is known to occur in different compartments [238]. For further information, see 2.2.2.
- As ATP-generating enzyme, I also found the Succinyl-CoA ligase, which is an isoenzyme that can catalyse three different reactions (succinyl-CoA synthetase, succinyl-CoA ligase ADP or GDP forming). I didn't find any distinct reactions in BiGG so I left « SUCLm » reversible.
- Finally, the last ATP-generating reaction is the ATP synthase (ATPS). Although it is known that ATPS also resides in the mitochondrial membrane of eukaryotic cells and creates ATP by oxidative phosphorylation, only the chloroplastic ATPS is present in PhotoEukStein, and its flux is regulated according to the photon uptake in the system (see 2.2.2.1).

Table 3: Heuristic constraints applied to prevent the uptake of CO_2 without light

- Important enzymes involved in carbon fixation in algae can include phosphoenolpyruvate (PEP) carboxylase (PEPC). This enzyme is involved in the fixation of carbon dioxide during C4 photosynthesis, a specialised type of photosynthesis found in some plants and algae. PEPC has a more effective carboxylase activity than RuBisCo. PEPC catalyses the carboxylation of PEP to form oxaloacetate, which is then converted to malate and transported to bundle sheath cells where carbon dioxide is released. This phenomen concentrates carbon dioxide around RuBisCo to improve its efficiency and limits then photorespiration. Indeed, photorespiration results from the fixation of one molecule of oxygen (O_2) by the oxygenase activity of RuBisCo in parallel with the carboxylase activity of this enzyme. Oxygen has a higher affinity than CO_2 for RuBisCo.

- Carbonic anhydrase (CA) is an enzyme found in photosynthetic organisms that catalyses the reversible reaction between carbon dioxide (CO_2) and water (H_2O) to form bicarbonate ions (HCO_3^-) and protons (H^+). This reaction is important for photosynthesis as it provides the carbon dioxide needed for carbon fixation. CA is found in the chloroplasts, where it plays a crucial role in supplying CO_2 to the Calvin cycle, which is the series of biochemical reactions that fix carbon dioxide and synthesise glucose. Without carbonic anhydrase, the rate of photosynthesis in vivo would be greatly reduced, as carbon dioxide would not be efficiently utilised by the photosynthetic machinery.

  CA and PEPC are known reactions which can use inorganic carbon as substrate without direct energy.

- There are also carboxylases. A carboxylase is an enzyme that catalyses the addition of a carboxyl group (-COOH) to a substrate molecule. This process is called carboxylation and is an important mechanism in many biological processes, including photosynthesis, lipid biosynthesis, and amino acid metabolism. Carboxylases are often dependent on co-factors, such as biotin, or ATP, to carry out their catalytic activity. For example, Acetyl-coA carboxylase catalyses irreversibly the carboxylation of acetyl-CoA to form malonyl-CoA, a key intermediate in the biosynthesis of fatty acids. The reverse reaction, which involves decarboxylation of malonyl-CoA to acetyl-CoA, is catalysed by another enzyme called malonyl-CoA decarboxylase. The pyruvate carboxylase, as for it, catalyses the carboxylation of pyruvate to form oxaloacetate, an important intermediate in the citric acid cycle and gluconeogenesis. Pyruvate carboxylase is a reversible enzyme catalysing the reaction in both directions depending on the concentration of the reactants and products in the cellular environment. I adjusted the constraints on the fluxes according to these data.

- Furthermore, in PhotoEukStein other reactions can assimilate inorganic carbon without energy. The irreversibility of these reactions and the missing energy source not being described, or my knowledge being limited, a bound of 0.001 is set for these reactions in order to limit the use of CO_2 without losing biological information. Metadata have been added to these reactions in order to be able to examine them later.

Table 4: Metabolic pathways found in PhotoEukStein based on KofamKOALA[172]

- Carbohydrate metabolism : Ascorbate and aldarate metabolism ; Glyoxylate and dicarboxylate metabolism ; Starch and sucrose metabolism ; Pyruvate metabolism ; C5-Branched dibasic acid metabolism ; Pentose and glucuronate interconversions ; Amino sugar and nucleotide sugar metabolism ; Glycolysis / Gluconeogenesis ; Butanoate metabolism ; Galactose metabolism ; Citrate cycle (TCA cycle).

- Lipid metabolism : Fatty acid degradation ; Biosynthesis of unsaturated fatty acids ; Glycerolipid metabolism ; Steroid hormone biosynthesis ; alpha-Linolenic acid metabolism ; Glycerophospholipid metabolism ; Fatty acid elongation ; Sphingolipid metabolism ; Fatty acid biosynthesis.

- Nucleotide metabolism : Pyrimidine metabolism ; Purine metabolism.

- Xenobiotics biodegradation and metabolism : Caprolactam degradation ; Benzoate degradation.

- Metabolism of cofactors and vitamins : Pantothenate and CoA biosynthesis ; Folate biosynthesis ; Nicotinate and nicotinamide metabolism ; One carbon pool by folate ; Thiamine metabolism.

- Energy metabolism : Photosynthesis ; Oxidative phosphorylation ; Methane metabolism ; Carbon fixation pathways in prokaryotes ; Nitrogen metabolism ; sulphur metabolism.

- Amino acid metabolism : Lysine biosynthesis ; Arginine and proline metabolism ; Tryptophan metabolism ; Lysine degradation ; Valine, leucine and isoleucine degradation ; Cysteine and methionine metabolism ; Valine, leucine and isoleucine biosynthesis.

- Metabolism of other amino acids : beta-Alanine metabolism ; Glutathione metabolism ; Selenocompound metabolism.

- Glycan biosynthesis and metabolism : N-Glycan biosynthesis ; Glycosphingolipid biosynthesis - ganglio series ; Glycosphingolipid biosynthesis - globo and isoglobo series ; Other glycan degradation ; Glycosylphosphatidylinositol (GPI)-anchor biosynthesis ; Glycosaminoglycan degradation ; Various types of N-glycan biosynthesis.

- Biosynthesis of other secondary metabolites : Monobactam biosynthesis ; Glucosinolate biosynthesis.

- Metabolism of terpenoids and polyketides : Terpenoid backbone biosynthesis.

Table 5: *Phaeodacylum tricornutum* culture and measurement of DMSP production under nitrogen stress.

- Growth conditions and sample collection :

  Culture of Phaeodactylum tricornutum (Pt1/CCMP2561) was grown in f/2 ASW medium[239] during 7 days in a growth chamber at 20°C in erlenmeyer flask shaken at 150 rpm under 80 umole photon * m^2 * sec -1 irradiance with a cold-white led light for 12-h dark / 12-h light photoperiod. For f/2-N, KNO3 was omitted and replaced by 9.9mM KCl. After centrifugation (3000 g, 15 min, at room temperature), the cells were recovered either in a fresh f/2 medium or in f/2-N medium. For the DMSP production kinetics, a volume of culture containing 1.108 cells was collected at different times by filtration onto a 47 mm PTFE filter (JH Omnipore, 0.45 µm) and proceeded after the metabolite extraction paragraph. At the same time, 1 ml of cells was also collected to analyse the extracellular content. The cells were centrifuged in 1.5 ml Eppendorf tube 10 minutes at 20,000g, 4°C. 800 µL of the supernatant were then collected and frozen at -80°C. Cell concentration was determined under optic microscope using a Thoma cell-counting chamber.

- Reversibility experiment :

  Phaeodactylum tricornutum (Pt1/CCMP2561) was grown in f/2-N medium for 4 days. Cells were recovered by centrifugation (3000 g, 15 min, room temperature) in either fresh f/2 medium or f/2-N medium. After 3 days, both cultures were processed as described below to quantify DMSP content.

- Metabolite extraction (adaptated from[240]) and LC/MS/MS analysis:

  Metabolism was quenched by placing this filter in 5 ml of a cold mixture of H2O/ Methanol/ Acetonitrile (1/3/1). After sonication to remove the cells from the filter, the solution was transferred into cryogenic vials and underwent 3 freeze/thaw cycles in liquid nitrogen/65 °C water to fully break the cells and extract the metabolites. The debris were removed by centrifugation (20 000 g, 10 mn, RT) and the supernatant was dried and first dissolved in 300 µl water.

  Before LC/MS analysis, the intra and extracellular samples were filtered on 0.22 µm (polytetrafluoroethylene; AcroPrep Advance, Pall) and finally diluted in a solution composed of 80% acetonitrile and 20% 10mM ammonium carbonate (pH 9.9).

  DMSP was detected by LC/MS/MS using a Dionex UltiMate TCC-3000RS chromatographic system (Thermo Fisher Scientific) coupledto a hybrid triple quadrupole linear ion trap mass spectrometer (QTRAP 5500 from ABSciex) equipped with a heated electrospray ionisation source. Chromatographic separation was achieved on a ZIC-pHILIC column (100 X 2.1 mm; 5µm ; Merck) thermostated at 40°C. The mobile phase flow rate was set at 0.2 ml/min and 5 µl of sample was injected. Mobile phase A consisted of 10mM ammonium carbonate, pH 9.9 and the mobile phase B consisted of acetonitrile. The gradient started at 80% B for 1 min followed by linear gradient to 40% B for 7 min, and remained at 40% B for 3 min. The system returned to the initial solvent composition in 3 min and reequilibrated under these conditions for 8.5 min.

  Mass spectrometry analyses were conducted with the following parameters : ion source 5000 V in positive mode, curtain gas 25 a.u., temperature 500°C, gas 1 60 a.u., gas 2 60 a.u., computer-aided-design medium. MS/MS experiments were performed in the triple quadripole mode using multiple reaction monitoring scan type. The optimisation of following MS parameters (declustering potential, collision energy and cell exit potential) was performed in order to establish the best intensity transitions (Table 6).

Table 6: MRM transitions and chromatographic retention time (RT) of DMSP detected by LC/MS/MS.

Parent mass m/z	Product mass m/z	DP Volts	CE Volts	CXP Volts	RT min
135.04748	73 62.8	60 60	17 17	6 8	5.25

DP, declustering potential ; CE, collision energy ; CXP, cell exit potential.

BIBLIOGRAPHY

1. De Meyer, T. Bruno Latour, Face à Gaïa. Huit conférences sur le Nouveau Régime Climatique. *Lectures* (2016) doi:10.4000/lectures.19763.
2. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).
3. Lovelock, J. E. & Margulis, L. Atmospheric homeostasis by and for the biosphere: the gaia hypothesis. *Tellus* **26**, 2–10 (1974).
4. *La fin d'un grand partage - Nature et société, de Durkheim à Descola - CNRS Editions*.
5. Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).
6. Buick, R. When did oxygenic photosynthesis evolve? *Philos Trans R Soc Lond B Biol Sci* **363**, 2731–2743 (2008).
7. Knoll, A. The Multiple Origins of Complex Multicellularity. *Annual Review of Earth & Planetary Sciences* **39**, 217–239 (2011).
8. Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in diverse microbial communities. *PNAS* **112**, 6449–6454 (2015).
9. *Ocean Productivity and Organic Carbon Flux: Overview and maps of primary production and export production*. (Scripps Institution of Oceanography, University of California, San Diego, 1987).
10. Falkowski, P. G., Laws, E. A., Barber, R. T. & Murray, J. W. Phytoplankton and Their Role in Primary, New, and Export Production. in *Ocean Biogeochemistry: The Role of the Ocean Carbon Cycle in Global Change* (ed. Fasham, M. J. R.) 99–121 (Springer, 2003). doi:10.1007/978-3-642-55844-3_5.
11. Watson, A. J. & Lovelock, J. E. Biological homeostasis of the global environment: the parable of Daisyworld. *Tellus B: Chemical and Physical Meteorology* **35**, 284–289 (1983).
12. Falkowski, P. G. *et al.* The Evolution of Modern Eukaryotic Phytoplankton. *Science* **305**, 354–360 (2004).
13. Keller, M. D., Bellows, W. K. & Guillard, R. R. L. Dimethyl Sulfide Production in Marine Phytoplankton. in *Biogenic sulphur in the Environment* vol. 393 167–182 (American Chemical Society, 1989).
14. Hanson, A. D., Rivoal, J., Paquet, L. & Gage, D. A. Biosynthesis of 3-dimethylsulfoniopropionate in Wollastonia biflora (L.) DC. Evidence that S-methylmethionine is an intermediate. *Plant Physiol* **105**, 103–110 (1994).
15. Kocsis, M. G. *et al.* Dimethylsulfoniopropionate biosynthesis in Spartina alterniflora1. Evidence that S-methylmethionine and dimethylsulfoniopropylamine are intermediates. *Plant Physiol* **117**, 273–281 (1998).
16. Otte, M. L., Wilson, G., Morris, J. T. & Moran, B. M. Dimethylsulphoniopropionate (DMSP) and related compounds in higher plants. *J Exp Bot* **55**, 1919–1925 (2004).
17. Raina, J.-B. *et al.* DMSP biosynthesis by an animal and its role in coral thermal stress response. *Nature* **502**, 677–680 (2013).
18. Ausma, T., Kebert, M., Stefels, J. & De Kok, L. J. DMSP: Occurrence in plants and response to salinity in Zea mays. in *sulphur Metabolism in Higher Plants - Fundamental, Environmental and Agricultural Aspects* (eds. De Kok, L. J., Hawkesford, M. J., Haneklaus, S. H. & Schnug, E.) 87–91 (Springer, 2017). doi:10.1007/978-3-319-56526-2_8.
19. Curson, A. R. J. *et al.* Dimethylsulfoniopropionate biosynthesis in marine bacteria and identification of the key gene in this process. *Nat Microbiol* **2**, 17009 (2017).
20. Curson, A. R. J. *et al.* DSYB catalyses the key step of dimethylsulfoniopropionate biosynthesis in many phytoplankton. *Nat Microbiol* **3**, 430–439 (2018).
21. Kageyama, H., Tanaka, Y., Shibata, A., Waditee-Sirisattha, R. & Takabe, T. Dimethylsulfoniopropionate biosynthesis in a diatom Thalassiosira pseudonana: Identification of a gene encoding MTHB-methyltransferase. *Archives of Biochemistry and Biophysics* **645**, 100–106 (2018).
22. Williams, B. T. *et al.* Bacteria are important dimethylsulfoniopropionate producers in coastal sediments. *Nat Microbiol* **4**, 1815–1825 (2019).
23. McParland, E. L., Lee, M. D., Webb, E. A., Alexander, H. & Levine, N. M. DMSP synthesis genes distinguish two types of DMSP producer phenotypes. *Environmental Microbiology* **23**, 1656–1669 (2021).
24. Smith, D. *et al.* Proteome Remodelling in Response to sulphur Limitation in " Candidatus Pelagibacter ubique". *mSystems* **1**, e00068-16 (2016).
25. DeBose, J., Lema, S. & Nevitt, G. Dimethylsulfoniopropionate as a Foraging Cue for Reef Fishes. *Science (New York, N.Y.)*

319, 1356 (2008).

26. Miller, T. R., Hnilicka, K., Dziedzic, A., Desplats, P. & Belas, R. Chemotaxis of Silicibacter sp. Strain TM1040 toward Dinoflagellate Products. *Applied and Environmental Microbiology* **70**, 4692–4701 (2004).

27. Seymour, J. R., Simó, R., Ahmed, T. & Stocker, R. Chemoattraction to dimethylsulfoniopropionate throughout the marine microbial food web. *Science* **329**, 342–345 (2010).

28. Zheng, Y. *et al.* Bacteria are important dimethylsulfoniopropionate producers in marine aphotic and high-pressure environments. *Nat Commun* **11**, 4658 (2020).

29. Karsten, U., Kueck, K., Vogt, C. & Kirst, G. Dimethylsulfoniopropionate Production in Phototrophic Organisms and its Physiological Functions as a Cryoprotectant. *Biological and Environmental Chemistry of DMSP and Related Sulfonium Compounds* (1996) doi:10.1007/978-1-4613-0377-0_13.

30. Husband, J. D., Kiene, R. P. & Sherman, T. D. Oxidation of dimethylsulfoniopropionate (DMSP) in response to oxidative stress in Spartina alterniflora and protection of a non-DMSP producing grass by exogenous DMSP+acrylate. *Environmental and Experimental Botany* **79**, 44–48 (2012).

31. Sunda, W., Kieber, D. J., Kiene, R. P. & Huntsman, S. An antioxidant function for DMSP and DMS in marine algae. *Nature* **418**, 317–320 (2002).

32. Trottmann, F. *et al.* Sulfonium Acids Loaded onto an Unusual Thiotemplate Assembly Line Construct the Cyclopropanol Warhead of a Burkholderia Virulence Factor. *Angewandte Chemie International Edition* **59**, 13511–13515 (2020).

33. Inventaire, poème écrit par Jacques Prévert.

34. Flynn, K. J. Reply to Horizons Article 'Plankton functional type modelling: running before we can walk' Anderson (2005): II. Putting trophic functionality into plankton functional types. *Journal of Plankton Research* **28**, 873–875 (2006).

35. Sir Peter Blake killed in Amazon pirate attack. *NZ Herald* https://www.nzherald.co.nz/nz/sir-peter-blake-killed-in-amazon-pirate-attack/XJPSKTOY5EN2C5VGG2SMWZJ2PY/ (2023).

36. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology* **9**, e1001177 (2011).

37. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology* 1–18 (2020) doi:10.1038/s41579-020-0364-5.

38. Rigonato, J. *et al.* Insights into biotic and abiotic modulation of ocean mesopelagic communities. 2021.02.26.433055 Preprint at https://doi.org/10.1101/2021.02.26.433055 (2021).

39. Watson, H. C. *et al.* Sequence and structure of yeast phosphoglycerate kinase. *EMBO J* **1**, 1635–1640 (1982).

40. The Nobel Prize in Physiology or Medicine 1910. *NobelPrize.org* https://www.nobelprize.org/prizes/medicine/1910/kossel/biographical/.

41. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

42. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

43. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).

44. Wohns, A. W. *et al.* A unified genealogy of modern and ancient genomes. *Science* **375**, eabi8264 (2022).

45. Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *NATURE* **408**, 796–815 (2000).

46. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

47. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidisation in major angiosperm phyla. *Nature* **449**, 463–467 (2007).

48. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804–810 (2007).

49. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

50. Aditya, L., Mahlia, T. M. I., Nguyen, L. N., Vu, H. P. & Nghiem, L. D. Microalgae-bacteria consortium for wastewater treatment and biomass production. *Science of The Total Environment* **838**, 155871 (2022).

51. "Candidatus Cloacamonas Acidaminovorans": Genome Sequence Reconstruction Provides a First Glimpse of a New Bacterial Division | Journal of Bacteriology. https://journals-asm-org.insb.bib.cnrs.fr/doi/10.1128/JB.01248-07.

52. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc Biol Sci* **270**, 313–321 (2003).

53. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5**, R245-249 (1998).

54. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088–5090 (1977).

55. LINNÉ - repères chronologiques - Encyclopædia Universalis. https://www.universalis.fr/encyclopedie/linne-reperes-chronologiques/.

56. De Vargas et al. - 2015 - Eukaryotic plankton diversity in the sunlit ocean.pdf.

57. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8**, (2017).

58. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **0**, (2022).

59. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).

60. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, (2015).

61. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).

62. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nature Communications* **9**, 373 (2018).

63. Royo-Llonch, M. *et al.* Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. *Nat Microbiol* **6**, 1561–1574 (2021).

64. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology* **3**, 804–813 (2018).

65. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **2**, 100123 (2022).

66. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).

67. Frémont, P. *et al.* Restructuring of plankton genomic biogeography in the surface ocean under climate change. *Nat. Clim. Chang.* **12**, 393–401 (2022).

68. Vorobev, A. *et al.* Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Res.* (2020) doi:10.1101/gr.253070.119.

69. Vorobev, A. *et al.* Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Res.* (2020) doi:10.1101/gr.253070.119.

70. Upton, J., Janeka, I. & Ferraro, N. The whole is more than the sum of its parts: Aristotle, metaphysical. *J Craniofac Surg* **25**, 59–63 (2014).

71. More Is Different, P. W. Anderson. Volume **177**, (1972).

72. Bertalanffy, L. von. *Théorie générale des systèmes*. (Dunod, 2012).

73. Kohl, P. & Noble, D. Systems biology and the virtual physiological human. *Mol Syst Biol* **5**, 292 (2009).

74. Gentleman, W. A chronology of plankton dynamics in silico: how computer models have been used to study marine ecosystems. *Hydrobiologia* **480**, 69–85 (2002).

75. *From Populations to Ecosystems,* Michel Loreau. (2010).

76. Heino, J. *et al.* Metacommunity organisation, spatial extent and dispersal in aquatic systems: patterns, processes and prospects. *Freshwater Biology* **60**, 845–869 (2015).

77. Patten, B. C. Mathematical Models of Plankton Production. *Internationale Revue der gesamten Hydrobiologie und Hydrographie* **53**, 357–408 (1968).

78. Follows, M. J. & Dutkiewicz, S. modelling diverse communities of marine microbes. *Ann Rev Mar Sci* **3**, 427–451 (2011).

79. The Oceans Their Physics, Chemistry, and General Biology. https://publishing.cdlib.org/ucpressebooks/view?docId=kt167nb66r.

80. Anderson, J. P. E. & Domsch, K. H. A physiological method for the quantitative measurement of microbial biomass in soils. *Soil Biology and Biochemistry* **10**, 215–221 (1978).

81. Martin, J. H. & Fitzwater, S. E. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331**, 341–343 (1988).

82. Fasham, M. J. R., Sarmiento, J. L., Slater, R. D., Ducklow, H. W. & Williams, R. Ecosystem behaviour at Bermuda Station “S” and ocean weather station “India”: A general circulation model and observational analysis. *Global Biogeochemical Cycles* **7**, 379–415 (1993).

83. Maier-Reimer, E., Mikolajewicz, U. & Winguth, A. Future ocean uptake of CO2: interaction between ocean circulation and

biology: *Climate Dynamics* **12**, 711–722 (1996).

84. Aumont: Dimethylsulfoniopropionate (DMSP) and dimethylsulfide (DMS) sea surface distributions simulated from a global three-dimensional ocean carbon cycle model (2002).

85. Ward, B. A., Dutkiewicz, S., Jahn, O. & Follows, M. J. A size-structured food-web model for the global ocean. *Limnology and Oceanography* **57**, 1877–1891 (2012).

86. Heinle, A. & Slawig, T. Theoretical Analysis and Optimisation of Nonlinear ODE Systems for Marine Ecosystem Models. in *System modelling and Optimisation* (eds. Hömberg, D. & Tröltzsch, F.) 501–510 (Springer, 2013). doi:10.1007/978-3-642-36062-6_50.

87. Siedlecki, S. A. *et al.* Seasonal and interannual oxygen variability on the Washington and Oregon continental shelves. *Journal of Geophysical Research: Oceans* **120**, 608–633 (2015).

88. McGill, B. J., Enquist, B. J., Weiher, E. & Westoby, M. Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution* **21**, 178–185 (2006).

89. Chai, F., Dugdale, R. C., Peng, T.-H., Wilkerson, F. P. & Barber, R. T. One-dimensional ecosystem model of the equatorial Pacific upwelling system. Part I: model development and silicon and nitrogen cycle. *Deep Sea Research Part II: Topical Studies in Oceanography* **49**, 2713–2745 (2002).

90. Smetacek, V. Diatoms and the Ocean Carbon Cycle. *Protist* **150**, 25–32 (1999).

91. Moore, J. K., Doney, S. C., Kleypas, J. A., Glover, D. M. & Fung, I. Y. An intermediate complexity marine ecosystem model for the global domain. *Deep Sea Research Part II: Topical Studies in Oceanography* **49**, 403–462 (2001).

92. Zeebe_CO2_In_Seawater_Ch_1.pdf.

93. Downward transport and fate of organic matter in the ocean: Simulations with a general circulation model - Najjar - 1992 - Global Biogeochemical Cycles - Wiley Online Library. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/91GB02718.

94. Maier-Reimer, E. Geochemical cycles in an ocean general circulation model. Preindustrial tracer distributions. *Global Biogeochemical Cycles* **7**, 645–677 (1993).

95. Gregg, W. W., Ginoux, P., Schopf, P. S. & Casey, N. W. Phytoplankton and iron: validation of a global three-dimensional ocean biogeochemical model. *Deep Sea Research Part II: Topical Studies in Oceanography* **50**, 3143–3169 (2003).

96. Aumont, O., Maier-Reimer, E., Blain, S. & Monfray, P. An ecosystem model of the global ocean including Fe, Si, P colimitations. *Global Biogeochemical Cycles* **17**, (2003).

97. Quéré, C. L. *et al.* Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology* **11**, 2016–2040 (2005).

98. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent Biogeography of Microbial Communities in a Model Ocean. *Science* **315**, 1843–1846 (2007).

99. Bopp, L. *et al.* Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* **10**, 6225–6245 (2013).

100. Faure, E., Ayata, S.-D. & Bittner, L. Towards omics-based predictions of planktonic functional composition from environmental data. *Nat Commun* **12**, 4361 (2021).

101. Vernette, C. *et al.* The Ocean Gene Atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. *Nucleic Acids Res* **50**, W516–W526 (2022).

102. Fang, X., Lloyd, C. J. & Palsson, B. O. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat Rev Microbiol* **18**, 731–743 (2020).

103. Budinich, M., Bourdon, J., Larhlimi, A. & Eveillard, D. A multi-objective constraint-based approach for modelling genome-scale microbial ecosystems. *PLOS ONE* **12**, e0171744 (2017).

104. Diener, C., Gibbons, S. M. & Resendis-Antonio, O. MICOM: Metagenome-Scale modelling To Infer Metabolic Interactions in the Gut Microbiota. *mSystems* **5**, e00606-19 (2020).

105. Régimbeau, A. *et al.* Contribution of genome-scale metabolic modelling to niche theory. *Ecology Letters* (2022) doi:10.1111/ele.13954.

106. Zomorrodi, A. R. & Maranas, C. D. OptCom: A Multi-Level Optimisation Framework for the Metabolic modelling and Analysis of Microbial Communities. *PLoS Comput Biol* **8**, e1002363 (2012).

107. Machado, D. *et al.* Polarisation of microbial communities between competitive and cooperative metabolism. *Nat Ecol Evol* **5**, 195–203 (2021).

108. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**, 329–342 (2012).

109. Thiele, I. & Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature*

Protocols **5**, 93–121 (2010).

110. Kschischo, M. A gentle introduction to the thermodynamics of biochemical stoichiometric networks in steady state. *Eur. Phys. J. Spec. Top.* **187**, 255–274 (2010).

111. Xavier, J. C., Patil, K. R. & Rocha, I. Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. *Metab Eng* **39**, 200–208 (2017).

112. Tasic, M., Rios, P., Santos, F., Filipini, F. & Maciel, F. Estimating the elemental biomass composition of Desmodesmus sp. cultivated in sugarcane stillage. *Adv techn* **5**, 33–37 (2016).

113. Levering, J. *et al.* Genome-Scale Model Reveals Metabolic Basis of Biomass Partitioning in a Model Diatom. *PLoS One* **11**, e0155038 (2016).

114. Lachance, J.-C. *et al.* BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Comput Biol* **15**, e1006971 (2019).

115. Gianchandani, E. P., Oberhardt, M. A., Burgard, A. P., Maranas, C. D. & Papin, J. A. Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics* **9**, 43 (2008).

116. Friis, J. C., Holm, C. & Halling-Sørensen, B. Evaluation of elemental composition of algal biomass as toxical endpoint. *Chemosphere* **37**, 2665–2676 (1998).

117. Feist, A. M. & Palsson, B. O. The Biomass Objective Function. *Curr Opin Microbiol* **13**, 344–349 (2010).

118. Cheung, C. Y. M. *et al.* A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *The Plant Journal* **75**, 1050–1061 (2013).

119. Beck, A. E., Hunt, K. A. & Carlson, R. P. Measuring Cellular Biomass Composition for Computational Biology Applications. *Processes* **6**, 38 (2018).

120. Feist, A. M. & Palsson, B. O. The Biomass Objective Function. *Curr Opin Microbiol* **13**, 344–349 (2010).

121. Geider, R. & Roche, J. L. Redfield revisited: variability of C:N:P in marine microalgae and its biochemical basis. *European Journal of Phycology* **37**, 1–17 (2002).

122. Edwards, J. S. & Palsson, B. O. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* **97**, 5528–5533 (2000).

123. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl Environ Microbiol* **60**, 3724–3731 (1994).

124. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* **99**, 15112–15117 (2002).

125. Lee, J. M., Gianchandani, E. P. & Papin, J. A. Flux balance analysis in the era of metabolomics. *Brief Bioinform* **7**, 140–150 (2006).

126. Maarleveld, T. R., Khandelwal, R. A., Olivier, B. G., Teusink, B. & Bruggeman, F. J. Basic concepts and principles of stoichiometric modelling of metabolic networks. *Biotechnol J* **8**, 997–1008 (2013).

127. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* **15**, 107–120 (2014).

128. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* **10**, 291–305 (2012).

129. Price, N., Reed, J. & Palsson, B. Genome-Scale Models Of Microbial Cells: Evaluating The Consequences Of Constraints. *Nature reviews. Microbiology* **2**, 886–97 (2004).

130. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* **44**, D515–D522 (2016).

131. Spielmeyer, A. & Pohnert, G. Influence of temperature and elevated carbon dioxide on the production of dimethylsulfoniopropionate and glycine betaine by marine phytoplankton. *Mar Environ Res* **73**, 62–69 (2012).

132. Niang, G. *et al.* METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES . *F1000Research* **9**, (2020).

133. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research* **46**, 7542–7553 (2018).

134. Richter, D. J. *et al.* EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. **2**, (2022).

135. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**, D1178–D1186 (2012).

136. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

137. Guiry, M. D. *et al.* AlgaeBase: An On-line Resource for Algae. *Cryptogamie, Algologie* **35**, 105–115 (2014).

138. Montsant, A., Bowler, C. & Lopez, P. Diatomics: Toward Diatom Functional Genomics. *Journal of nanoscience and nanotechnology* **5**, 5–14 (2005).

139. Schellenberger, J., Park, J. O., Conrad, T. M. & Palsson, B. Ø. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213 (2010).

140. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* **20**, 1085–1093 (2019).

141. Chang, R. L. *et al.* Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol Syst Biol* **7**, 518 (2011).

142. Knies, D. *et al.* modelling and Simulation of Optimal Resource Management during the Diurnal Cycle in Emiliania huxleyi by Genome-Scale Reconstruction and an Extended Flux Balance Analysis Approach. *Metabolites* **5**, 659 (2015).

143. Juneja, A., Chaplen, F. W. R. & Murthy, G. S. Genome scale metabolic reconstruction of Chlorella variabilis for exploring its metabolic potential for biofuels. *Bioresour Technol* **213**, 103–110 (2016).

144. Levering, J. *et al.* Genome-Scale Model Reveals Metabolic Basis of Biomass Partitioning in a Model Diatom. *PLoS One* **11**, e0155038 (2016).

145. Broddrick, J. T. *et al.* Cross-compartment metabolic coupling enables flexible photoprotective mechanisms in the diatom Phaeodactylum tricornutum. *New Phytologist* **222**, 1364–1379 (2019).

146. Nègre, D. *et al.* Genome-Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae Saccharina japonica and Cladosiphon okamuranus. *Antioxidants (Basel)* **8**, 564 (2019).

147. Ravikrishnan, A. & Raman, K. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Brief. Bioinformatics* **16**, 1057–1068 (2015).

148. Moretti, S., Tran, V. D. T., Mehl, F., Ibberson, M. & Pagni, M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Res* **49**, D570–D574 (2021).

149. Kroth, P. G. *et al.* A Model for Carbohydrate Metabolism in the Diatom Phaeodactylum tricornutum Deduced from Comparative Whole Genome Analysis. *PLOS ONE* **3**, e1426 (2008).

150. Río Bártulos, C. *et al.* Mitochondrial Glycolysis in a Major Lineage of Eukaryotes. *Genome Biol Evol* **10**, 2310–2325 (2018).

151. Vilhena, M. do P. S. P., Costa, M. L. da, Berrêdo, J. F., Paiva, R. S. & Almeida, P. D. Chemical composition of phytoplankton from the estuaries of Eastern Amazonia. *Acta Amazonica* **44**, 513–526 (2014).

152. Rodrigues, D. Chemical composition of red, brown and green macroalgae from Buarcos bay in Central West Coast of Portugal. *Food Chemistry* 11 (2015).

153. Tasić, M. B. (Faculty of T., Rios, P. L. F. (Faculty of C. E. (FEQ), Fernandes, D. S. (Faculty of C. E. (FEQ), Ferreira, G. F. (Faculty of C. E. (FEQ) & Filho, R. M. (Faculty of C. E. (FEQ). Estimating the elemental biomass composition of Desmodesmus Sp. cultivated in sugarcane stillage. *Savremene tehnologije / Advanced Technologies* (2016).

154. Nikdel, A., Braatz, R. D. & Budman, H. M. A systematic approach for finding the objective function and active constraints for dynamic flux balance analysis. *Bioprocess Biosyst Eng* **41**, 641–655 (2018).

155. Lachance, J.-C. *et al.* BOFdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Comput Biol* **15**, e1006971 (2019).

156. Davidsen, T. *et al.* The comprehensive microbial resource. *Nucleic Acids Res* **38**, D340–D345 (2010).

157. de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).

158. Kramer, D. M., Sacksteder, C. A. & Cruz, J. A. How acidic is the lumen? *Photosynthesis Research* **60**, 151–163 (1999).

159. Trinh, M. D. L. & Masuda, S. Chloroplast pH Homeostasis for the Regulation of Photosynthesis. *Front Plant Sci* **13**, 919896 (2022).

160. Poburko, D., Santo-Domingo, J. & Demaurex, N. Dynamic regulation of the mitochondrial proton gradient during cytosolic calcium elevations. *J Biol Chem* **286**, 11672–11684 (2011).

161. Hou, Y. *et al.* Ca2+-associated triphasic pH changes in mitochondria during brown adipocyte activation. *Mol Metab* **6**, 797–808 (2017).

162. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **42**, D459–D471 (2014).

163. Xu, Z., Sun, X. & Yu, S. Genome-scale analysis to the impact of gene deletion on the metabolism of E. coli: constraint-based simulation approach. *BMC Bioinformatics* **10**, S62 (2009).

164. Shaw, D. K., Sekar, J. & Ramalingam, P. V. Recent insights into oceanic dimethylsulfoniopropionate biosynthesis and catabolism. *Environmental Microbiology* **24**, 2669–2700 (2022).

165. Gage, D. A. *et al.* A new route for synthesis of dimethylsulphoniopropionate in marine algae. *Nature* **387**, 891–894 (1997).

166. Summers, P. S. *et al.* Identification and Stereospecificity of the First Three Enzymes of 3-Dimethylsulfoniopropionate Biosynthesis in a Chlorophyte Alga1. *Plant Physiology* **116**, 369–378 (1998).

167. Otte, M. L., Wilson, G., Morris, J. T. & Moran, B. M. Dimethylsulphoniopropionate (DMSP) and related compounds in higher plants. *Journal of Experimental Botany* **55**, 1919–1925 (2004).

168. Lyon, B. R., Lee, P. A., Bennett, J. M., DiTullio, G. R. & Janech, M. G. Proteomic Analysis of a Sea-Ice Diatom: Salinity Acclimation Provides New Insight into the Dimethylsulfoniopropionate Production Pathway. *Plant Physiology* **157**, 1926–1941 (2011).

169. Uchida, A., Ooguri, T., Ishida, T., Kitaguchi, H. & Ishida, Y. Biosynthesis of Dimethylsulfoniopropionate in Crypthecodinium Cohnii (Dinophyceae). in *Biological and Environmental Chemistry of DMSP and Related Sulfonium Compounds* (eds. Kiene, R. P., Visscher, P. T., Keller, M. D. & Kirst, G. O.) 97–107 (Springer US, 1996). doi:10.1007/978-1-4613-0377-0_9.

170. Summers, P. S. *et al.* Identification and Stereospecificity of the First Three Enzymes of 3-Dimethylsulfoniopropionate Biosynthesis in a Chlorophyte Alga. *Plant Physiol* **116**, 369–378 (1998).

171. Wang, J. *et al. Novel dimethylsulfoniopropionate biosynthesis enzymes in diverse marine bacteria, cyanobacteria and abundant algae*. https://www.researchsquare.com/article/rs-2678769/v1 (2023) doi:10.21203/rs.3.rs-2678769/v1.

172. Guérin, N. *et al.* Genomic adaptation of the picoeukaryote Pelagomonas calceolata to iron-poor oceans revealed by a chromosome-scale genome sequence. *Commun Biol* **5**, 1–14 (2022).

173. O'Brien, J. *et al.* Biogeographical and seasonal dynamics of the marine Roseobacter community and ecological links to DMSP-producing phytoplankton. *ISME Commun* **2**, 16 (2022).

174. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

175. Box, G. E. P. Science and Statistics. *Journal of the American Statistical Association* **71**, 791–799 (1976).

176. Dutreuil, S. What good are abstract and what-if models? Lessons from the Gaïa hypothesis. *HPLS* **36**, 16–41 (2014).

177. Bernstein, D. B., Sulheim, S., Almaas, E. & Segrè, D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology* **22**, 64 (2021).

178. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

179. de Vienne, D. What is a phenotype? History and new developments of the concept. *Genetica* **150**, 153–158 (2022).

180. van Tol, H. M. & Armbrust, E. V. Genome-scale metabolic model of the diatom Thalassiosira pseudonana highlights the importance of nitrogen and sulphur metabolism in redox balance. *PLOS ONE* **16**, e0241960 (2021).

181. Roig, X. Aquatic Photosynthesis. Paul G. Falkowski, John A. Raven. *International Microbiology* **3**, 259 (2000).

182. Broddrick, J. T. *et al.* Cross-compartment metabolic coupling enables flexible photoprotective mechanisms in the diatom Phaeodactylum tricornutum. *New Phytologist* **222**, 1364–1379 (2019).

183. Saks, V., Beraud, N. & Wallimann, T. Metabolic Compartmentation – A System Level Property of Muscle Cells. *Int J Mol Sci* **9**, 751–767 (2008).

184. Hinzpeter, F., Gerland, U. & Tostevin, F. Optimal Compartmentalisation Strategies for Metabolic Microcompartments. *Biophys J* **112**, 767–779 (2017).

185. van Tol, H. M. & Armbrust, E. V. Genome-scale metabolic model of the diatom Thalassiosira pseudonana highlights the importance of nitrogen and sulphur metabolism in redox balance. *PLoS ONE* **16**, e0241960 (2021).

186. *Biological and Environmental Chemistry of DMSP and Related Sulfonium Compounds*. (Springer US, 1996). doi:10.1007/978-1-4613-0377-0.

187. Dickschat, J. S., Rabe, P. & Citron, C. A. The chemical biology of dimethylsulfoniopropionate. *Org Biomol Chem* **13**, 1954–1968 (2015).

188. Kempf, B. & Bremer, E. Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. *Arch Microbiol* **170**, 319–330 (1998).

189. Eitinger, T., Rodionov, D. A., Grote, M. & Schneider, E. Canonical and ECF-type ATP-binding cassette importers in prokaryotes: diversity in modular organisation and cellular functions. *FEMS Microbiol Rev* **35**, 3–67 (2011).

190. Ziegler, C., Bremer, E. & Krämer, R. The BCCT family of carriers: from physiology to crystal structure. *Mol Microbiol* **78**, 13–34 (2010).

191. Vila-Costa, M. *et al.* Dimethylsulfoniopropionate Uptake by Marine Phytoplankton. *Science* **314**, 652–654 (2006).

192. Alcolombri, U. *et al.* MARINE sulphur CYCLE. Identification of the algal dimethyl sulfide-releasing enzyme: A missing link in

the marine sulphur cycle. *Science* **348**, 1466–1469 (2015).

193. Bullock, H. A., Luo, H. & Whitman, W. B. Evolution of Dimethylsulfoniopropionate Metabolism in Marine Phytoplankton and Bacteria. *Front Microbiol* **8**, 637 (2017).

194. Bernstein, D. B., Sulheim, S., Almaas, E. & Segrè, D. Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology* **22**, 64 (2021).

195. Noble, D. Claude Bernard, the first systems biologist, and the future of physiology: Systems biology and the future of physiology. *Experimental Physiology* **93**, 16–26 (2008).

196. Gottesman, I. I. & Gould, T. D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**, 636–645 (2003).

197. Dawkins, R. *The Extended Phenotype: The Long Reach of the Gene*. (Oxford University Press, 1999).

198. Dawkins, R. *The selfish gene*. (Oxford University Press, 2006).

199. Charlesworth, B. Levels of Selection in Evolution. *Heredity* **84**, 493–493 (2000).

200. Gould, S. J. *The Structure of Evolutionary Theory:* (Belknap Press, 2002).

201. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

202. LA LOGIQUE DU VIVANT - Fiche de lecture - Encyclopædia Universalis. https://www.universalis.fr/encyclopedie/la-logique-du-vivant/.

203. Monod, J. & Jacob, F. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb Symp Quant Biol* **26**, 389–401 (1961).

204. Morgan, T. H. The mechanism of Mendelian heredity. http://www.columbia.edu/cu/lweb/digital/collections/cul/texts/ldpd_5998129_000/.

205. Waddington, C. H. An introduction to modern genetics. *An introduction to modern genetics.* (1939).

206. Waddington, C. H. Organisers and Genes. *Organisers and Genes.* (1940).

207. Waddington, C. H. *The Strategy of the Genes*. (Routledge, 2014).

208. Stephens, N., Flynn, K. J. & Gallon, J. R. Interrelationships between the pathways of inorganic nitrogen assimilation in the cyanobacterium Gloeothece can be described using a mechanistic mathematical model. *New Phytologist* **160**, 545–555 (2003).

209. Cavalier-Smith, T. Membrane heredity and early chloroplast evolution. *Trends in Plant Science* **5**, 174–182 (2000).

210. Noble, D. Biophysics and systems biology. *Philos Trans A Math Phys Eng Sci* **368**, 1125–1139 (2010).

211. The membranome and membrane heredity in development and evolution. in *Organelles, Genomes and Eukaryote Phylogeny* (ed. Horner, R. P. H., David S.) (CRC Press, 2004).

212. Michie, K. A. & Löwe, J. Dynamic Filaments of the Bacterial Cytoskeleton. *Annual Review of Biochemistry* **75**, 467–492 (2006).

213. Fuerst, J. A. Intracellular Compartmentation in Planctomycetes. *Annual Review of Microbiology* **59**, 299–328 (2005).

214. Sun, Y.-H., Chen, S.-P., Wang, Y.-P., Hu, W. & Zhu, Z.-Y. Cytoplasmic Impact on Cross-Genus Cloned Fish Derived from Transgenic Common Carp (Cyprinus carpio) Nuclei and Goldfish (Carassius auratus) Enucleated Eggs1. *Biology of Reproduction* **72**, 510–515 (2005).

215. Gugliotta, G. Cloned Ox Dies From Infection. *Washington Post* (2001).

216. Keller, E. *The Century of the Gene*. (2002).

217. Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. (Belknap Press, 1985).

218. Woese, C. R. & Goldenfeld, N. How the Microbial World Saved Evolution from the Scylla of Molecular Biology and the Charybdis of the Modern Synthesis. *Microbiol Mol Biol Rev* **73**, 14–21 (2009).

219. Goldenfeld, N. & Woese, C. Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium. *Annual Review of Condensed Matter Physics* **2**, 375–399 (2011).

220. Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat Rev Genet* **11**, 855–866 (2010).

221. Bhattacharya, S., Zhang, Q. & Andersen, M. E. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC Systems Biology* **5**, 85 (2011).

222. Thummarati, P. *et al.* Recent Advances in Cell Sheet Engineering: From Fabrication to Clinical Translation. *Bioengineering (Basel)* **10**, 211 (2023).

223. Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat Rev Genet* **11**, 855–866 (2010).

224. Kingsolver, J. G., Gomulkiewicz, R. & Carter, P. A. Variation, selection and evolution of function-valued traits. *Genetica* **112–**

113, 87–104 (2001).

225. Popper, CONJECTURES AND REFUTATIONS (1962).

226. Kohl, P., Noble, D., Winslow, R. & Hunter, P. Computational modelling of biological systems: Tools and visions. *Philosophical Transactions of The Royal Society B Biological Sciences* **358**, 579 (2000).

227. Saab, M. A.-A. Day-to-day variation in phytoplankton assemblages during spring blooming in a fixed station along the Lebanese coastline. *Journal of Plankton Research* **14**, 1099–1115 (1992).

228. Djurhuus, A. *et al.* Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications* **11**, (2020).

229. Kavanaugh, M. *et al.* Seascapes as a new vernacular for pelagic ocean monitoring, management and conservation. *ICES Journal of Marine Science: Journal du Conseil* **73**, fsw086 (2016).

230. Fay, A. R. & McKinley, G. A. Global open-ocean biomes: mean and temporal variability. *Earth System Science Data* **6**, 273–284 (2014).

231. Reygondeau, G. *et al.* Dynamic biogeochemical provinces in the global ocean. *Global Biogeochemical Cycles* **27**, 1046–1058 (2013).

232. Dutkiewicz, S. *et al.* Dimensions of marine phytoplankton diversity. *Biogeosciences* **17**, 609–634 (2020).

233. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model | Science. https://www.science.org/doi/10.1126/science.1254421.

234. Laso-Jadart, R. *et al.* Investigating population-scale allelic differential expression in wild populations of Oithona similis (Cyclopoida, Claus, 1866). *Ecology and Evolution* **10**, 8894–8905 (2020).

235. Elshourbagy, N. A. *et al.* Cloning and expression of a human ATP-citrate lyase cDNA. *Eur J Biochem* **204**, 491–499 (1992).

236. Sun, T., Hayakawa, K., Bateman, K. S. & Fraser, M. E. Identification of the Citrate-binding Site of Human ATP-Citrate Lyase Using X-ray Crystallography *. *Journal of Biological Chemistry* **285**, 27418–27428 (2010).

237. Wiegand, G. & Remington, S. J. Citrate synthase: structure, control, and mechanism. *Annu Rev Biophys Biophys Chem* **15**, 97–117 (1986).

238. Hippmann, A. A. *et al.* Proteomic analysis of metabolic pathways supports chloroplast–mitochondria cross-talk in a Cu-limited diatom. *Plant Direct* **6**, e376 (2022).

239. Guillard, R. R. & Ryther, J. H. Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt, and Detonula confervacea (cleve) Gran. *Can J Microbiol* **8**, 229–239 (1962).

240. Stuani, L. *et al.* Novel metabolic features in Acinetobacter baylyi ADP1 revealed by a multiomics approach. *Metabolomics* **10**, 1223–1238 (2014).

241. Rousk, J. & Bengtson, P. Microbial regulation of global biogeochemical cycles. *Frontiers in Microbiology* **5**, (2014).

242. The Great Oxidation Event: How Cyanobacteria Changed Life. *ASM.org* https://asm.org:443/Articles/2022/February/The-Great-Oxidation-Event-How-Cyanobacteria-Change.

243. Overmann, J. & Lepleux, C. Marine Bacteria and Archaea: Diversity, Adaptations, and Culturability. in *The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential* (eds. Stal, L. J. & Cretoiu, M. S.) 21–55 (Springer International Publishing, 2016). doi:10.1007/978-3-319-33000-6_2.

244. Bar-On, Y. M., Phillips, R. & Milo, R. The biomass distribution on Earth. *Proc Natl Acad Sci U S A* **115**, 6506–6511 (2018).

245. Brussaard, C. P. D. Viral control of phytoplankton populations--a review. *J Eukaryot Microbiol* **51**, 125–138 (2004).

246. Stefels, J. Physiological aspects of the production and conversion of DMSP in marine algae and higher plants. *Journal of Sea Research* **43**, 183–197 (2000).

247. Escobar-Zepeda, A., Vera-Ponce de León, A. & Sanchez-Flores, A. The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Front Genet* **6**, 348 (2015).

248.Three Prochlorococcus Cyanophage Genomes: Signature Features and Ecological Interpretations - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1079782/.

249. Fleming, R. H. The Control of Diatom Populations by Grazing. *ICES Journal of Marine Science* **14**, 210–227 (1939).

250. Palsson, B. The challenges of in silico biology. *Nat Biotechnol* **18**, 1147–1150 (2000).

251. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* **5**, 264–276 (2003).

252. The EcoCyc Database in 2021 - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8357350/.

253. Mendoza, S. N., Olivier, B. G., Molenaar, D. & Teusink, B. A systematic assessment of current genome-scale metabolic

reconstruction tools. *Genome Biology* **20**, 158 (2019).

254. Li, Y. *et al.* High-throughput phenotyping analysis of maize at the seedling stage using end-to-end segmentation network. *PLOS ONE* **16**, e0241528 (2021).

255. Sagan, L. On the origin of mitosing cells. *J Theor Biol* **14**, 255–274 (1967).

256. Zakharova, L., Seifan, M. & Meyer, K. Trait-based modelling in ecology: A review of two decades of research. *Ecological Modelling* **407**, (2019).

257. Reuveny, Z., Dougall, D. K. & Trinity, P. M. Regulatory coupling of nitrate and sulphate assimilation pathways in cultured tobacco cells. *Proc Natl Acad Sci U S A* **77**, 6670–6672 (1980).

258. McParland, E. L., Wright, A., Art, K., He, M. & Levine, N. M. Evidence for contrasting roles of dimethylsulfoniopropionate production in Emiliania huxleyi and Thalassiosira oceanica. *New Phytol* **226**, 396–409 (2020).

259. Cooke, R. J., Oliver, J. & Davies, D. D. Stress and Protein Turnover in Lemna minor. *Plant Physiol* **64**, 1109–1113 (1979).

260. Kettles, N. L., Kopriva, S. & Malin, G. Insights into the Regulation of DMSP Synthesis in the Diatom Thalassiosira pseudonana through APR Activity, Proteomics and Gene Expression Analyses on Cells Acclimating to Changes in Salinity, Light and Nitrogen. *PLoS One* **9**, e94795 (2014).

261. HilleRisLambers, J., Adler, P. B., Harpole, W. S., Levine, J. M. & Mayfield, M. M. Rethinking Community Assembly through the Lens of Coexistence Theory. *Annual Review of Ecology, Evolution, and Systematics* **43**, 227–248 (2012).

262. Jousset, A., Eisenhauer, N., Materne, E. & Scheu, S. Evolutionary history predicts the stability of cooperation in microbial communities. *Nat Commun* **4**, 2573 (2013).

263. Helliwell, K. E. The roles of B vitamins in phytoplankton nutrition: new perspectives and prospects. *New Phytol* **216**, 62–68 (2017).

264. Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J. & Smith, A. G. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**, 90–93 (2005).

265. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* **20**, 947–959 (2010).

266 Basterretxea, G., Font-Muñoz, J. S., Hernández-Carrasco, I. & Sañudo-Wilhelmy, S. *Global variability of high nutrient low chlorophyll regions using neural networks and wavelet coherence analysis*. https://egusphere.copernicus.org/preprints/2022/egusphere-2022-827/ (2022) doi:10.5194/egusphere-2022-827.

267. Martin, J. H. & Fitzwater, S. E. Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331**, 341–343 (1988).

268. Moënne-Loccoz, Y., Mavingui, P., Combes, C., Normand, P. & Steinberg, C. Microorganisms and Biotic Interactions. *Environmental Microbiology: Fundamentals and Applications* 395–444 (2014) doi:10.1007/978-94-017-9118-2_11.

269. Goyal, A. & Maslov, S. Diversity, Stability, and Reproducibility in Stochastically Assembled Microbial Ecosystems. *Phys. Rev. Lett.* **120**, 158102 (2018).

270. Klamt, S. & Stelling, J. Combinatorial Complexity of Pathway Analysis in Metabolic Networks. *Mol Biol Rep* **29**, 233–236 (2002).

271. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat Biotechnol* **28**, 245–248 (2010).

272. Coles, V. J. *et al.* Ocean biogeochemistry modelled with emergent trait-based genomics. *Science* **358**, 1149–1154 (2017).

273. Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

274. Belcour, A. *et al.* AuCoMe: inferring and comparing metabolisms across heterogeneous sets of annotated genomes. 2022.06.14.496215 Preprint at https://doi.org/10.1101/2022.06.14.496215 (2022).

275. Aite, M. *et al.* Traceability, reproducibility and wiki-exploration for “à-la-carte” reconstructions of genome-scale metabolic models. *PLoS Comput Biol* **14**, e1006146 (2018).

DEFINITIONS

- Aerobic : a condition or process that occurs in presence of oxygen.
- Albedo : refers to the measure of the reflectivity of a surface, specifically how much sunlight or solar radiation is reflected back into space.
- Anabolism : metabolic process in living organisms in which complex molecules are synthesised or built up from simpler molecules, requiring energy input.
- Anaerobic : a condition or process that occurs in absence of oxygen.
- Atmosphere : the layer of gases that surround the Earth, including air, water vapor, and other gases.
- Autotrophy : biological process in which organisms synthesise organic compounds from inorganic substances allowing a self-feeding.
- Auxotrophy : inability to produce essential compounds, requiring external supply.
- Biosphere : all living organisms and their interactions with each other and with their environment.
- Blocked reaction : reaction in a metabolic model that cannot proceed under the given conditions, resulting in the absence of flux through that reaction.
- Bloom : rapid increase or accumulation of planktonic organisms, such as algae or other microscopic organisms, in a body of water. Blooms can be caused by various factors including nutrient availability, favorable environmental conditions, and ecological dynamics, and they can have significant impacts on marine ecosystems.
- Catabolism : metabolic process in living organisms that break down complex molecules into simpler ones, releasing energy in the process.
- Chemosynthesis : process where certain organisms use chemical reactions to produce organic molecules as a source of energy, instead of relying on sunlight.
- Chloroplaste : specialised organelle found in plant cells and some protists, responsible for photosynthesis.
- Coding sequence : part of a genome that contains the instructions for a protein.
- *De novo* (in genomics) : refers to the process of assembling or constructing something from scratch or without reference to a pre-existing template or sequence.
- Duplicated reaction : redundant or replicated metabolic reaction in a metabolic model.
- Ecosystem : community of living organisms (including plants, animals, or microorganisms) in conjunction with their physical environment, interacting as a system.
- Electron transport chain : series of protein complexes and molecules in the inner thylakoid membrane in chloroplasts (or mitochondrial membrane) that transfer electrons during photosynthesis (or cellular respiration) to produce a proton gradient for ATP synthesis.

- Eukaryote : organism whose cells have a nucleus and other membrane-bound organelles. Eukaryotes include organisms such as plants, animals, fungi, and protists.
- Gene expression : process by which the information encoded in a gene is used to synthesise a functional gene product, such as a protein or RNA molecule. It involves transcription of the gene into messenger RNA (mRNA) and the subsequent translation of mRNA into a protein. Gene expression is tightly regulated and can be influenced by various factors, including cellular signals, environmental cues, and developmental stages.
- Geosphere : the solid Earth, including rocks, minerals, and the Earth's interior.
- Heterotrophy : mode of nutrition in which an organism obtains its energy and nutrients by consuming organic matter from other organisms.
- Holobiont : concept emphasising the interconnectedness and mutual dependencies between the host and its associated microbiota. It recognizes that many organisms exist in symbiotic relationships with their microbial partners and that these partnerships play a significant role in the overall health, development, and function of the host organism (e.g. gut microbiota).
- Hydrosphere: all the water on Earth, including oceans, lakes, rivers, groundwater, and ice.
- *in silico :* performed or occurring in a computational environment.
- *in vitro :* in a controlled laboratory environment outside of a living organism.
- *in vivo :* inside a living organism.
- Ions : atom or molecule with a net electrical charge due to the gain or loss of electrons.
- Isomer : different forms of the same molecule.
- Lithosphere : the outermost layer of the Earth's crust, including the solid and brittle portion of the Earth's surface.
- Lithotrophy : metabolic process in which certain microorganisms derive energy by oxidising inorganic compounds (such as ammonia, nitrites, sulphur compounds, or iron compounds as their energy source instead of organic matter).
- Metabolism : chemical processes that occur within an organism to sustain life and enable its growth, development, and maintenance.
- Microbe : a microorganism.
- Pangenome : entire set of genes found in all individuals of a particular species, including both core genes present in all individuals and accessory genes that are unique to specific individuals or subsets of the population.
- Phenome : set of observable characteristics and traits of an organism, resulting from the interaction of its genotype with the environment.
- Photosynthesis (oxygenic) : process by which some organisms convert sunlight, carbon dioxide, and water into glucose and oxygen.
- Phototrophy : ability of organisms to obtain energy from light.
- Phycosphere : region immediately surrounding a phytoplankton cell where microorganisms

and organic matter interact.

- Pleiotropy : phenomenon where a single gene or genetic variant influences multiple traits or phenotypic characteristics.
- Polygeny : refers to the inheritance of a trait or phenotype that is controlled by multiple genes.
- Polyoside : refers to a polysaccharide, which is a complex carbohydrate composed of multiple sugar units linked together (cellulose, starch, glycogen, chitin).
- Polyphyletic group : artificial group that includes multiple species but does not include the most recent common ancestor of those species (plankton).
- Prokaryote : single-celled organism that lacks a distinct nucleus (bacteria).
- Protists :diverse group of eukaryotic microorganisms that are not classified as plants, animals, or fungi.
- Protozoa : single-celled organisms found in water and soil that can be parasitic or free-living, playing roles in nutrient cycling and some causing diseases (part of the animal kingdom).
- sulphur reduction : microbial process using sulphur compounds as electron acceptors.
- Transcription : process of synthesising RNA from a DNA template.
- Translation : process of protein synthesis where the sequence of mRNA is decoded to produce a specific amino acid sequence.
- Trophism : refers to the nutritional or energy requirements of an organism, particularly in relation to its interactions with its environment or other organisms.

ACRONYMS

- ATP : Adenosine TriPhosphate
- ATPS : ATP Synthase (enzyme)
- CBM : Constraint-Based metabolic Model
- DM : Demand reaction (metabolic modelling)
- DMS : DiMethyl Sulfide
- DMSP : DiMethylSulfonioPropionate
- GSM : Genome-Scale metabolic Model (i.g. constraint-based)
- NADPH : Nicotinamide Adenine Dinucleotide Phosphate
- QSSA : Quasi-Steady-State Assumption
- SK : Sink reaction (metabolic modelling)
- MAT : Methionine AminoTransferase
- MTOB : 4-MethylThio-2-OxoButyrate
- MTHB : 4-MethylThio-2-Hydro-oxyButyrate
- DMSHB : 4-DiMethylSulfonio-2-HydroxyButyrate
- MHM : 4-Methylthio-2-Hydro-oxybutyrate Methyltransferase
- PSI : PhotoSystem I
- PSII : PhotoSystem II
- PEPC : PhosphoEnolPyruvate Carboxylase (enzyme)
- PE : PhosphoEnolpyruvate
- PFT : Planktonic Functional Trait or Type (depending of the context)
- CA : Carbonic Anhydrase (enzyme)
- RuBisCo : Ribulose-1,5-Bisphosphate Carboxylase/oxygenase.
- BOF : Biomass Objective Function (metabolic modelling)

5 RÉSUMÉ DÉTAILLÉ EN FRANÇAIS

Cette thèse s'est principalement focalisée sur le développement de PhotoEukStein, un méta-modèle permettant la reconstruction entièrement automatique de modèles métaboliques à base de contraintes (CBMs) pour les microalgues eucaryotes à l'échelle du génome.

5.1 CHAPITRE 1 : CONTEXTES BIOLOGIQUES ET DE MODÉLISATION

Le chapitre 1 de cette thèse sert d'introduction, fournissant des concepts essentiels pour comprendre cette recherche. La première partie met en évidence l'importance de la caractérisation des communautés planctoniques microbiennes, en particulier leur rôle dans la régulation du système terrestre. La partie suivante nous éclaire sur les données disponibles et la modélisation mathématique utilisée aujourd'hui pour décrire les populations planctoniques et leurs fonctions. Pour combler le fossé entre les données environnementales et les modèles existants qui manquent de descriptions détaillées des processus métaboliques, nous proposons d'utiliser des CBMs à l'échelle du génome (GSMs). Bien que la modélisation métabolique ait déjà fait d'importants progrès en écologie, peu a été fait pour les microorganismes eucaryotes.

5.1.1 Les communautés planctoniques marines

Les cycles biogéochimiques sont des processus essentiels qui impliquent la transformation, le transport et le recyclage des molécules sur notre planète. Ces cycles jouent un rôle vital dans le maintien de la vie en régulant la disponibilité des éléments essentiels. En effet, les organismes dépendent d'un approvisionnement continu en molécules spécifiques, tels que les nutriments et les ions, essentiels pour se développer et prospérer. Les composés biochimiques subissent des processus métaboliques au sein des organismes et se transforment en molécules vitales telles que l'ADN, les protéines, les lipides et les glucides. Ces processus métaboliques et transformations matérielles contribuent au fonctionnement global et à la survie des organismes. Les écosystèmes microbiens en particulier jouent un rôle crucial dans le maintien d'un environnement stable et habitable. Ces écosystèmes interagissent avec l'environnement, façonnant le développement et l'évolution des organismes de manière complexe. Les organismes planctoniques marins notamment, jouent un rôle central dans la régulation des grands cycles biogéochimiques, influençant le climat de la Terre. De plus, ils occupent une position cruciale dans les chaînes alimentaires marines, servant de source principale de nourriture pour de nombreux organismes aquatiques. Cependant, ces organismes essentiels sont vulnérables face à divers facteurs de stress environnementaux, tels que la pollution, l'acidification des océans et le changement climatique. Les impacts de ces facteurs de stress peuvent s'étendre à l'ensemble de l'écosystème marin, affectant sa santé et sa stabilité globale. Pour améliorer notre compréhension de la diversité du plancton et de son importance profonde dans la dynamique du système terrestre, un effort collaboratif regroupant divers domaines de recherche est essentiel.

5.1.2 Données disponibles et modélisation mathématique

Actuellement, la disponibilité des métagénomes et métatranscriptomes environnementaux offre des informations précieuses sur la vaste diversité et les rôles fonctionnels du plancton procaryotique et eucaryotique au sein des écosystèmes complexes directement à partir d'échantillons environnementaux. Cependant, il est important de reconnaître que les données omiques seules ne peuvent pas répondre à tous les défis en cours. Bien que ces ensembles de données offrent une mine d'informations, leur intégration dans des modèles mathématiques présente un grand potentiel pour faire avancer notre compréhension. Les GSMs offrent une approche mécaniste en établissant des relations génotype-environnement-phénotype quantitatives et calculables pour les organismes cibles.

En effet, les réseaux métaboliques font référence à l'ensemble des réactions métaboliques (dans les limites des connaissances actuelles) qui se produisent au sein d'une cellule ou d'un organisme. À partir d'un génome, il est possible de prédire les gènes codés et donc d'identifier les enzymes correspondantes et leurs réactions métaboliques associées. Ces réactions constituent l'ensemble des transformations biochimiques vitales pour les organismes (photosynthèse, respiration, anabolisme du DMSP...) qui leur permettent de croître, de se reproduire, de maintenir leur structure et de répondre à leur environnement.

Les GSMs sont des représentations mathématiques de réseaux métaboliques qui tiennent compte des contraintes imposées par la thermodynamique, la stœchiométrie et d'autres facteurs physiologiques. Ces modèles utilisent des algorithmes d'optimisation pour prédire les flux métaboliques ou les taux de croissance dans différentes conditions environnementales ou génotypiques. Ils supposent que le système métabolique est à quasi-état stable, ce qui signifie que les taux de production et de consommation de tous les métabolites intracellulaires sont équilibrés. Cette hypothèse permet le calcul des flux métaboliques sans avoir besoin de données cinétiques détaillées.

L'approvisionnement continu en métabolites depuis et vers le milieu est facilité par les réactions d'échange. Elles sont responsables de l'absorption ou de la sécrétion de nutriments, de produits de déchets ou de molécules de signalisation par les cellules. Les métabolites d'échange sont importants en modélisation métabolique car ils représentent l'interface entre le réseau métabolique et l'environnement externe, et ils peuvent avoir un impact significatif sur le comportement et les propriétés du réseau. Leurs taux d'absorption ou de sécrétion peuvent être contraints en fonction de mesures expérimentales ou estimés à l'aide de méthodes d'optimisation. Si l'échange de métabolites n'était pas possible, alors pour chaque réaction, le seul état possible serait l'équilibre chimique, avec tous les flux nets égaux à zéro.

Initialement, les GSMs sont utilisés pour modéliser la physiologie cellulaire et la croissance d'organismes modèles. Cependant, des extensions de ces approches basées sur les contraintes émergent pour prédire et comprendre les communautés microbiennes. Évidemment, il y a encore place à l'amélioration, notamment en obtenant un nombre suffisant de GSMs capables de représenter avec précision la vaste diversité taxonomique et fonctionnelle du plancton. Actuellement, de nombreux GSMs écologiquement pertinents sont disponibles pour les procaryotes, mais les modèles pour les eucaryotes sont en retard. Ce retard peut être attribué à plusieurs facteurs, notamment la disponibilité limitée d'organismes modèles avec des génomes entièrement séquencés pour le plancton eucaryotique. De plus, la curation manuelle nécessaire pour construire des GSMs efficaces peut être particulièrement fastidieuse et chronophage, en particulier dans les approches

traditionnelles « ascendantes » où la curation doit être effectuée pour chaque nouvelle reconstruction de modèle. Pour relever ces défis, l'approche « descendante » offre une solution prometteuse. Cette approche implique le développement d'un méta-modèle générique qui subit une curation une seule fois. À partir de ce méta-modèle, des modèles spécifiques à chaque organisme peuvent être dérivés, préservant la curation manuelle et les propriétés structurales importantes. Jusqu'à présent, cette technique était uniquement appliquée aux procaryotes.

5.1.3 Objectif principal de la thèse

L'objectif principal de ma recherche était de développer PhotoEukStein, un nouveau méta-modèle générique conçu spécifiquement pour la reconstruction entièrement automatique de modèles métaboliques d'algues eucaryotes. Ce méta-modèle représente une avancée significative dans le domaine en simplifiant le processus de reconstruction de modèles pour les eucaryotes.

5.2 Chapitre 2 : Reconstruction et validation de PhotoEukStein

Le chapitre 2 offre un aperçu complet des étapes impliquées dans la reconstruction de PhotoEukStein, en commençant par la curation manuelle et en poursuivant par la validation des prédictions du modèle. Le processus de reconstruction du méta-modèle implique des étapes de curation minutieuses, où divers aspects du modèle sont soigneusement affinés et optimisés. Cela implique la collecte et l'intégration d'informations bioinformatiques et biochimiques provenant de sources disponibles, et en s'assurant que le modèle représente avec précision les caractéristiques métaboliques des microalgues eucaryotes. Ce processus de reconstruction et de validation du modèle englobe non seulement des aspects techniques, mais soulève également des considérations philosophiques et épistémologiques.

5.2.1 Reconstruction de PhotoEukStein

Historiquement, la plupart des informations détaillées sur les processus photosynthétiques eucaryotes proviennent d'études sur les plantes supérieures et quelques algues modèles, notamment *Synechocystis*, *Chlamydomonas*, *Chlorella*, *Thalassiosira* et *Phaeodactylum*. Traditionnellement, la plupart des organismes modèles ont été choisis parce qu'ils sont facilement cultivables ou peuvent être manipulés génétiquement plutôt que leur pertinence écologique.

Les réseaux métaboliques des algues eucaryotes peuvent être trouvés dans des bases de données telles que BiGG et BioCyc, ou directement à partir de sources bibliographiques. Dans ces deux bases de données, j'ai spécifiquement ciblé les organismes photoautotrophes. Pour affiner davantage la sélection, j'ai exclu les modèles liés aux plantes terrestres, à l'exception d'*Arabidopsis thaliana*, qui fait l'objet d'études approfondies et est bien documenté. De plus, j'ai exclu les organismes parasitaires qui possèdent probablement des voies métaboliques uniques associées à leurs stratégies adaptatives. Après examen attentif de toutes les données disponibles, les informations biochimiques et génomiques de 15 algues eucaryotes et d'une plante terrestre ont été choisies comme matière première pour la construction de PhotoEukStein.

Dans le contexte des bases de données biologiques et de l'intégration des données, l'utilisation d'identifiants différents pour une même entité (comme les réactions, les métabolites, les gènes) peut créer de la confusion et rendre difficile la fusion des données provenant de ces différentes sources. Afin de garantir que différentes bases de données utilisent le même identifiant pour une entité donnée, il est possible de créer un espace de noms qui fait référence à un système d'identifiants uniques attribués aux entités afin de les standardiser. MetaNetX est une plateforme en ligne qui fournit des tables pour la correspondance des identifiants des métabolites et des réactions enzymatiques. Malgré les efforts de conciliation des bases de données métaboliques, une telle hétérogénéité nécessite encore une curation manuelle minutieuse. La gestion des identifiants au sein et entre les bases de données présente des défis significatifs, et ce processus de nettoyage est chronophage. Cependant, grâce à mon travail, une table plus complète pour la conversion des identifiants de BiGG et BioCyc vers BiGG est désormais disponible, complétant celle de MetaNetX pour ces deux bases de données (Chers collègues curateurs, vous trouverez cette précieuse ressource à ce lien : https://www.genoscope.cns.fr/PhotoEukStein/photoeukstein_manual_curation/). Un total de 2870 doublons a finalement été reconnu au sein de PhotoEukStein. De plus, 123 réactions ont été modifiées, 160 métabolites inconnus et 250 réactions (soit non trouvées dans la base de données, soit considérées comme des entités fictives) ont été supprimées au cours de ce processus.

L'objectif ultime de la curation est de préparer un modèle métabolique pour une analyse basée sur des contraintes. Les modèles basés sur des contraintes reposent sur la loi de conservation de la masse (Antoine Lavoisier, 1789), qui suppose que le système métabolique est dans un état quasi-stationnaire. Selon cette loi, la masse totale d'un système clos reste constante au fil du temps et ne peut être créée ni détruite. Ce principe est crucial pour les modèles métaboliques car il garantit l'équilibre de la stœchiométrie des réactions. Cet équilibre est essentiel pour la prédiction des flux métaboliques, sans avoir besoin de données cinétiques détaillées. Une réaction est équilibrée en masse si le nombre d'éléments est le même des deux côtés de la réaction. J'ai ajouté les formules chimiques manquantes (3406 formules manquantes sur 7467 métabolites) en utilisant une combinaison de méthodes, notamment MetaNetX, la curation manuelle et un algorithme de prédiction que j'ai developpé. Cette phase du processus de reconstruction s'est révélée particulièrement exigeante. L'aspect le plus chronophage a été la vérification manuelle et la recherche des formules manquantes, ainsi que l'identification des métabolites doublons (voir paragraphe précédent). Bien que j'aie réussi à prédire 672 formules, au moins autant de recherches internet ont dû être effectuées manuellement sur les différentes bases de données tout au long de cette étape. Il est important de noter que j'ai collecté 24945 identifiants de métabolites avec leurs formules chimiques respectives afin d'éviter ce travail fastidieux la prochaine fois (Chers collègues curateurs, vous trouverez cette ressource précieuse au lien ci-dessus).

5.2.2 Contenu de PhotoEukStein

PhotoEukStein englobe les informations biochimiques et génomiques disponibles sur 15 algues eucaryotes et une plante terrestre. Associé à CarveMe, ce nouveau modèle générique permet la reconstruction entièrement automatique de GSMs pour les microalgues eucaryotes (Chers collègues modélisateurs, vous trouverez cette ressource précieuse à ce lien : https://www.genoscope.cns.fr/PhotoEukStein/photoeukstein_for_carveme/).

PhotoEukStein contient 5831 métabolites et 11229 réactions. Deux types de réactions sont

distingués : 2067 réactions de bordure (comprenant 360 réactions d'échange, 674 réactions de puits et 1033 réactions de demande) et 9162 transformations biochimiques internes. Sur les 9162 réactions enzymatiques de PhotoEukStein, 7599 d'entre elles sont associées à 20468 séquences de protéines issues de génomes de référence. Quant aux 1563 autres réactions internes de PhotoEukStein qui n'ont pas de gènes associés, soit elles sont "spontanées" (se produisent sans influence ou intervention extérieure), soit aucun gène n'a été trouvé pour catalyser les réactions. Un troisième cas serait que les gènes sont connus, mais les modèles de base de PhotoEukStein ne les ont pas incorporés.

Dans PhotoEukStein, 15 fonctions objectives de biomasse ont été incorporées à partir de *Chlamydomonas reinhardtii*, *Chlorella variabilis* et *Phaeodactylum tricornutum*. Les réactions incluses consistent principalement en des réactions de biomasse autotrophes pendant les périodes lumineuses et sombres. Une limitation de la modélisation métabolique réside dans la dépendance à une fonction objective de biomasse qui est généralement paramétrée pour une espèce algale spécifique cultivée en laboratoire, ce qui la rend moins applicable à un large éventail d'algues dans leur environnement naturel. La supposition que les cellules microbiennes maximisent leur croissance peut être adaptée à des fins de bio-ingénierie, mais pas nécessairement à des applications écologiques, où les contraintes nutritionnelles sont courantes. De plus, dans une étude utilisant la modélisation multi-objectifs d'un petit écosystème, il a été démontré que lorsque chaque espèce croît à son taux maximal, les autres guildes échouent à produire de la biomasse. Pour maintenir la stabilité de l'écosystème, les espèces doivent croître à des taux sous-optimaux. Dans le cadre de recherches futures, on pourrait améliorer la réaction de biomasse en ne considérant que les exigences minimales de chaque algue, tout en minimisant la consommation d'énergie, par exemple. Les autres contenus moléculaires seraient ensuite produits indépendamment. Cependant, « ce que les organismes vivants cherchent à optimiser ? », en particulier pour les organismes eucaryotes dans leur environnement naturel, est une question débattue qui soulève des questions philosophiques profondes dont les conclusions possibles dépassent presque l'approche scientifique.

5.2.3 Capacités de PhotoEukStein

PhotoEukStein combine les caractéristiques métaboliques des eucaryotes photosynthétiques, c'est-à-dire que l'absorption des photons permet la production d'adénosine-triphosphate (ATP) par l'ATP-synthase chloroplastique, qui alimente à la fois la fixation du CO_2 par la RuBisCo et son intégration dans les composants organiques essentiels à la croissance. L'intégration de compartiments intracellulaires tels qu'un chloroplaste (12 métabolites et 44 réactions) et un thylakoïde (3 métabolites et 4 réactions) permet cette fine synchronisation entre ces réactions clés du métabolisme photoautotrophe. En effet, les membranes biologiques servent à de nombreuses fins. L'une d'entre elles est de contrôler les flux de soluté entre les compartiments à l'intérieur des cellules et entre les cellules, une autre est de faciliter l'organisation spatiale des réactions chimiques et ainsi favoriser l'émergence de nouveaux phénotypes.

Mon travail s'est principalement concentré sur la fonction carboxylase de la RuBisCo. Cependant, il est important de noter que cette enzyme possède également une autre activité enzymatique appelée oxygénation. La carboxylation et l'oxygénation se produisent toutes deux dans le même site actif de la RuBisCo, ce qui crée une relation de compétition entre les deux activités. L'activité dominante dépend des concentrations relatives des substrats à proximité immédiate de l'enzyme. Ce processus, qui implique la capture de l'oxygène et la libération du dioxyde de carbone, est appelé photorespiration.

La respiration cellulaire est un autre processus métabolique complexe comprenant trois étapes principales : la glycolyse, le cycle de Krebs et la phosphorylation oxydative. Lors de la phosphorylation oxydative, l'énergie est libérée à partir des électrons transportés par des molécules réduites, telles que le NADH. Ces électrons sont transportés à travers une séquence de protéines de transport d'électrons situées dans la membrane mitochondriale interne, ce qui génère un gradient transmembranaire de protons. Finalement, ce gradient de protons facilite la production d'ATP par l'ATP-synthase mitochondriale. Cependant, cette ATP-synthase nécessite la présence à la fois des compartiments mitochondriaux et des peroxysomes pour permettre le bon fonctionnement et la coordination des réactions impliquées dans la respiration cellulaire. Alors que les principales voies de la respiration, telles que la glycolyse et le cycle de Krebs, sont déjà incorporées dans PhotoEukStein, la validation du métabolisme respiratoire pendant la nuit permettrait de modéliser les algues en fonction de leur horloge circadienne.

Je suis convaincu qu'il est possible d'améliorer davantage le métabolisme primaire et secondaire de PhotoEukStein. Cependant, malgré ces limitations, les GSMs dérivés de PhotoEukStein pour diverses espèces d'algues, telles que *P. tricornutum*, *C. variabilis* et *T. pseudonana*, présentent une efficacité comparable aux modèles basés sur l'expertise pour capturer les connaissances biologiques essentielles. Nous avons échantillonné de manière extensive les niches métaboliques photoautotrophes pour les 6 GSM (c'est-à-dire environ 10000 conditions environnementales générées de manière aléatoire pour chacun d'eux) et avons comparé leurs taux de croissance prédits. Pour chaque paire de GSMs, les taux de croissance prédits sont fortement corrélés, ce qui montre que les GSM dérivés de PhotoEukStein sont aussi efficaces que les modèles basés sur l'expertise pour capturer les connaissances biologiques fondamentales, et correspondent aux observations faites avec les cultures. Pour examiner plus en détail la cohérence interne des GSM dérivés de PhotoEukStein, nous avons examiné les corrélations entre les flux des réactions par les deux modèles pour *Phaeodactylum tricornutum*. Les cartes de corrélation obtenues sont très similaires tant pour le modèle de référence que pour le GSM dérivé de PhotoEukStein, ce qui indique que les deux modèles relient de manière très similaire les différents flux. Lorsque les paires de réactions sont fortement corrélées dans un GSM, elles sont également fortement corrélées dans l'autre GSM, et les paires de réactions peu liées ont des caractéristiques similaires dans les deux modèles. L'approche automatisée de type « descendant » appliquée à *P. tricornutum* capture donc les mêmes éléments essentiels que le modèle basé sur l'expertise et représente les mêmes caractéristiques biologiques, même lorsqu'on considère la distribution des flux métaboliques dans le GSM.

Ainsi, les GSM dérivés de PhotoEukStein démontrent la capacité à reproduire des phénotypes physiologiques attendus et sont susceptibles de fournir des connaissances biologiques utiles en intégrant le contenu génétique des organismes dans les GSM, même si aucune autre connaissance que leur contenu génétique n'est disponible. Bien qu'il soit reconnu que PhotoEukStein peut comporter des imprécisions dans divers aspects, il reste le modèle générique le plus complet et raffiné actuellement disponible pour les eucaryotes photoautotrophes. Son développement et sa curation ont impliqué l'intégration de diverses données expérimentales et sources bibliographiques, ainsi qu'une curation et un affinement manuels approfondis, ce qui a donné un modèle qui capture un large éventail de processus métaboliques et d'interactions. Par conséquent, PhotoEukStein représente une étape importante vers la compréhension et la modélisation du métabolisme, de la physiologie, de la biogéochimie et de l'écologie des eucaryotes photoautotrophes, et constitue donc une ressource précieuse pour les chercheurs. De plus, PhotoEukStein peut facilement être étendu pour incorporer de

nouvelles connaissances métaboliques afin de suivre le développement des études sur les organismes unicellulaires eucaryotes phototrophes, que ce soit en identifiant de nouvelles réactions métaboliques ou en accumulant des séquences protéiques de référence associées à une réaction donnée.

5.3 Chapitre 3 : Une base de données des modèles métaboliques pour les microeucaryotes phototrophes marins

Le chapitre 3 est représenté par mon article de thèse et résume brièvement les concepts des chapitres 1 et 2. Plus important encore, il présente la toute nouvelle ressource de 549 GSMs pour les microeucaryotes phototrophes dérivés des génomes et transcriptomes environnementaux (Chers collègues modélisateurs, vous trouverez cette ressource précieuse à ce lien : https://www.genoscope.cns.fr/PhotoEukStein/photoeukstein_DB/). Ces GSMs offrent de nouvelles opportunités pour comprendre les réseaux métaboliques complexes et les implications écologiques de ces organismes eucaryotes dans différents contextes environnementaux.

Cet article met l'accent sur l'importance d'adopter une approche holistique lors de l'étude des systèmes biologiques. Actuellement, la caractérisation des fonctions planctoniques est souvent limitée à 1) l'annotation génique, 2) les corrélations statistiques, ou 3) les proxies taxonomiques. Cependant, ces approches ont leurs limites et ne fournissent pas une compréhension complète des interactions complexes et des fonctions émergentes au sein de ces systèmes. 1) La dépendance exclusive à l'annotation génique réduit la vision au déterminisme génétique et néglige le réseau complexe d'interactions qui contribuent aux résultats fonctionnels. 2) Les approches statistiques qui corrélent les gènes ou l'abondance des organismes avec les paramètres environnementaux fournissent des informations précieuses, mais n'établissent pas de liens causaux et ne répondent pas à la question de "qui fait quoi et comment". 3) La modélisation des traits fonctionnels à l'échelle océanique et la prise en compte des dynamiques temporelles sont très puissantes. Cependant, ces modèles simplifient souvent les processus biologiques en associant la fonction à des proxies taxonomiques, ignorant la variabilité intra-individuelle et la complexité des processus physiologiques. Par conséquent, l'objectif à long terme consiste à prédire les processus physiologiques, tels que la croissance des organismes planctoniques ou la production de molécules clés, en considérant l'ensemble des réactions biochimiques plutôt que de les simplifier à des équations physiques. De plus, l'objectif est de s'éloigner des associations systématiques entre taxons et fonction, en reconnaissant que des traits fonctionnels peuvent être présents chez des organismes divers, et d'intégrer la variabilité intra-individuelle, la plasticité phénotypique.

Les 549 nouveaux GSMs représentent une ressource précieuse qui met en évidence l'importance d'adopter une approche holistique pour l'étude des systèmes biologiques et souligne les limites des approches actuelles de caractérisation des fonctions planctoniques. Plutôt que de se fier à des corrélations simplistes ou à des observations isolées, l'adoption d'une perspective holistique nous permet de capturer la nature multifacette des systèmes biologiques. Cela nous permet d'explorer l'interaction complexe entre la diversité génétique, la dynamique environnementale et le fonctionnement des écosystèmes, conduisant finalement à une représentation plus nuancée et précise des complexités inhérentes à ces écosystèmes vitaux. L'utilisation de GSMs ouvre plusieurs dimensions à la définition de la "fonction".

En effet, lors de l'exploration de notre nouvelle base de données de GSMs, nous avons observé que

lorsque nous nous concentrons uniquement sur l'annotation fonctionnelle des gènes dans les génomes, un signal phylogénétique est apparent. Une observation intrigante est que lorsque nous excluons les annotations structurales et ne considérons que les annotations fonctionnelles codant pour les enzymes, une partie de ce signal est perdue. Cependant, lorsque nous examinons le contenu des réactions parmi les modèles, le signal phylogénétique reste relativement cohérent. Ces trois étapes indiquent collectivement que la diminution du signal phylogénétique n'est pas attribuable à la prédiction du contenu réactionnel par PhotoEukStein, mais plutôt à la spécificité taxonomique inhérente à la structure cellulaire.

Ce qui ajoute un intérêt supplémentaire, c'est l'examen de la manière dont les différentes composantes du système, en particulier les réactions, sont interconnectées. Le signal phylogénétique est totalement absent et aucun schéma spécifique n'émerge. Au lieu de cela, chaque réseau semble être unique, ce qui entraîne une répartition dispersée dans l'espace. Il est possible que l'absence de compartiments dans PhotoEukStein contribue à une perte d'information et que la structure des réseaux métaboliques puisse correspondre à un schéma phylogénétique. Bien qu'il soit important de ne pas exclure cette hypothèse, je suis fermement convaincu que l'interprétation peut aller au-delà de ces aspects. En se concentrant sur les aspects fonctionnels, cette dispersion peut également indiquer un niveau significatif d'adaptabilité lorsque l'on considère l'ensemble des réseaux collectivement.

Ce qui est vraiment puissant, c'est de prendre en compte les flux réactionnels dans les modèles. Les organismes ont été testés pour leur croissance et la production d’un métabolite particulier dans de nombreux environnements aléatoires. Cette fois, des groupes clairs émergent. Nous les décrivons comme des groupes fonctionnels, c'est-à-dire des groupes d'organismes qui suivent le même profil de variation en fonction du milieu, en termes de production de composés ou de taux de croissance. Ce sont des groupes d'organismes qui répondent de manière similaire aux conditions environnementales et qui ne sont pas du tout liés sur le plan phylogénétique : des organismes étroitement apparentés avec un répertoire similaire de réactions métaboliques peuvent présenter des profils fonctionnels différents, tandis que des organismes distants apparentés avec des ensembles différents de réactions métaboliques peuvent masquer des similitudes métaboliques. Le profilage des organismes en fonction de traits fonctionnels spécifiques conduit à des classifications distinctes qui ne peuvent être réduites uniquement à la taxonomie ou à la présence/absence d'un gène.

Nous soutenons l'idée de considérer PhotoEukStein et ses GSMs dérivés comme une ressource permettant de mettre en évidence des catégories améliorées de phénotypes omiques pouvant être considérées comme des traits potentiels dans les futurs modèles des systèmes océaniques.

5.4 Chapitre 4 : PhotoEukStein ouvre la voie à la modélisation métaboliques des microeucaryotes phototrophes

Le chapitre 4 se compose de deux parties distinctes, abordant chacune des aspects importants de la recherche. La première partie se plonge dans la complexité inhérente des systèmes biologiques, en abordant la relation sophistiquée entre le génotype et le phénotype. Mais elle met aussi en évidence les limitations de l'intégration de tous les paramètres dans un seul type de modèle. Cette reconnaissance de la complexité souligne la nécessité de recourir à des approches de modélisation alternatives qui capturent les imperfections et les incertitudes, ce qui peut à son tour conduire à la génération de nouvelles hypothèses et de nouvelles connaissances. En discutant de ces concepts

philosophiques et épistémologiques, le chapitre favorise une compréhension plus profonde des motivations et des justifications sous-jacentes des méthodologies choisies. Il met l'accent sur l'importance de la pensée critique et de l'interprétation dans la recherche scientifique, encourageant les chercheurs à reconnaître les limitations et les hypothèses intégrées aux modèles.

Dans la deuxième partie du chapitre, l'accent est mis sur le potentiel de la modélisation métabolique pour élucider les caractéristiques et les fonctions des organismes planctoniques. A) Des résultats préliminaires sont présentés pour démontrer la valeur de l'intégration de ces approches de modélisation avec des manipulations expérimentales, mettant en évidence les effets synergiques qui découlent de leur combinaison. B) De plus, la discussion s'étend aux orientations et aux idées de recherche futures potentielles, mettant en évidence les voies d'exploration et d'investigation supplémentaires dans les domaines écologiques.

5.4.1 Relations génotypes x environnements - phénotypes

Une habitude des biologistes est la caractérisation des phénotypes, que ce soit pour la santé humaine (comme les endophénotypes), la productivité des cultures ou la surveillance des écosystèmes (biomarqueurs), par exemple. La relation était initialement considérée comme simple. Pour chaque « caractère phénotypique » héréditaire, il était postulé qu'il existe un élément génétique discret (un gène) le transmettant à travers les générations. Cette approche réductrice de la biologie moléculaire et de la génomique est souvent comprise comme une chaîne causale mécaniste perpétuant des raccourcis tels que « le(s) gène(s) X pour le trait Y ». Ainsi, peu importe le point de vue adopté, le génotype et le phénotype sont effectivement équivalents. En fait, le phénotype est souvent un indicateur imparfait du génotype : le même génotype peut donner lieu à une large gamme de phénotypes, et le même phénotype peut provenir de différents génotypes. Il est crucial de reconnaître que les fonctions biologiques de haut niveau impliquent souvent l'activité coordonnée de nombreux gènes, jusqu'à des centaines ou plus (phénomène appelé polygénie). De même, des gènes individuels peuvent participer à de multiples fonctions (pléiotropie). Cette complexité rend difficile l'attribution d'étiquettes fonctionnelles univoques aux gènes en se basant uniquement sur les protéines qu'ils codent. La fonction biologique émerge des interactions complexes entre les protéines et les autres composants cellulaires.

Une vision globale et une perspective à l'échelle des systèmes sont essentielles pour démêler de manière exhaustive les complexités de la fonction des gènes et leur contribution aux caractéristiques phénotypiques. Cela nécessite de prendre en compte la logique et les principes qui opèrent à différents niveaux, sans se concentrer uniquement sur les niveaux inférieurs. De plus, une grande partie de la logique des systèmes vivants se trouve à des niveaux supérieurs, car c'est souvent à ces niveaux que la sélection a lieu, déterminant ainsi si les organismes vivent ou meurent (en fonction de leur adaptation à leur environnement). Chaque niveau a sa propre intégration de fonctions, et il incombe aux biologistes de déterminer à quel niveau une fonction spécifique est intégrée.

Tous les niveaux d'organisation biologique sont influencés par l'environnement externe. Il a été rapporté divers cas où une différence génétique n'est pas visible au niveau phénotypique en raison des influences environnementales. Par exemple, la différence génétique rouge-blanc dans la couleur des fleurs de primevère n'est plus visible lorsque les plantes sont cultivées à 30°C-35°C car à haute température, toutes les fleurs sont blanches. Comme autre exemple populaire, les études

développementales de Waddington ont montré que l'embryon de drosophile pouvait présenter différentes structures du thorax et des ailes simplement en modifiant la température environnementale ou un stimulus chimique. Pour revenir au plancton, les diazotrophes ont la capacité génétique de fixer le gaz diatomique N_2 en tant que source d'azote grâce à l'enzyme nitrogenase. Il peut donc être tentant de supposer automatiquement qu'ils fixent toujours le N_2, cependant l'expression de la nitrogenase se produit uniquement lorsque les diazotrophes ne peuvent pas obtenir suffisamment d'azote à partir d'autres sources inorganiques telles que NH_4. Cette caractéristique fonctionnelle dépend des conditions environnementales et est un événement d'acclimatation. Nous discuterons également de la production de DMSP sous stress azoté chez *Phaeodactylum tricornutum* (Chapitre 4). En plus de ces facteurs abiotiques, les interdépendances métaboliques avec d'autres organismes (comme les relations de coopération) permettent également l'émergence de phénotypes particuliers.

L'environnement immédiat du système, tel que les informations structurelles et l'histoire évolutive, est également un composant essentiel pour comprendre la complexité des systèmes biologiques et leur comportement. Ces contraintes sont encodées à la fois dans les séquences d'ADN et dans l'architecture cellulaire héritée. La vision « centrée sur les gènes » dans sa version forte suggère que la structure complète d'un organisme est en quelque sorte encodée dans l'information génétique. Cependant, cette vision est considérée comme implausible et non étayée par les connaissances actuelles. En effet, l'ADN n'est pas le seul porteur de l'hérédité. Alors que les séquences d'ADN déterminent les séquences d'acides aminés dans les protéines, l'architecture cellulaire influence leur localisation, leurs mouvements et leurs interactions. Les machines cellulaires, y compris les mitochondries, le réticulum endoplasmique, les microtubules, les membranes et les arrangements chimiques spécifiques au sein des compartiments, déterminent également le comportement des protéines. Ces composants hérités ne sont pas principalement dictés par les séquences d'ADN. Les gènes n'ont pas besoin de coder tous les aspects de la fonction cellulaire. Les cellules eucaryotes, en particulier, sont hautement structurées, avec des organites membranaires et d'autres compartiments qui contribuent à leur complexité. Les propriétés biophysiques et les processus d'auto-organisation des molécules et des structures jouent un rôle important dans le développement phénotypique.

5.4.2 Directions de recherche et idées potentielles

5.4.2.1 Coopération synergique entre les approches expérimentales et de modélisation métabolique

Le composé tertiaire de sulfonium, le diméthylsulfonio-propionate (DMSP), a suscité un intérêt particulier en tant que précurseur biogénique du principal gaz sulphuré, le diméthylsulphure (DMS). Lorsque le DMS est libéré dans l'atmosphère, il peut agir comme un noyau de condensation des nuages, ce qui signifie qu'il peut attirer la vapeur d'eau pour former de petites gouttelettes qui finissent par former des nuages. La couverture nuageuse est importante pour réguler le climat car elle affecte la quantité de rayonnement solaire absorbée par la surface terrestre et l'atmosphère. La production de DMSP a été observée chez divers organismes planctoniques, notamment les algues, les bactéries, les dinoflagellés hétérotrophes, mais aussi les plantes et les animaux tels que les coraux, entre autres. Ainsi, le DMSP est présent dans tous les écosystèmes marins et peut être utilisé à diverses fins, non seulement par les organismes producteurs, mais également par d'autres espèces

qui vivent dans le même habitat que les producteurs de DMSP.

De nombreux facteurs peuvent affecter la biosynthèse du DMSP, tels que la lumière, la salinité ou la température, en fonction de ses fonctions physiologiques. En plus de ceux-ci, d'autres facteurs semblent également affecter les quotas cellulaires de DMSP, mais les mécanismes de régulation exacts ne sont pas encore clairs. Une hypothèse est présentée selon laquelle la production de DMSP est décrite comme un mécanisme de débordement pour les composés réduits du carbone et du soufre en excès. Chez les plantes supérieures, il existe un couplage régulatoire réciproque entre les voies de réduction des sulphates et des nitrates assimilatoires afin de maintenir des proportions appropriées d'acides aminés pour la synthèse des protéines. Cependant, dans la littérature, il a été observé que la limitation en azote peut entraîner une production accrue de DMSP chez de nombreuses algues et plantes productrices de DMSP, ce qui entraîne une incorporation plus élevée de soufre par rapport à l'incorporation d'azote. Fait intéressant, le DMSP ne contient pas d'azote. Le mécanisme de débordement peut être considéré comme une réponse de la cellule dans des conditions de croissance déséquilibrée, produisant et éliminant des composés pour assurer la poursuite d'autres voies métaboliques. Ce mécanisme permet une assimilation continue des sulphates même en présence de limitations en azote. Ainsi, l'augmentation de l'excrétion dans le milieu peut servir de moyen de dissipation de l'excès de soufre et de carbone.

Par conséquent, nous avons comparé la capacité de *Phaeodacylum tricornutum* à produire du DMSP sous stress azoté à la fois *in silico* avec le modèle dérivé de PhotoEukStein, ainsi que *in vivo* avec la culture d'algues réalisée au Genoscope. Les résultats ci-dessous sont préliminaires et des explorations supplémentaires sont nécessaires avant de tirer des conclusions.

Nous avons montré que l'absence de la voie métabolique du DMSP dans le modèle de *Phaeodactylum* pénalise grandement sa croissance dans un environnement riche en soufre. Nous avons ensuite montré que cette augmentation du soufre dans l'environnement est relative à d'autres éléments présents dans l'environnement et peut également être perçue comme un stress azoté. En effet, pour maintenir une croissance maximale même en cas de stress azoté, le modèle doit augmenter sa sécrétion de DMSP dans l'environnement. Pour évaluer cette hypothèse, nous avons réalisé une expérience *in vivo*. En utilisant une culture standard de *Phaeodactylum*, nous avons divisé les cellules en deux groupes : l'un placé dans un milieu frais dépourvu de nitrate, et l'autre dans un milieu frais contenant du nitrate. Nous avons mesuré la concentration intracellulaire de DMSP dans les cultures sans nitrate et dans les cultures avec nitrate. L'augmentation de la production de DMSP par *Phaeodactylum* pendant le stress azoté est clairement observée. Cependant, des expériences plus approfondies sont nécessaires pour déterminer plus précisément les mécanismes.

La principale distinction entre un biologiste qui utilise la modélisation mathématique et un autre qui ne le fait pas est que le premier explore quantitativement les implications de ses idées, y compris en menant des expériences computationnelles pour évaluer leur plausibilité. Les avantages potentiels d'une telle approche sont évidents, car des prédictions quantitativement plausibles améliorent la recherche expérimentale subséquente axée sur les hypothèses. À l'inverse, il est également possible d'intégrer de nouvelles données issues d'expériences pour affiner nos prédictions. Les modèles que nous reconstruisons généralement sont basés sur le potentiel global de l'organisme, en supposant que toutes les protéines encodées dans le génome peuvent être utilisées par le modèle. Cependant, cette approche peut conduire à une surestimation des capacités métaboliques de l'organisme. En réalité, toutes les protéines ne sont pas exprimées simultanément dans différentes conditions. En tenant compte des transcriptomes, nous pouvons observer un sous-ensemble du réseau métabolique,

capturé au moment de l'échantillonnage et dans des conditions environnementales spécifiques, ce qui peut fournir une image plus précise des stratégies utilisées par l'organisme dans des conditions particulières. Cette approche est extrêmement innovante et j'ai hâte de voir les travaux futurs qui pousseront cette approche encore plus loin, de la culture en laboratoire à l'échelle de l'océan.

5.4.2.2 L'échelle mésoscopique

La compétition pour les ressources métaboliques peut affecter la composition des communautés en excluant certaines espèces compétitrices ou en favorisant la différenciation des niches. Les interactions coopératives et syntrophiques, telles que l'échange métabolique bénéfique, sont également susceptibles de jouer un rôle important, car elles peuvent modifier significativement la qualité nutritionnelle de l'habitat. Un aspect fascinant de ces interactions réside dans l'échange mutuel de nutriments, tels que les vitamines, entre différents organismes. Les vitamines sont des composés organiques essentiels nécessaires à divers processus biologiques. De nombreuses enzymes qui ont un coenzyme B12 sont connues chez les eucaryotes, notamment la méthionine synthase dépendante de la B12, mais la plupart de ces organismes ne peuvent pas la synthétiser *de novo* (auxotrophie). Cela signifie qu'ils dépendent de sources externes pour leur approvisionnement. Ainsi, les procaryotes forment souvent des partenariats avec les microalgues, leur fournissant les vitamines nécessaires. En retour, les microalgues offrent aux procaryotes un environnement stable et des nutriments qu'ils peuvent synthétiser. Cet échange coopératif de ressources illustre le pouvoir de la symbiose dans le maintien de l'équilibre écologique.

Disposer de plus de 549 modèles d'algues à notre disposition nous permet d'aller beaucoup plus loin dans l'étude de ces interactions, qu'elles soient virales, parasitaires ou synergiques, et ouvre également la voie aux concepts de l'holobionte ou de la phycosphère. En approfondissant l'étude des interactions entre les microorganismes, nous acquérons une meilleure compréhension de l'interdépendance des différentes espèces, de leur rôle dans le cycle des nutriments, de la façon dont cela façonne la composition des communautés en excluant certaines espèces compétitrices ou en favorisant la différenciation des niches, de la façon dont cela modifie la qualité nutritionnelle de l'habitat, etc.

Il est possible d'utiliser des techniques de co-occurrence pour capturer des modules d'espèces susceptibles d'interagir et mettre en évidence leurs interdépendances métaboliques à l'aide d'outils adaptés aux GSMs. Cependant, passer à des comportements, au niveau de la population, qui émergent de ces interactions individuelles nécessite un temps de calcul important et est donc limité aux petites communautés (jusqu'à 4 espèces à ma connaissance). Afin de passer à une modélisation à l'échelle de l'océan, l'utilisation d'approches basées sur les traits et dite « soupe » peut permettre de surmonter la demande de calcul des modèles individuels.

5.4.2.3 L'échelle océanique

Les modèles du système terrestre (ESM) sont devenus de plus en plus sophistiqués, permettant des simulations détaillées qui fournissent des informations précieuses sur l'océan, ses ressources et son avenir. Ces modèles, tels que NEMO-PISCES, utilisent des équations différentielles pour représenter la croissance d'organismes emblématiques en reliant la disponibilité des nutriments au taux de croissance à l'échelle de l'océan. Ils intègrent des processus physiques pour prédire la disponibilité

des nutriments à l'échelle mondiale et dans le temps. Cependant, ces équations nécessitent de nombreuses valeurs de paramètres qui sont souvent difficiles à obtenir expérimentalement. De plus, bien que les ESM soient efficaces du point de vue computationnel, ils ne prennent pas pleinement en compte les données omiques récentes, telles que les gènes et les fonctions associées, limitant leur capacité à capturer toute la variabilité intra-individuelle et les processus moléculaires. De plus, ces modèles simplifient excessivement l'association des caractéristiques fonctionnelles à la phylogénie, ce qui est connu comme une approche réductionniste. Une avancée significative dans ce domaine est l'intégration des modèles à l'échelle du génome (GSM) dans les ESM, comme proposé dans le prochain article "Modelling genome-scale knowledge in the global ocean" par Regimbeau *et al*. Cette intégration permet de relever le défi de l'estimation des taux de croissance tout en considérant de manière holistique le métabolisme de l'organisme. Ils exploitent également les conditions environnementales de l'ESM pour explorer l'espace des niches, révélant les propriétés physiologiques des organismes modélisés. C'est la première fois que les connaissances omiques sont appliquées aux ESM, ouvrant la voie à des études basées sur les omiques et à la théorie de l'évolution.

Pour faciliter l'intégration des GSMs avec les ESMs, il est crucial d'améliorer le lien entre ces deux types de modèles. Bien que les ESMs puissent avoir un ensemble de métabolites plus restreint par rapport aux GSMs, il existe néanmoins certains métabolites absents des GSMs. Par exemple, dans le cas de PhotoEukStein, le fer et le silicate ne sont pas inclus dans le modèle. Actuellement, les connexions entre les GSMs et les ESMs impliquent principalement la prise en compte de trois facteurs clés : l'azote, le phosphore et la lumière. Ces trois points de connexion ne sont pas du tout suffisants pour prédire correctement la croissance de certains organismes. Par exemple, dans la plupart des écosystèmes océaniques ouverts, il existe généralement une corrélation positive entre les concentrations en macronutriments et la biomasse du phytoplancton, en particulier lorsque l'ensoleillement est suffisant. Cependant, cette compréhension traditionnelle ne s'applique pas dans certaines régions de l'océan mondial, notamment le Pacifique subarctique, le Pacifique équatorial oriental et central, et l'océan Austral. Ces régions, appelées zones à haute teneur en nutriments et faible chlorophylle, présentent des concentrations élevées de nitrate et de phosphate tout au long de l'année, mais des niveaux relativement faibles de phytoplancton. En effet, la croissance des grandes cellules de phytoplancton, en particulier les diatomées, est limitée non seulement par le phosphate, mais aussi par la disponibilité de fer ou de silicate, ce qui explique l'activité autotrophe limitée dans ces régions. Élargir la gamme de métabolites considérés dans les GSMs et les aligner sur les composants pertinents des ESMs sera une étape importante pour parvenir à une représentation plus complète et précise de la dynamique des écosystèmes. En plus d'intégrer les métabolites de l'ESM dans PhotoEukStein, nous pouvons également proposer inversement de nouveaux métabolites clés à intégrer dans l'ESM. Cela nécessitera une revue approfondie de toutes les réactions de puits (SK), de demande (DM) et d'échange (EX) de PhotoEukStein.

Au-delà des concepts traditionnels des traits fonctionnels planctoniques, les GSMs peuvent servir d'outils précieux pour définir des traits fonctionnels spécifiques à certaines conditions environnementales, indépendamment des considérations taxonomiques ou phylogénétiques, comme proposé dans notre article (Chapitre 3). Cette approche nous permet d'explorer les caractéristiques fonctionnelles des organismes de manière plus nuancée et dépendante du contexte.

Aujourd'hui, l'intégration des GSMs avec les ESMs est réalisée pour un seul organisme à la fois. Cependant, les microorganismes existent rarement isolément et dépendent souvent d'interactions synergiques avec d'autres organismes. Les associations complexes au sein de ces communautés

contribuent à leur stabilité dans des environnements divers et variables. À cet égard, la prochaine frontière en modélisation métabolique réside dans l'utilisation de réseaux métaboliques avec une focalisation sur la modélisation des systèmes multi-organismes. Cependant, la complexité des réseaux métaboliques en tant que structures de données présente des défis. Il est essentiel d'adapter les réseaux métaboliques à la modélisation des écosystèmes. Une stratégie intéressante consisterait à changer l'échelle biologique et à ne pas considérer un seul compartiment (un GSM) par organisme, mais plutôt à construire un modèle qui contient la diversité fonctionnelle de plusieurs organismes partageant le même trait. L'idée serait de reconstruire un méta-modèle contenant l'ensemble des réactions des modèles regroupés dans un même groupe fonctionnel. L'objectif de cette approche est une fois de plus de dépasser la classification taxonomique.

En résumé, il est crucial de souligner l'importance de l'intégration des organismes divers et des traits communautaires, de relever les défis liés à l'intégration des données omiques et de comprendre la variabilité et la structure biogéographique des communautés planctoniques dans la modélisation des écosystèmes. Malgré les obstacles, l'incorporation des données omiques dans les modèles d'écosystèmes a le potentiel d'améliorer notre compréhension des écosystèmes planctoniques et de leurs réponses aux changements environnementaux à l'échelle océanique. Pour approfondir notre compréhension de la diversité planctonique et de ses contributions au fonctionnement du système terrestre, des efforts collaboratifs entre plusieurs domaines de recherche et le développement d'approches et de technologies innovantes sont essentiels.

5.5 CONCLUSION

En conclusion, cette thèse présente PhotoEukStein, qui est le premier modèle générique pour les eucaryotes. Il peut être facilement étendu à mesure que de nouvelles connaissances émergent. Son utilisation avec une approche descendante permet la reconstruction entièrement automatique de modèles métaboliques basés sur des contraintes pour les microeucaryotes phototrophes.

Nous mettons à disposition une toute nouvelle ressource de 549 GSMs pour les microeucaryotes phototrophes dérivés de génomes environnementaux et de transcriptomes. Nous avons montré qu'il est possible d'utiliser ces modèles pour mener des expériences *in vivo*, et aussi pour définir des traits fonctionnels qui incluent à la fois l'environnement et le génotype.

Les possibilités d'utilisation de PhotoEukStein sont en fait beaucoup plus vastes, et j'ai hâte de voir toutes les nouvelles recherches que PhotoEukStein rendra possibles.