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

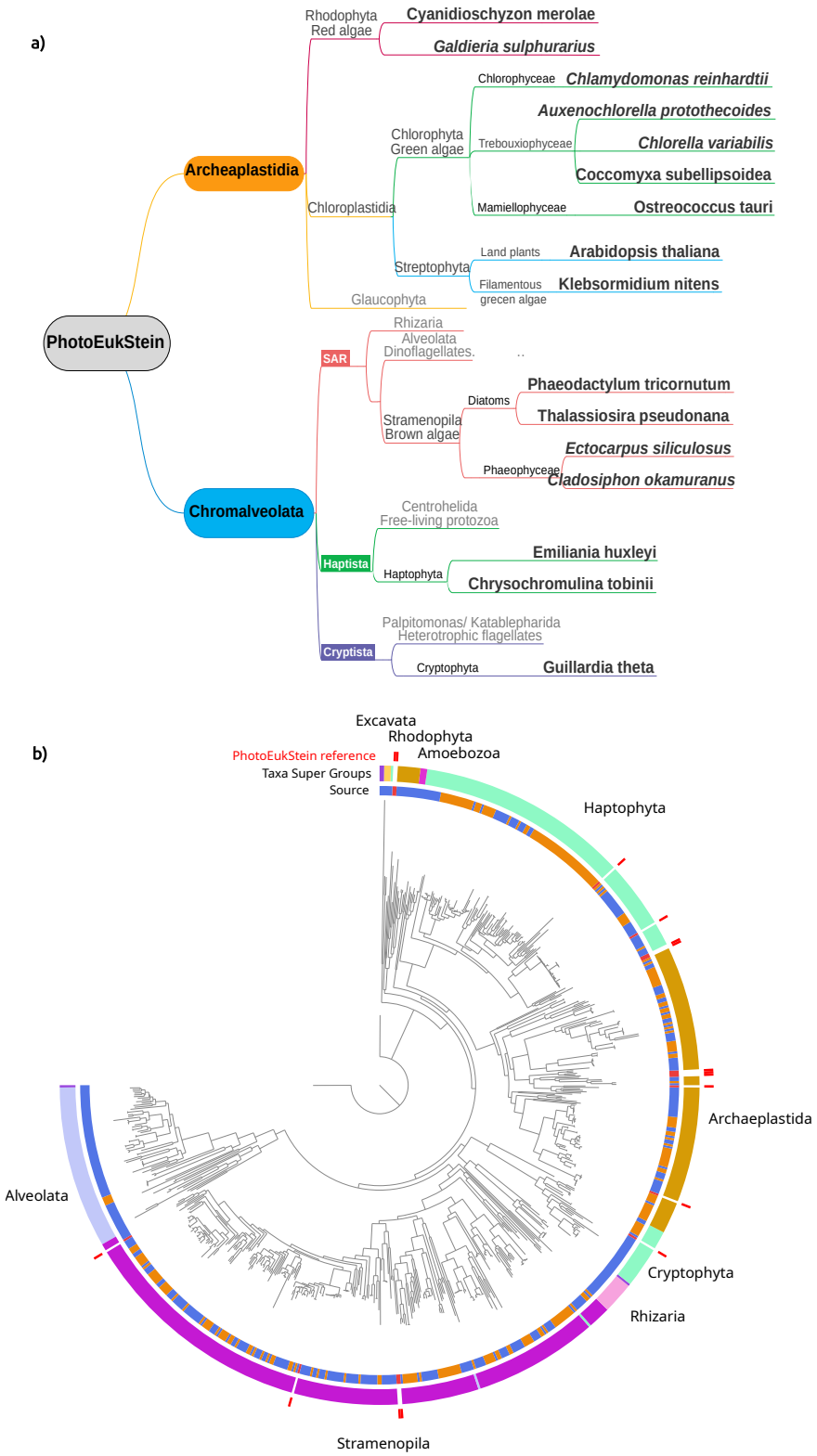


Figure 1: a) PhotoEukStein construction. Taxonomic diversity of the 16 existing GSMs that were combined to generate PhotoEukStein. b) Taxonomic diversity of the 553 PhotoeukStein-derived GSMs applied on 259 MAGs from Tara Oceans data (Delmont 2022, in orange in the inner circle) and 274 transcriptomes 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.

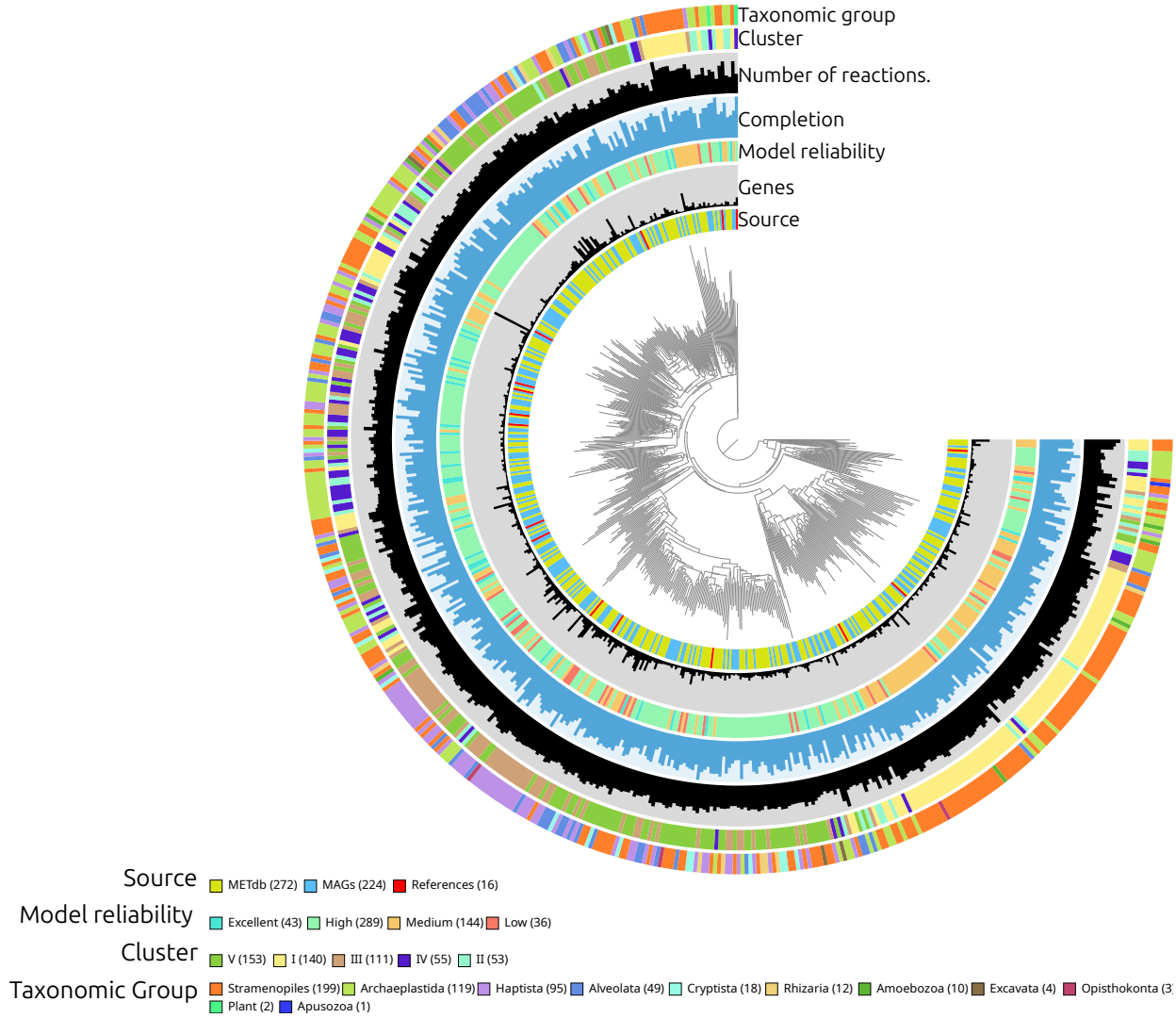


Figure 2: Main characteristics of PhotoEukStein derived GSMs. Central dendrogram represent Jaccard distance between GSMs reactions composition. Inner circle indicate the source of the sequence supporting the GSMs. Model reliability is defined by. Completion is the Busco-based evaluation of genome completion (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).

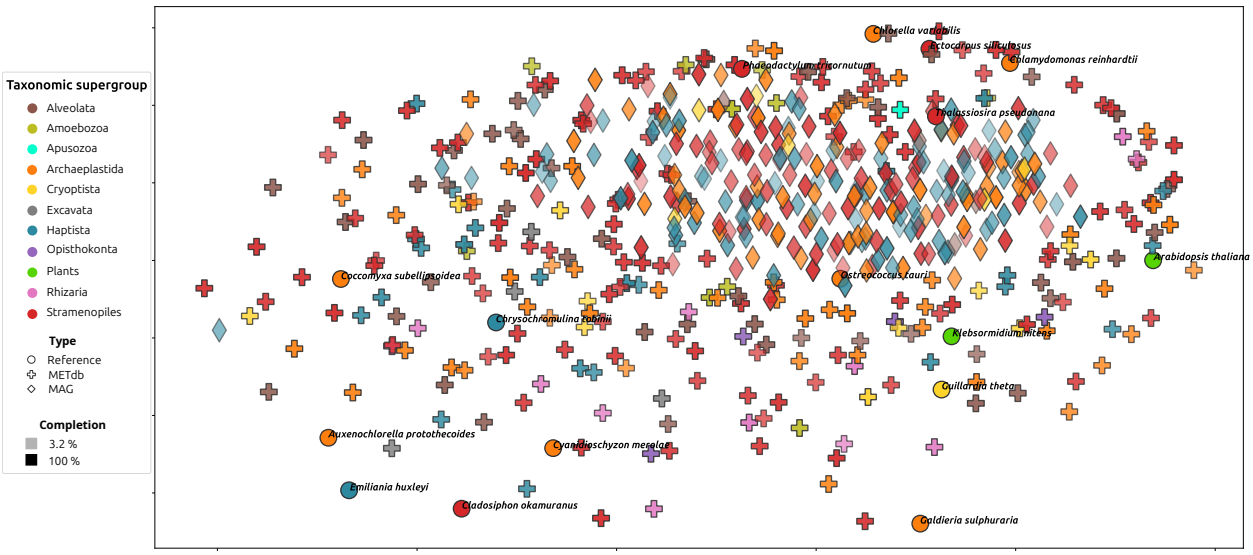


Figure 3: UMAP representation of diffusion map analysis of the 553 GSMs topology. MAGs are symbolised by diamonds, METdb by crosses, and reference genomes by circles. Colours indicate the taxonomic groups of each supporting genome, while transparency represents Busco-based genome completion estimation.

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].

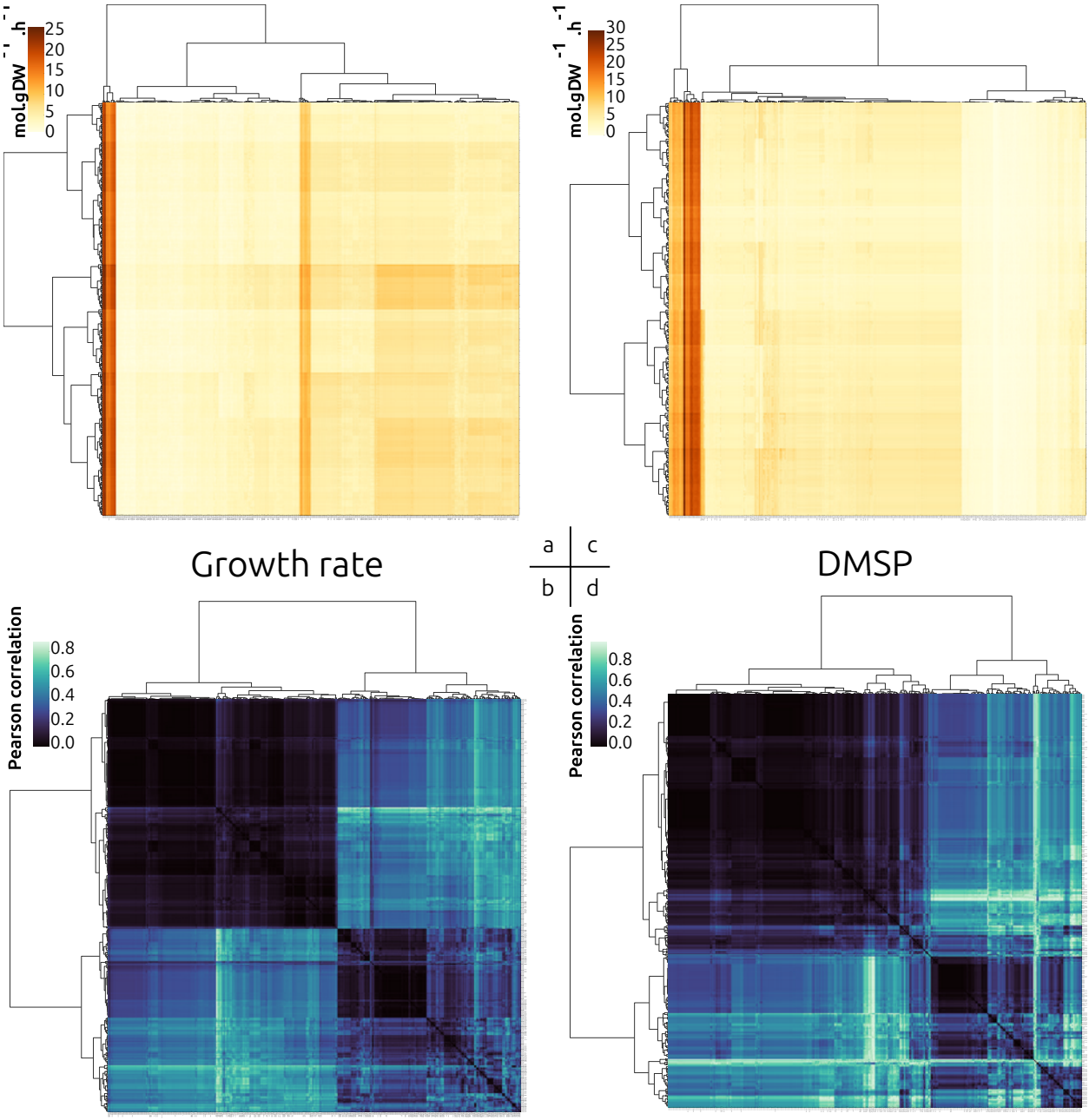


Figure 4: Metabolic niche exploration. The 553 GSM models are exposed to medium modification by systematic permutations of 1 up to 9 extra metabolites (listed in Supplementary Table S4 and S5), and growth rate is computed a). b) represents the clustered correlation matrix of growth rates modification 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.

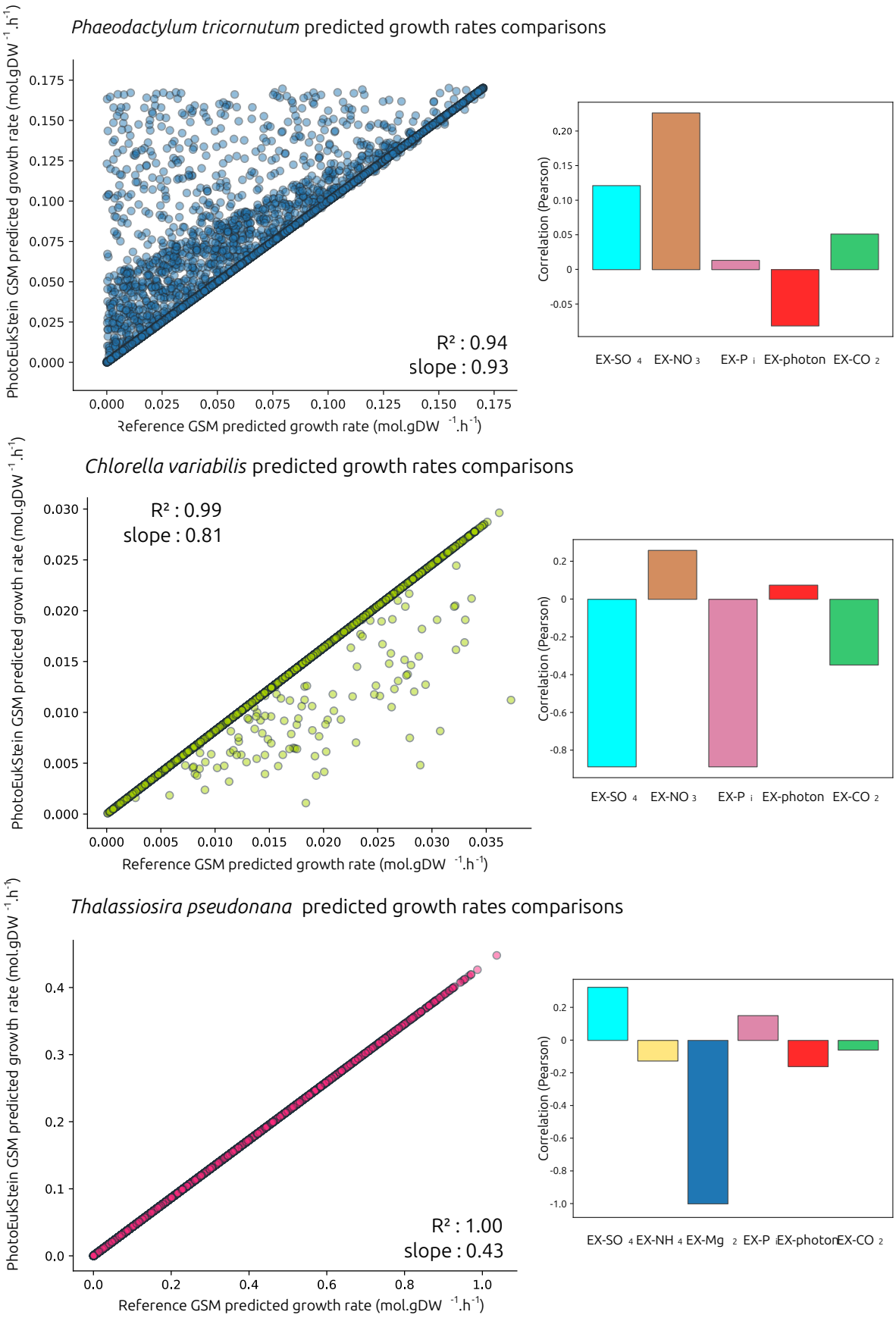
Supplementary Figure S2: Correlation matrix comparing original (iLB1034) and PhotoEukStein-derived GSMs reaction fluxes of *Phaeodactylum tricornutum*. The 434 shared reaction showing at least 1% or their correlations with other reactions fluxes greater to 0.2 when randomly sampling the whole metabolic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

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

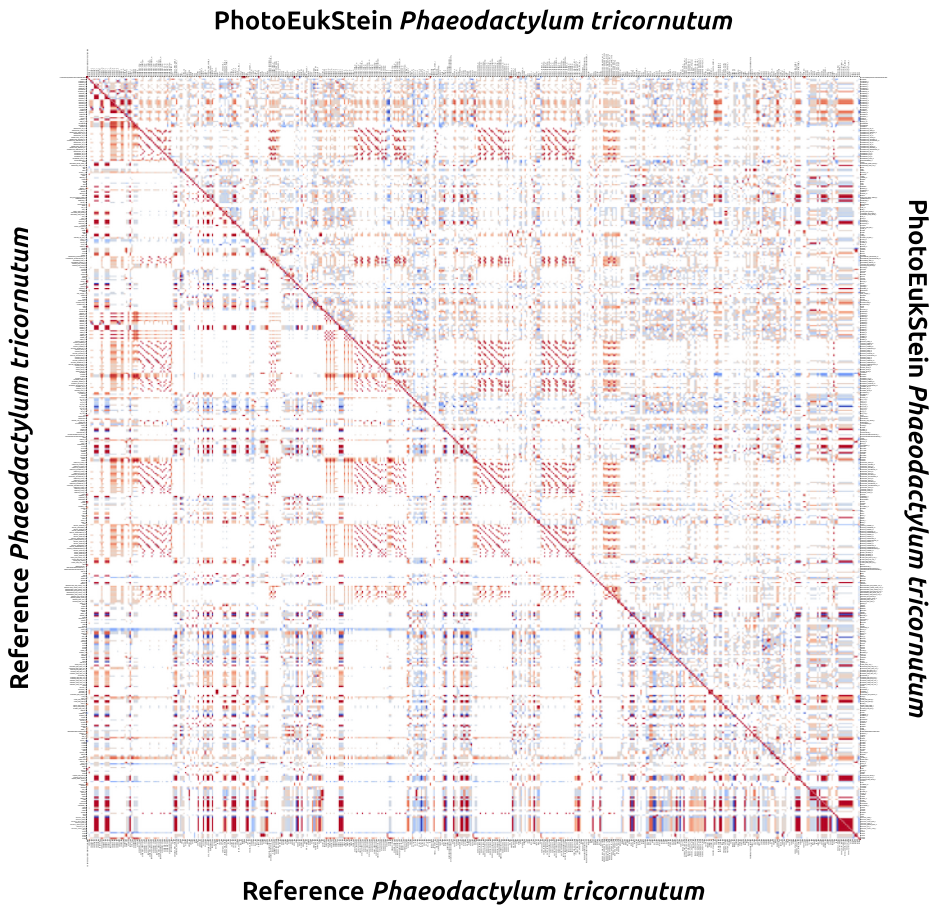
Supplementary Figure S3: HexBin representation of distribution of compared reaction fluxes correlations between PhotoEukStein and iLB1034 reference GSMs for *Phaeodactylum tricornutum*. Effective of correlations pairs of the 434 common reactions showing a possible correlation displayed in Figure S2 (see 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

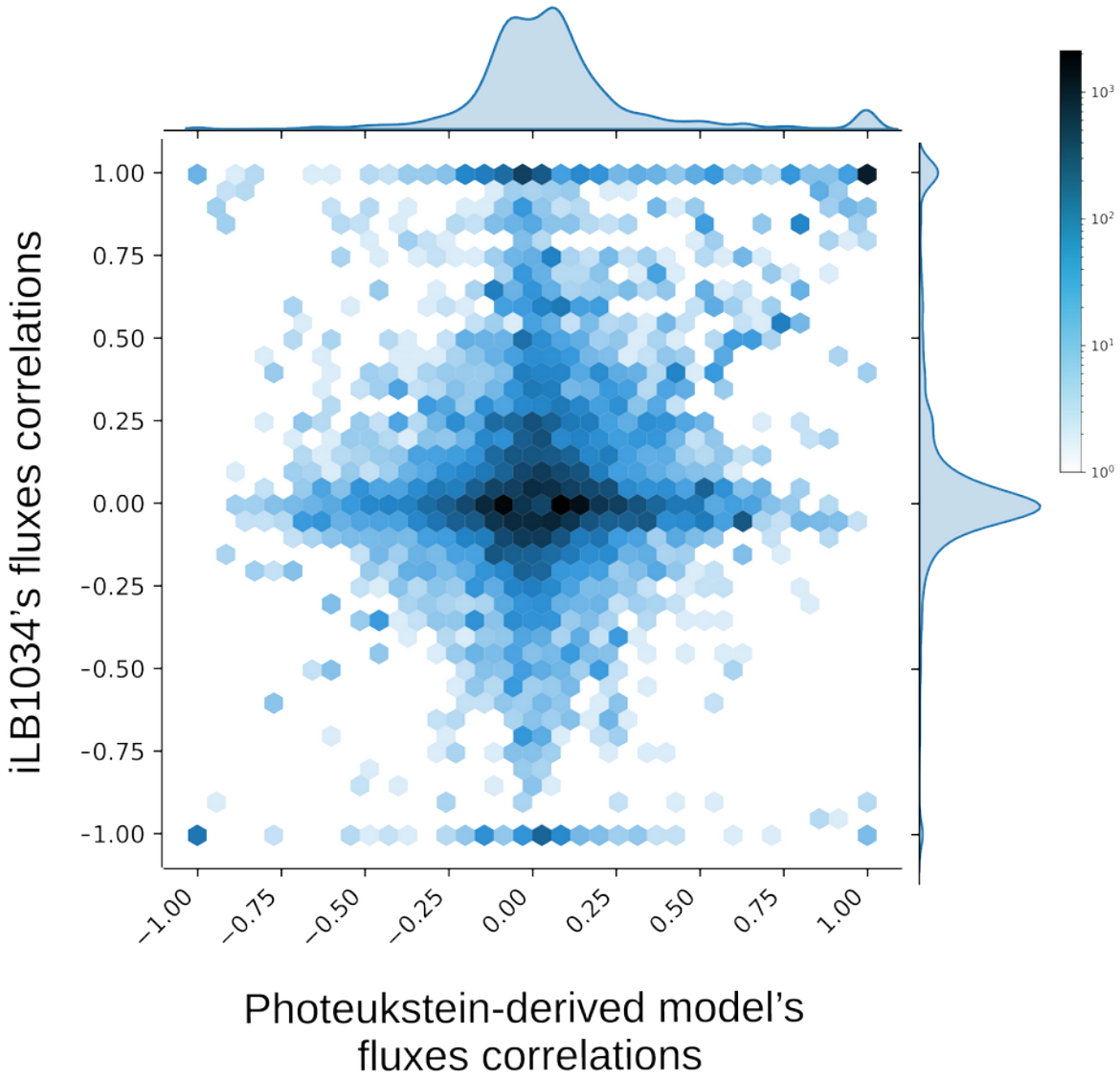
Supplementary Figure S5: Compositional analysis of the 549 PhotoEukStein derived GSMs. a) Diffusion map multiscale geometric analysis of GSMs topology followed by umap reduction of dimension processing have been performed to study the distribution of GSMs topologies in the functional space. Shapes indicate 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.



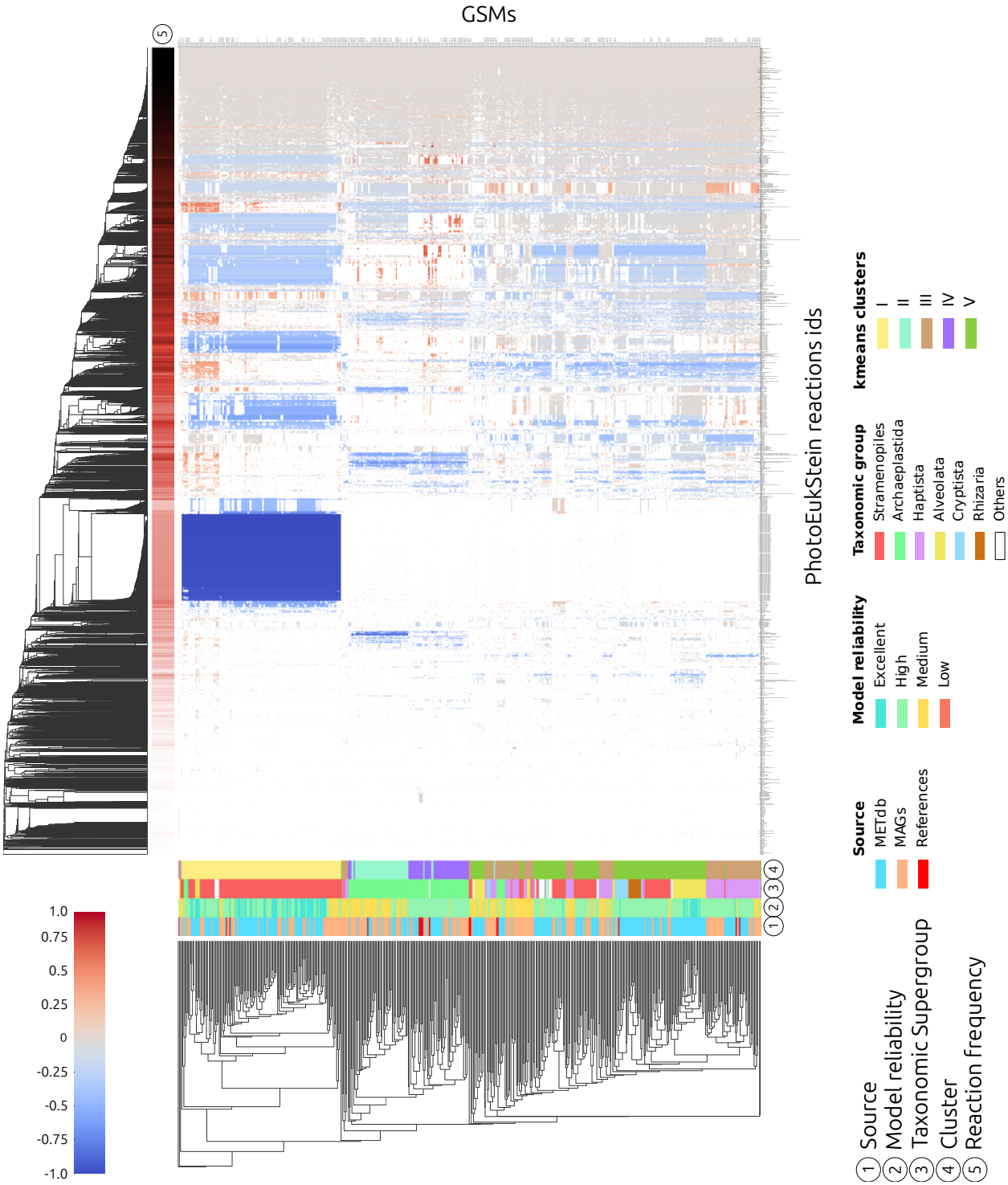
Supplementary Figure S1: Comparison of predicted growth rates from PhotoEukStein or reference GSMs for Phaeodactylum tricornutum (top left), Chlorella variabilis (middle left) and Thalassiosira pseudonana (bottom left). In each case, 10,000 iterations of random sampling within each GSMs' niche space 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.



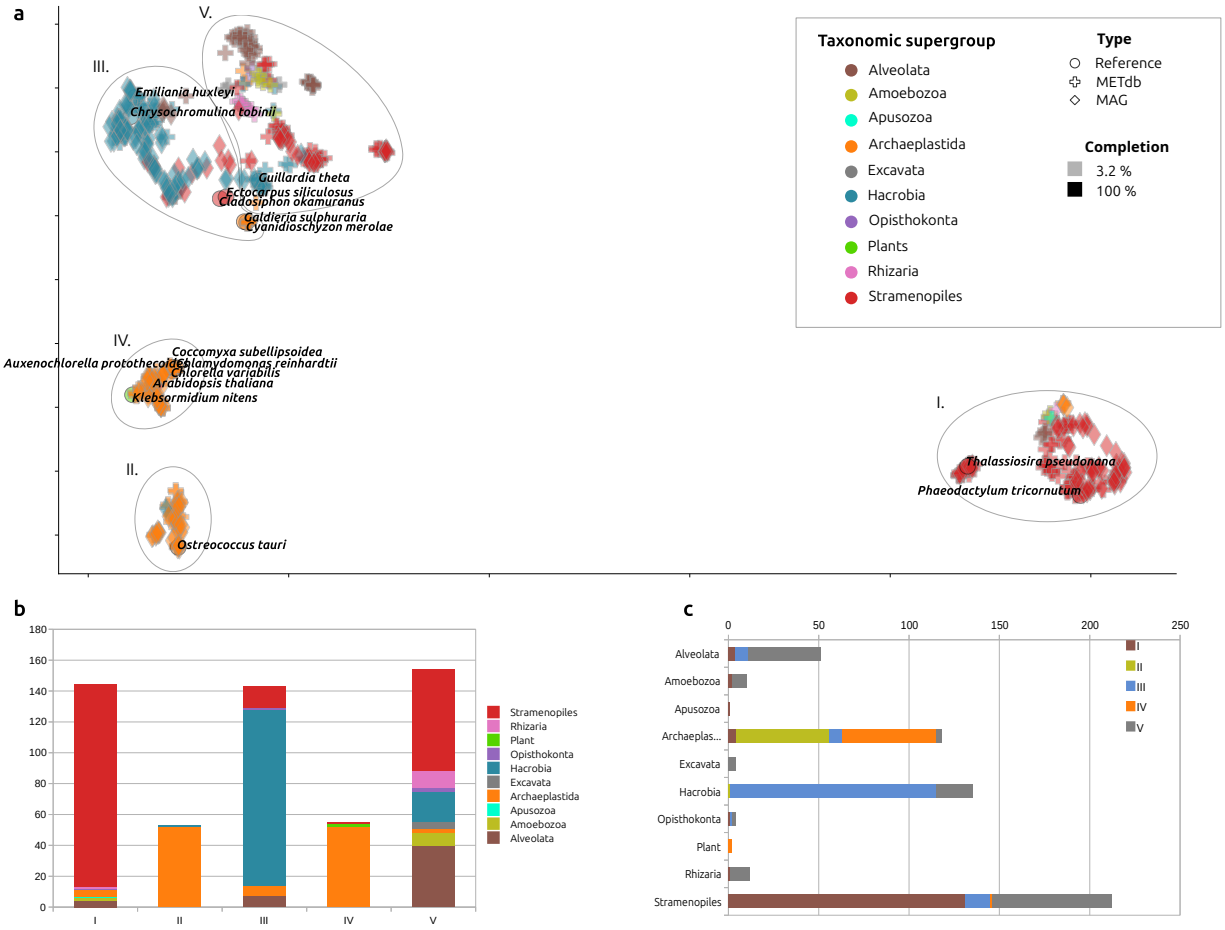
Supplementary Figure S2: Correlation matrix comparing original (iLB1034) and PhotoEukStein-derived GSMs reaction fluxes of Phaeodactylum tricornutum. The 434 shared reaction showing at least 1% or their correlations with other reactions fluxes greater to 0.2 when randomly sampling the whole metabolic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

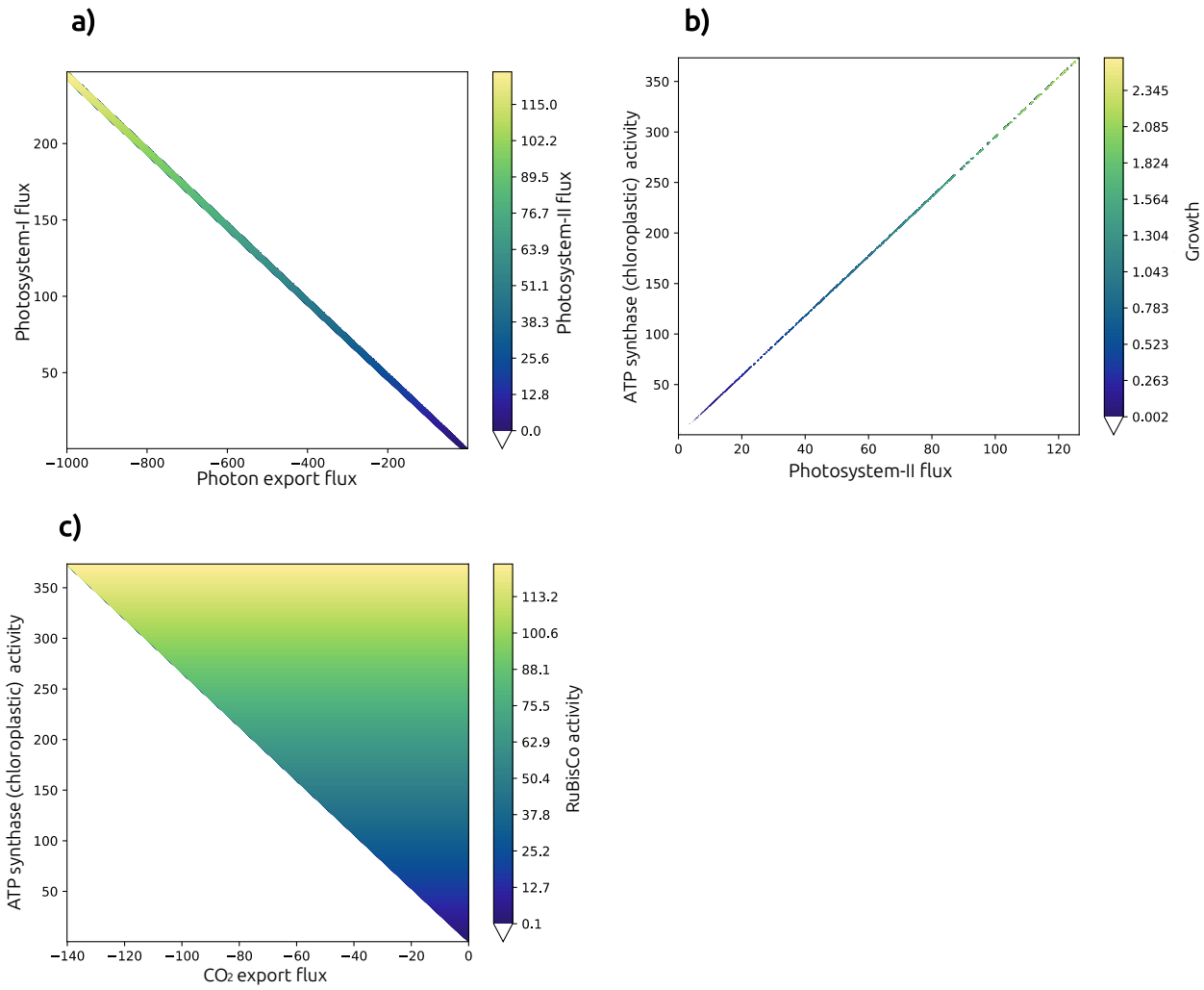
18. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
19. Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nat Commun* **11**, 4887 (2020).
20. Eren,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

Extended Data

1 - PhotoEukStein validation

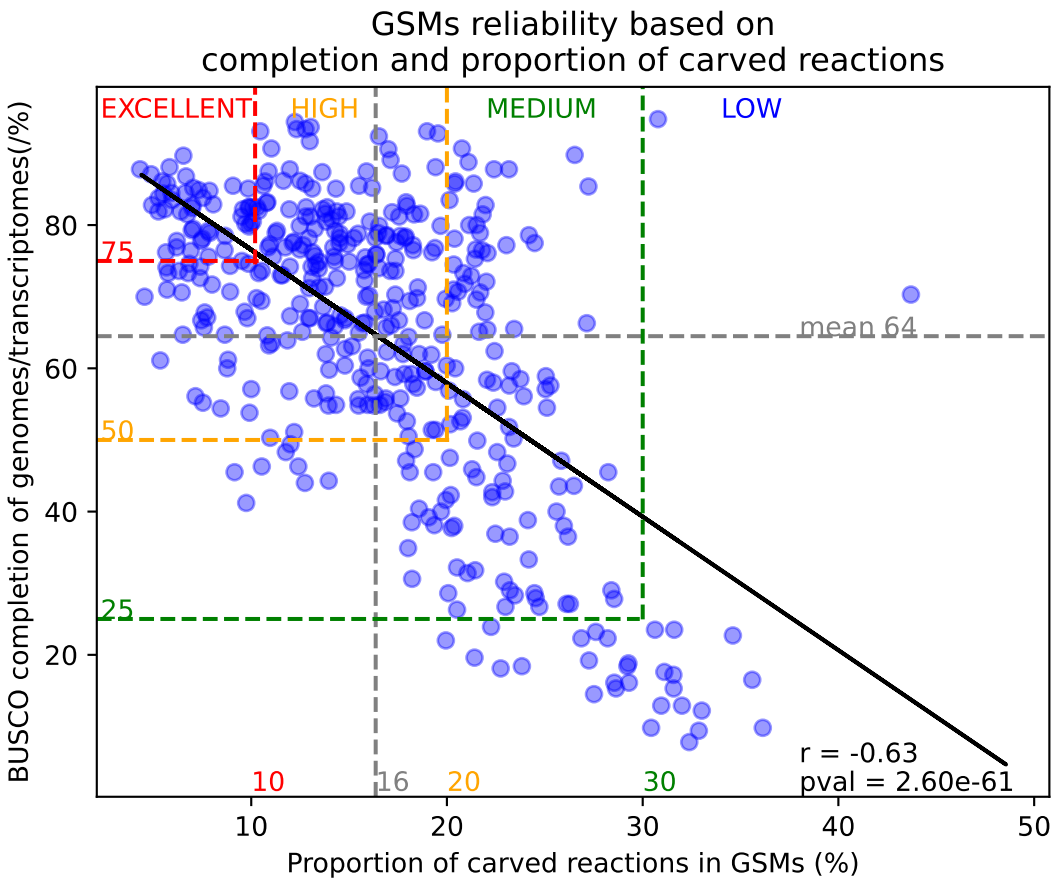
In an epistemological context, model validation and sensitivity analyses are critical steps to ensure the robustness and reliability of the model's predictions. The validation process generally consists of comparing the model's predictions with observed data, the literature or other reliable models, 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.


a)
Photosystem-I flux
Photon export flux
Photosystem-II flux
115.0
102.2
89.5
76.7
63.9
51.1
38.3
25.6
12.8
0.0
b)
ATP synthase (chloroplastic) activity
Photosystem-II flux
Growth
2.345
2.085
1.824
1.564
1.304
1.043
0.783
0.523
0.263
0.002
c)
ATP synthase (chloroplastic) activity
CO2 export flux
RuBisCo activity
113.2
100.6
88.1
75.5
62.9
50.4
37.8
25.2
12.7
0.1

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