Title: Planetary-scale marine community modeling predicts metabolic synergy and viral impacts

Authors:

Régimbeau A.[1]†, Tian F.[2,3]†, Smith G.[2,3], Riddell V J.[2,3,4], Andreani C.[5,15,23], Bordron P.[6,7], Budinich M.[17], Howard-Varona C.[2,3], Larhlimi A.[1], Ser-Giacomi E.[8], Trottier C.[9], Guidi L.[17], Hallam S.J.[10,11], Iudicone D.[12,17], Karsenti E.[17,20], Maass A.[15,17,18,19], Sullivan M.B.[2,3,4,16,17]* & Eveillard, D.[1,17]*

Affiliations:

1. Nantes Université, Centrale Nantes, CNRS, LS2N, 2 rue de la Houssinière, 44322 Nantes, France
2. Department of Microbiology, Ohio State University, Columbus, OH 43210, USA
3. Center of Microbiome Science, Ohio State University, Columbus, OH 43210, USA
4. EMERGE Biology Integration Institute, Ohio State University, Columbus, OH 43210, USA
5. Sorbonne University, CNRS, Laboratoire d'Océanographie de Villefranche, Villefranche-sur-Mer, France
6. Université Fédérale de Toulouse, INRAE, MIAT, Auzeville, CS 52627, 31326 Castanet Tolosan cedex.
7. Université Fédérale de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France

8. IFISC (CSIC-UIB), Institute for Cross-Disciplinary Physics and Complex Systems, Palma de Mallorca, Spain
9. Nantes Université, Inserm, TENS, The Enteric Nervous System in Gut and Brain Diseases, IMAD, 1 rue Gaston Veil, 44035 Nantes, France
10. Department of Microbiology & Immunology, University of British Columbia, Vancouver, BC V6T 1Z1, Canada.
11. Life Sciences Institute, University of British Columbia, Vancouver, BC V6T 1Z3, Canada.
12. Stazione Zoologica Anton Dohrn, 80135 Napoli, Italy.
13. European Molecular Biology Laboratory, Heidelberg, Germany
14. IBENS Département de Biologie, Institut de Biologie de l'ENS, École Normale supérieure, CNRS, INSERM, Université PSL, Paris, France
15. Center for Mathematical Modeling, University of Chile and IRL 2807 CNRS, Santiago, Chile
16. Department of Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA
17. Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara GO-SEE, Paris, France
18. Department of Mathematical Engineering, University of Chile, Santiago, Chile
19. Millennium Institute Center for Genome Regulation, University of Chile, Santiago, Chile
20. Institut de Biologie de l'ENS, Département de Biologie (IBENS), École Normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

† = these authors contributed equally to this work

* = corresponding authors: sullivan.948@osu.edu , damien.eveillard@univ-nantes.fr

Abstract:

The oceans buffer against climate change via biogeochemical cycles underpinned by microbial metabolic networks. While planetary-scale surveys provide baseline microbiome data, inferring metabolic and biogeochemical impacts remains challenging. Here, we constructed a metabolic model for each TARA Ocean metagenome or metatranscriptome, quantified the importance of each metabolic reaction, and used this to assess planetary-scale heterotrophic prokaryotic metabolic phenotypes as a proxy for marine biogeochemistry. This revealed metabolism-inferred ecological zones that matched taxonomy- and function-inferred ones, connections between microbial metabolism and diversity, and predictions about long-sought virus ecological roles, including that virus-encoded metabolic genes target important reactions and identify viral shunt- or shuttle-enriched ocean regions. Together, this framework is agile and resolves planetary-scale community metabolic features to better incorporate microbes and viruses into future predictive ecosystem and climate models.

Main Text:

Connecting "omic" data (DNA, RNA, protein, and metabolites) to ecosystem-level nutrient fluxes and energy flows is a grand challenge in the life sciences (*1–3*). Leveraging petascale microbiome data, a fast growing microbial metabolic knowledgebase (*4*, *5*), and matched physical and chemical information provides opportunities to develop constrained models to interrogate active metabolic processes. However, this will require new modeling approaches as current options cannot yet simultaneously model myriad taxa across the many orders of

magnitude of biological organization (e.g., individual, population, and community levels) required to establish ecosystem-, regional- and even planetary-scale predictions.

Fortunately, advances in several areas offer promise as follows. *First*, the rapid growth in genome-resolved '-omics' observations makes them increasingly available to parameterize large-scale biogeochemical models (*6*) and simulate ocean ecosystem behavior (*7*). *Second*, community-level gene expression data (metatranscriptomes) can now be used to identify transcribed genes that predict regional ocean response patterns (*8*). *Third*, metabolic mapping (*9*) and graph-based (*10*, *11*) approaches – even driven simply by reaction presence or absence – are increasingly used to analyze compositional data (*12*) or semi-quantitative read counts for inferring organismal or multi-organismal metabolic phenotypes (*13*, *14*). *Finally*, metabolic modeling has started using gene and reaction lists to tackle ecosystem-level questions (*15*) where models were effective for assessing mutualistic interactions (*16*), distinguishing competitive or cooperative communities (*17*, *18*), and exploring trade-offs between biological objectives in a community (*19*).

Here, we sought to build upon these efforts by using metabolic modeling, in particular Genome-Scale Models (GSMs), to mechanistically connect microbial metabolic abilities to biogeochemical processes at unprecedented, ecosystem- and planetary-wide scales. In order to build GSMs that can approximate biogeochemistry across such vast scales, we had to 'consider a spherical cow' (*20*) by simplifying how ocean microbiota were represented as follows. *First*, each sample's community metabolism was modeled as a single superorganism (*21*, *22*), where all taxa data were concatenated to form one "community cell" per sample. *Second*, the model relied upon imperfect functional annotations where 18 million (M) "hypothetical" predicted open

reading frames (or ORFs) (*8*) contributed little to the model, such that only 26% (of 47M) and 7% (of 70M) of the ORFs from prokaryote- and virus-enriched samples, respectively, could be mapped to physicochemically defined metabolic networks. To compensate for this, we leveraged a top-down algorithm to "carve" (*23*) a universal metabolic model down to one that matches reaction-mapped gene or transcript data while still satisfying constraints imposed by steady-state thermodynamics. *Third*, we modeled only heterotrophic prokaryotes as no universal model is available for photosynthetic organisms (*23*).

While these simplifications and assumptions were essential due to computational and biochemical challenges, they enabled a new framework for us to apply towards assessing marine biogeochemistry via microbial metabolic modeling at planetary scales. To this end, we applied the approach to genes and transcripts derived from prokaryotes and their viruses from globally distributed *Tara* Oceans metagenomes (*8*, *24*) and metatranscriptomes (*8*) – totaling 127 prokaryote-enriched metagenomes paired with 94 prokaryote-enriched metatranscriptomes and 93 virus-enriched metagenomes that spanned multiple depths from 62 stations sampled throughout the global ocean (**Fig. 1a**, see **Methods**, and **Supplementary Text 1**).

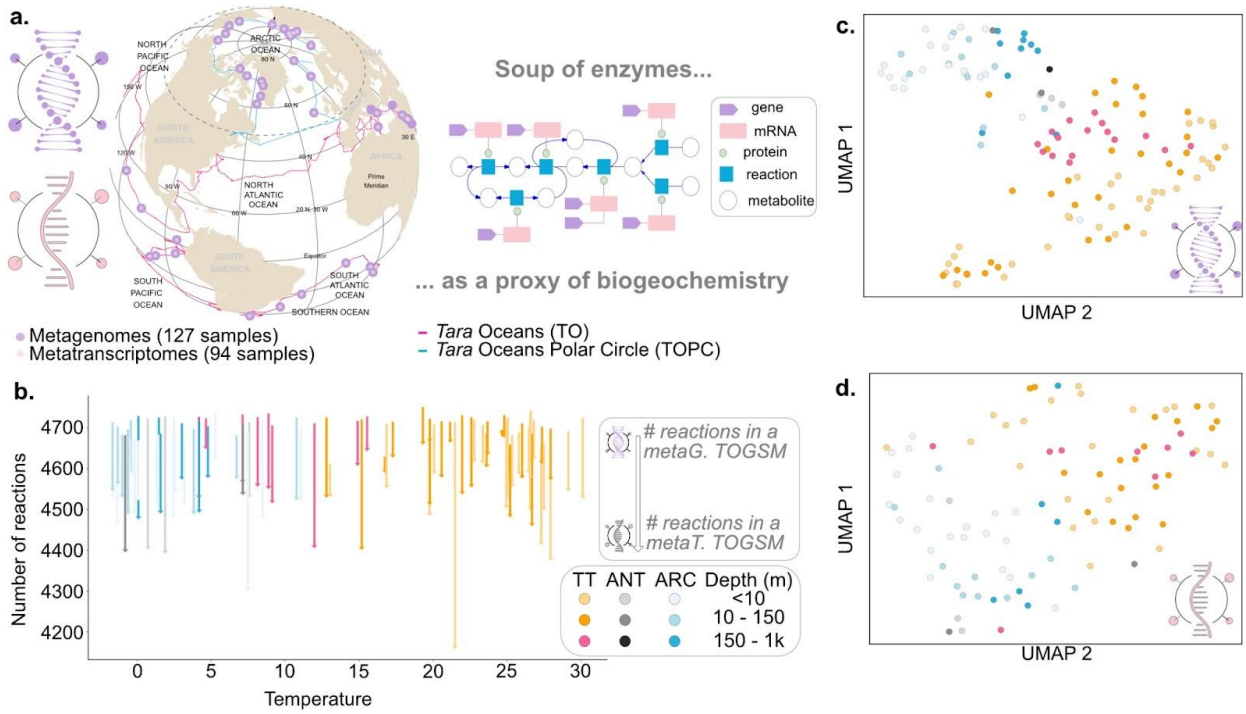

Figure 1. Construction and ecological exploration of *Tara* Ocean Genome-Scale Metabolic Models (TOGSMM). a. Globally distributed metagenomes (n=127) and metatranscriptomes (n=94) (left) were used for genome-scale metabolic model reconstruction. For this, known physics and chemistry of metabolic reactions are captured into metabolic networks (schematic right, where genes are depicted in purple, enzymes in pink, and metabolic reactions in blue that transform metabolites in white). **b.** The number of reactions for each metabolic model was reconstructed per sample from its metagenome, and the matching metatranscriptome was plotted against temperature. In all cases, the former metagenome-based model is larger than the latter metatranscriptome-based model (see directionality of the arrow), which was interpreted as the per-sample historical versus instantaneously sampled functional repertoires, respectively. Color coding represents previous ecological zones defined by microbial and virus community data (Sunagawa et al. 2015; Gregory et al., 2019; TT = Temperate and Tropical, ANT = Antarctic, ARC = Arctic), and the degree of color represents sampling depth as indicated in the figure. **c** and **d.** Ordination (UMAP, or Uniform Manifold Approximation and Projection) of per-sample presence/absence reaction predictions show that genome-scale metabolic models generated from

both metagenomes (c) and metatranscriptomes (d) structure along known ecological zones. Colors are the same as **b.**

In spite of the compute-required simplifications, these reconstructed *Tara* Oceans Genome-Scale Metabolic Models (TOGSMMs) recovered both known and unexpected metabolic features. *First*, the per-sample metagenome-constructed TOGSMMs contained more reactions with genetic evidence than metatranscriptome-constructed TOGSMMs with an average of 4,058 (+/- 28 stdev.) and 3,900 (+/- 79 stdev.) reactions, respectively (**Fig. 1b** and **Supplementary Text 2** for further discussion). These findings are consistent with the idea that a snapshot of a community's expressed transcripts (i.e., the metatranscriptome) will represent only the active subset of genes in the fuller genomic repertoire (i.e., the metagenome). *Second*, the ordination of reaction presence/absence in metagenome- (**Fig. 1c**) and metatranscriptome- (**Fig. 1d**) constructed TOGSMMs accurately recovered global ocean ecological zones known from taxonomy or genome content (*24–26*) (see **Supplementary Text 2**) such that every sample could be consistently categorized into an ecological zone whether the underlying data were derived from microbial or viral taxa, genomes, or now from TOGSMMs. These findings suggest that encoded and/or expressed genes result in a community metabolic signal that directly relates to the microbial and viral community structure used to previously define global ocean ecological zones.

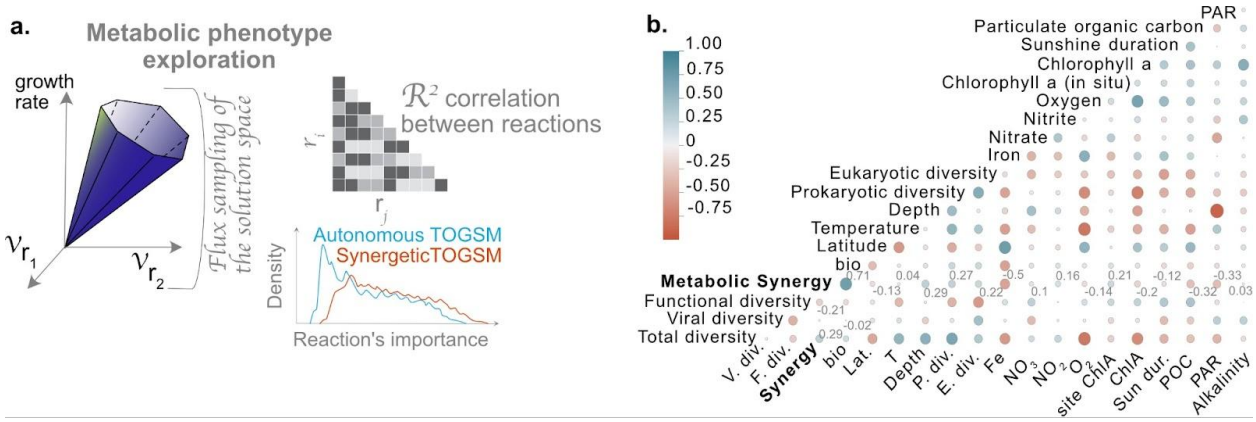

Figure 2 | Computing metabolic synergy across the global ocean and comparing other parameters. **a.** Metabolic solution space drawn from random flux distributions with constraints. A pairwise correlation was computed per reaction pair within a network. The density of reaction correlations determines the synergism of a network. **b.** Correlogram of metabolic reaction synergy and environmental variables.

We next sought to provide a mathematical representation of any given sample's observed metabolic phenotype, which required overcoming the challenges of solving TOGSMMs. While each GSM has a solution space depicting all mathematically feasible reaction fluxes that satisfy the modeling constraints (**Fig. 2a**), the dimensions of the solution space are equal to the number of reactions present in the GSM. This makes TOGSMM analyses impossible to solve exhaustively, given the multidimensionality implied by over four thousand reactions for each. Thus, we explored the theoretical per-sample metabolic solution space through a metabolic flux sampling analysis (*27*) that performs random walks to sample the solution space (**Fig. 2a**, see **Methods** for details).

Through this numerical exploration, we developed approaches to establish, for each TOGSMM, its reaction importance (i.e., a reaction's ability to control fluxes within a metabolic network) and its synergy rank (i.e., biogeochemical interdependence). Towards the former,

reaction importance was calculated from the solution space sampling as the sum of its pairwise flux correlations with all other reactions (see **Methods**), as partially done previously (*28*). Towards the latter, we compared reaction importance density distributions between each TOGSMM (**Supplementary Fig. 14**) to compute a per-sample 'synergy rank' (see **Methods**). When the distribution is shifted towards lower-importance reactions, the synergy rank is lower (i.e., closer to zero) and we assume more autonomous activity between reaction fluxes (illustrated in blue in **Fig. 2a**). Conversely, when the distribution is shifted towards higher-importance reactions, synergy rank is higher (i.e., closer to one) and we assume more cooperative activity, or synergy, between metabolic reactions in the biogeochemical cycles (illustrated in red in **Fig. 2a**). These results allow us to characterize each metabolic phenotype of the heterotrophic prokaryotic community, which are in turn used as estimates of the biogeochemistry process occurring at each sample location.

Phenotype modeling opens a novel window into ocean metabolism

With the modeling framework in place, we next explored how the synergy ranks related to metadata (**Fig. 2b**). This revealed associations between synergy rank and abiotic factors such as iron concentrations (estimated from the NEMO-PISCES Earth System Model, Spearman r: -0.5, p-value: $2x10^{-5}$), as well as biotic factors such as alpha diversity (Spearman r: 0.27, p-value: $1x10^{-3}$). Although only a first-generation model approach, these findings are already consistent with the idea that synergy rank can link and predict metabolic synergy patterns in the ocean based on a combination of taxonomic diversity and limiting factor feature information (*29*).

Pragmatically, this suggests that autonomous biogeochemical systems have relatively fewer important reactions, sustained by a community of low diversity, which are less sensitive to environmental changes, with the opposite true for synergistic biogeochemical systems. Our biogeochemical focus provides a mechanistic and functional contribution to the "diversity-stability debate" by emphasizing the need to assess the nutrient needs of each organism and genome-scale functions of ecosystems as a complement to organismal interaction simulations (*30*).

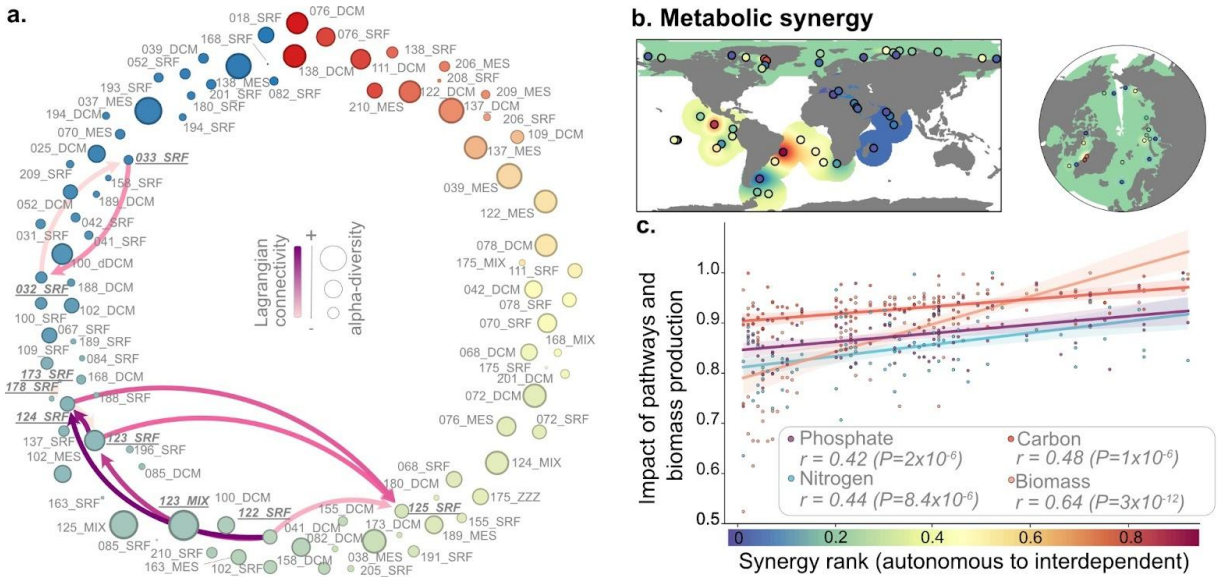

Figure 3 | Metabolic synergy across the global ocean. **a.** Distribution of *Tara* samples based on metabolic reaction synergy from autonomous (blue) to interdependent (red) systems. The node size is proportional to the per-sample prokaryotic organism diversity. Directed edges represent Lagrangian connections between samples realized in less than two months to describe the potential acclimation of water masses. The darker the arrow's color is, the stronger the Lagrangian connection is. In other words, the darker the arrow is, the higher the connection probability between the water masses is. **b.** Projection on the metabolic synergy across the global ocean visualized from an equator- (left) and pole- (right) centric perspectives. The data points were interpolated at each station (halo color on each map), and actual data values are the colored dots. **c.** Relationships between the metabolic synergy of the microbial ecosystem against the

normalized impact of phosphate (purple), nitrogen (blue), and carbon (red) reactions and the potential prokaryotic ecosystem biomass production (orange).

Given this, metabolic synergy also revealed biogeographical patterns (**Fig. 3a** and **3b,** and **Supplementary Text 3**). For example, the metabolic phenotypes of most oligotrophic waters, such as the Mediterranean Sea and the Indian Ocean, appear autonomous-shifted. In contrast, those from nutrient-enriched water masses that follow bloom decay (e.g., Stations 76 and 138) or nutrient release via island effects (e.g., Stations 122 to 125) appear synergetic-shifted. Such patterns could emphasize a relationship between metabolic synergy and microbial productivity via nutrient availability (*29*). Indeed, when we computed the maximum theoretical growth rate associated with each TOGSMM (using standard organismal GSM approaches (*31*)), this revealed a correlation with metabolic synergy rank (**Fig. 3c**; Spearman r: 0.70, p-value: 10^{-14}).

Beyond these global observations, we also explored local metabolic features, where higher-resolution sampling was available. For example, Lagrangian connectivity between each TOGSMM (computed as done previously (*32*)) allowed us to 'connect' samples to assess how metabolic phenotypes changed over short time scales (60 days or less; see edges in **Fig. 3a** and **Methods**). Most Earth System Models simulate a continuous transition between conditions (i.e., each point on the grid) (*33*). However, we observed uneven shifts across the synergy rank gradient rather than seeing a continuous transition from an autonomous state to a synergetic state (or vice versa) as a water mass is transported. This was exemplified in the vicinity of the Marquesas archipelagos, where natural fertilization and the island effect have a significant role in the modeled community from upstream to downstream of the islands (*29*) (i.e., from oligotrophic

Tara Oceans Stations 122, fertilized in 123, and resilience in 124-125 in **Fig. 3a** and **Supplementary Fig. 6**). Problematically, outside this example, the current *Tara* Oceans compendium lacked high-resolution sample sets, emphasizing the need for more transect and time series integrated multi-omic studies *(2)* to constrain metabolic phenotype and synergetic status across representative water masses and environmental gradients.

Beyond providing a new window into the metabolic phenotype of seawater (i.e., metabolic synergy rank), we wondered whether canonical biogeochemical signatures such as the Redfield Ratio (*32*) could be explored through our framework. This decades-old ratio refers to the surprisingly constant stoichiometric ratios between carbon, nitrogen, and phosphorus found in the ocean's plankton and seawater. In our TOGSMMs, we sought to compute the importance of these elements' cycles as assigned using the KEGG database (*34*). While the importance of each cycle scales with synergy rank (**Fig. 3c** and **Supplementary Text 4.c** for discussion on the statistical validity of this correlation), we also observed a remarkably constant proportional importance (std < 0.1 of the mean value for the cycles C:N, C:P and P:N) just as observed in the Redfield Ratio, at least at first principles. However, future applications will benefit from improved data in several areas, including considering other plankton size fractions, moving beyond heterotrophic reactions, and resolving current pathway annotation limitations (see **Supplementary Text 4.a**).

Finally, we evaluated the use of expressed transcripts (regrouped in KO (*8*)) correlating with synergy rank to inform our understanding of how metabolic phenotypes relate to biogeochemical cycles. We specifically focused on expression levels showing a correlation with synergy rank that shifted from positive to negative between polar and temperate subsystems (**Supplementary**

Fig. 13). As expected, this analysis revealed functions associated with prokaryotes and involved in core metabolism (i.e., oxidoreductase activity) or membrane activity (i.e., transferase, molecular transducer, membrane transport, and transmembrane transporter activity). However, surprisingly, this also revealed genes including hydrolases and lyases, potentially encoded by viruses, that we interpreted as suggesting that viruses might strongly impact community-level metabolic phenotypes. To assess this, we analyzed this signal further.

Virus modulated metabolisms and shunt/shuttle biogeography

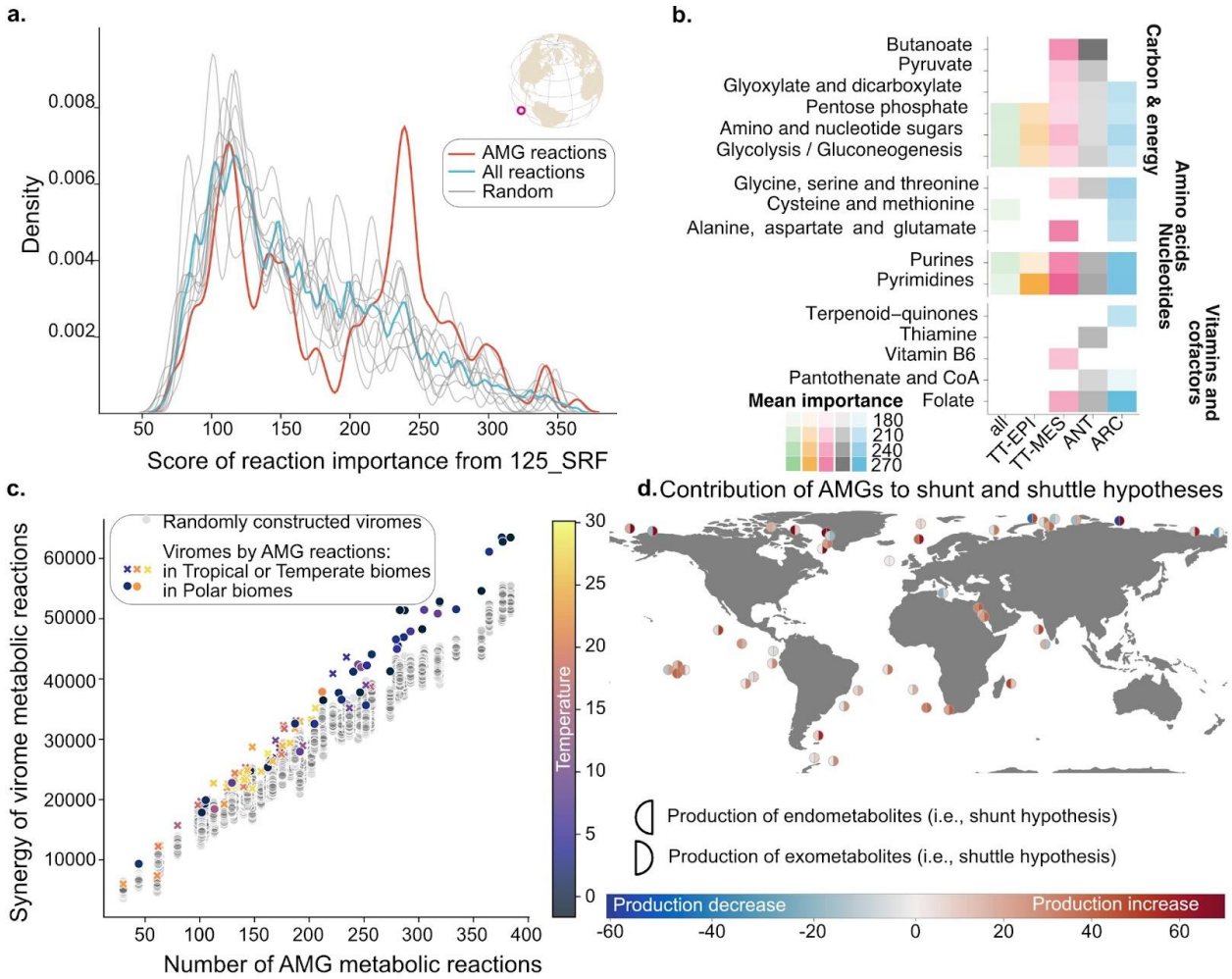

Figure 4 | Importance and impact of reactions targeted by virus-encoded Auxiliary Metabolic Genes (AMGs). **a.** Representative example from sample 125_SRF of the importance of all (blue) and randomly sampled (gray, see **Materials and Methods**) metabolic reactions as compared to the reactions targeted by virus-encoded AMGs (red). All samples showed a similar, statistically significant high-importance AMG spike (see **Supplementary Fig. 14)**. **b.** Metabolic pathways associated with important AMG-targeted reactions across ecological zones. The importance scores of reactions targeted by AMGs, present in all networks and statistically belonging to the high-importance spike, were averaged per KEGG pathway (i.e., lines) and ecological zone (i.e., columns). Importances are then displayed by opacity. Note that a single reaction may be shared among several KEGG pathways. **c.** Comparison between the summed importance of reactions targeted by AMGs and the number of reactions per sample. Each dot represents a sample. Its color code indicates the temperature associated, and its shape

characterizes the system: a circle for polar systems and a cross for temperate systems. Grey dots represent the importance of randomly defined viromes. **d.** Contribution of AMGs to shunt and shuttle hypotheses across the global upper ocean, based on the impact of AMG reactions on exometabolites and endometabolites through increased (i.e., red) or decreased (i.e., blue) metabolite production. Here we consider metabolites that are only exometabolites (n=26), or only endometabolites (n=28) (Moran *et al*. 2022) (*35*).

To explore the role of viruses in modulating community-level metabolic phenotypes, we looked at the presence of conservatively-identified, virus-encoded AMGs (*41*) using cognate viral metagenomes within TOGSMMs (*36*). Strikingly, this revealed that virus-encoded AMGs do not randomly target reactions, but instead predominantly target the most important reactions needed to sustain a metabolic phenotype (**Fig. 4a**, and **Supplementary Fig. 14**). Thus, even though viruses obtain host genes at random through error-prone genomic replication and select for those that improve viral fitness (*37*), this evolutionarily selective process manifests at a community-level towards the important reactions that control the overall metabolic phenotype of the collective microbiome and its associated biogeochemical cycles.

Exploring this further functionally, we found that virus-encoded purine and pyrimidine AMGs – already known to be enriched in viral metagenomes (*38*, *39*) and differentially expressed in marine virocell experiments (*40*) – were among the most important reactions in TOGSMMs (**Fig. 4b**). Presumably this reflects their critical role in the energy demand of viral replication (*41*), and now TOGSMM-contextualization suggests nucleotide-synthesis AMGs significantly impact community level metabolic phenotypes. Other energy-related AMGs targeted fermentation, glycolysis and gluconeogenesis, pentose phosphate pathway, and

two-carbon compound reactions. Presumably, such AMG-impacted metabolisms arise from varied viral host cell reprogramming responses across gradients of nutrient-impacted host physiological states requiring diverse energy resources to drive the viral life cycle, again signatures observed in virocell experiments now placed into *in situ* conditions (*42*, *43*). Biogeographically, AMG-targeted reactions that varied across ecological zones included amino acid metabolism, cell wall biosynthesis, and vitamin biosynthesis pathways, which suggests that virus modulation of host cell wall structures and altering of resources (amino acids and vitamins) in turn also impact community metabolic phenotypes. Further, we found that AMG reactions were significantly more abundant and important in polar systems by estimating the relationship between the number of AMG metabolic reactions and their importance for each TOGSMM (**Fig. 4c**). Such a finding might indicate a higher impact of viruses in the polar community metabolic phenotype and associated biogeochemistry (*44*). Together, such TOGSMM-powered findings highlight critical metabolic pathways for viral replication that can help constrain virocell studies and metabolic models of viral infection and place them into a biogeographical context.

Finally, we wondered whether we could use TOGSMMs to predict virus ecosystem "shunt versus shuttle" effects, which is of collective interest (*35*) due to the ocean's climate buffering role. For a quarter-century, the paradigm was that viruses help heterotrophs rapidly remineralize carbon by making it more available via cell lysis – i.e., the viral shunt hypothesis (*45*). However, aggregating lysis products (*46*, *47*) and viral roles in carbon cycling beyond lysis (e.g., AMGs and metabolic reprogramming) have emerged as additional contributors to biogeochemical cycling (*48*). At least in some instances, viruses are now thought to 'shuttle' carbon to the deep sea via sinking aggregated cells and/or altering grazing efficiency (*42*, *45*, *49*, *50*). Accordingly,

previous *Tara* Oceans work using statistical taxon-based modeling suggested that virus abundances, more than those of prokaryotes or eukaryotes, best predict key biogeochemical processes like carbon flux (*44*).

Expanding upon these prior findings, we used TOGSMMs to explore per-sample shunt-versus-shuttle emphasis interpreted through the lens of the relative contribution of AMGs previously categorized as exo- or endo-metabolites (*35*). For this analysis, we assumed that (i) exometabolites represent sticky materials that increase carbon export via aggregation invoked in the shuttle hypothesis (*51*) and (ii) endometabolites, as intracellular metabolites, might instead be linked to community-level reprogramming and, thus, the remineralization-related shunt hypothesis (**Fig. 4d,** and **Supplementary Fig. 15** for details). Global TOGSMM modeling then revealed that epipelagic layer AMG reactions contribute to both, but with varied per-sample proportions, clear hot-spots biogeographically, and an overall larger contribution to the shuttle hypothesis that was emphasized by a significant correlation between exometabolite productions and Particulate Organic Carbon (Spearman r: 0.34, p-val $4x10^{-3}$). Though an imperfect approach currently – due to poorly characterized seawater metabolites (*35*) and the need to much more broadly characterize metabolic reactions into shunt-versus-shuttle effects – the approach already provides a critically needed map for hotspots to guide hypothesis-testing experiments. If robust and generalizable, these results reinforce the need to better assess biological products' stickiness (*51*) and provide a roadmap for improving understanding of the complex interplay between viral infection and metabolic reprogramming as metabolite measurements improve.

Conclusions

Biology underpins complex systems in nature, where its rules span 12 orders of magnitude, ranging from molecular and cellular to interacting and changing individuals, populations, and communities – all resulting in predictable ecosystem- and planetary-scale outputs. Extensive, systems-level measurements at any of these scales can be challenging, and modeling those data is typically restricted to one or two or, at best, a few scales. This study introduces a modeling approach that can simultaneously study community-wide heterotrophic bacterial metabolism at scales ranging from molecules and cells to the planet. Conceptually, our modeling builds on foundational studies that explore how biological systems self-organize to reveal emerging properties (*52*). When applied to a metabolic abstraction of oceanic heterotrophic prokaryotes and considering physical and chemical laws, we see that such self-organization provides a new window into understanding established patterns in the global ocean.

From a practical viewpoint, the approach leverages rapidly growing high-throughput multi-omics datasets and decades of cumulative metabolism knowledge. It is remarkably agile because it minimizes time-consuming and often grossly data-limited parameterization tasks that can take years to develop and integrate before assessments can be made. Even in this simplified form (see above), the approach already provides (i) new metrics (e.g., ocean metabolic phenotype and its synergetic rank) that could feed into trait-based or biogeochemical models and (ii) metabolic-phenotype-derived hypotheses that offer insights into several ecological grand challenge questions like metabolic reaction drivers of ecosystem stability, functional redundancy of organisms carrying these reactions, the origins of ocean biogeochemical ratios, and the role of

viruses in ecosystem processes. Such effectiveness and agility will only grow as our approach is designed to leverage growing large-scale efforts to characterize functional dark matter (*53*, *54*) and curate reaction databases (reduces gap filling), and there are ready development opportunities for targeted TOGSMM improvements that incorporate autotrophic reactions, abundance instead of presence/absence, and interacting MAG-constrained GSMs. As the approach matures, we envision that this ocean-focused work will help guide realistic societal actions surrounding ocean-based carbon storage as a geoengineering solution to climate change. However, the approaches are generalizable such that they should also help transform our understanding anywhere microbiomes and viromes are driving ecosystems.

References

1. Z. N. Oltvai, A.-L. Barabási, Life's Complexity Pyramid. *Science* **298**, 763–764 (2002).
2. A. S. Hahn, K. M. Konwar, S. Louca, N. W. Hanson, S. J. Hallam, The information science of microbial ecology. *Curr. Opin. Microbiol.* **31**, 209–216 (2016).
3. A. Tagliabue, 'Oceans are hugely complex': modelling marine microbes is key to climate forecasts. *Nature* **623**, 250–252 (2023).
4. Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, N. E. Lewis, BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–D522 (2016).
5. I. Dukovski, D. Bajić, J. M. Chacón, M. Quintin, J. C. C. Vila, S. Sulheim, A. R. Pacheco, D. B. Bernstein, W. J. Riehl, K. S. Korolev, A. Sanchez, W. R. Harcombe, D. Segrè, A metabolic modeling platform for the computation of microbial ecosystems in time and space (COMETS). *Nat. Protoc.* **16**, 5030–5082 (2021).
6. T. Mock, S. J. Daines, R. Geider, S. Collins, M. Metodiev, A. J. Millar, V. Moulton, T. M. Lenton, Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. *Glob. Change Biol.* **22**, 61–75 (2016).
7. V. J. Coles, M. R. Stukel, M. T. Brooks, A. Burd, B. C. Crump, M. A. Moran, J. H. Paul, B. M. Satinsky, P. L. Yager, B. L. Zielinski, R. R. Hood, Ocean biogeochemistry modeled with emergent trait-based genomics. *Science* **358**, 1149–1154 (2017).
8. G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M.

Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).

9. A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, B. Ø. Palsson, Reconstruction of Biochemical Networks in Microbial Organisms. *Nat. Rev. Microbiol.* **7**, 129–143 (2009).

10. P. Bordron, M. Latorre, M.-P. Cortés, M. González, S. Thiele, A. Siegel, A. Maass, D. Eveillard, Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *MicrobiologyOpen* **5**, 106–117 (2016).

11. K. M. Konwar, N. W. Hanson, A. P. Pagé, S. J. Hallam, MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* **14**, 202 (2013).

12. M. Royo-Llonch, P. Sánchez, C. Ruiz-González, G. Salazar, C. Pedrós-Alió, K. Labadie, L. Paoli, T. O. Coordinators, S. Chaffron, D. Eveillard, E. Karsenti, S. Sunagawa, P. Wincker, L. Karp-Boss, C. Bowler, S. G. Acinas, Ecogenomics of key prokaryotes in the arctic ocean. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.06.19.156794.

13. M. Kim, I.-T. Cha, K.-E. Lee, M. Li, S.-J. Park, Pangenome analysis provides insights into the genetic diversity, metabolic versatility, and evolution of the genus Flavobacterium. *Microbiol. Spectr.* **11**, e01003-23 (2023).

14. I. Veseli, Y. T. Chen, M. S. Schechter, C. Vanni, E. C. Fogarty, A. R. Watson, B. Jabri, R. Blekhman, A. D. Willis, M. K. Yu, A. Fernàndez-Guerra, J. Füssel, A. M. Eren, Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. *bioRxiv*, 2023.05.10.540289 (2023).

15. H. Scott, D. Segrè, Metabolic Flux Modeling in Marine Ecosystems. doi: 10.1146/annurev-marine-032123-033718 (2024).

16. S. Stolyar, S. Van Dien, K. L. Hillesland, N. Pinel, T. J. Lie, J. A. Leigh, D. A. Stahl, Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* **3**, 92 (2007).

17. D. Machado, O. M. Maistrenko, S. Andrejev, Y. Kim, P. Bork, K. R. Patil, K. R. Patil, Polarization of microbial communities between competitive and cooperative metabolism. *Nat. Ecol. Evol.* **5**, 195–203 (2021).

18. N. Giordano, M. Gaudin, C. Trottier, E. Delage, C. Nef, C. Bowler, S. Chaffron, Genome-scale community modelling reveals conserved metabolic cross-feedings in epipelagic bacterioplankton communities. *Nat. Commun.* **15**, 2721 (2024).

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

20. J. Harte, *Consider A Spherical Cow: A Course in Environmental Problem Solving* (University Science Books, 1988).

21. K. Liautaud, E. H. van Nes, M. Barbier, M. Scheffer, M. Loreau, Superorganisms or loose collections of species? A unifying theory of community patterns along environmental gradients. *Ecol. Lett.* **22**, 1243–1252 (2019).

22. M. B. Biggs, G. L. Medlock, G. L. Kolling, J. A. Papin, Metabolic network modeling of microbial communities. *WIREs Syst. Biol. Med.* **7**, 317–334 (2015).
23. D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46**, 7542–7553 (2018).
24. A. C. Gregory, A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-É. Tremblay, D. Vik, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, M. B. Sullivan, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
25. J.-F. Ghiglione, P. E. Galand, T. Pommier, C. Pedrós-Alió, E. W. Maas, K. Bakker, S. Bertilson, D. L. Kirchman, C. Lovejoy, P. L. Yager, A. E. Murray, Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc. Natl. Acad. Sci.* **109**, 17633–17638 (2012).
26. G. Dominguez-Huerta, A. A. Zayed, J. M. Wainaina, J. Guo, F. Tian, A. A. Pratama, B. Bolduc, M. Mohssen, O. Zablocki, E. Pelletier, E. Delage, A. Alberti, J.-M. Aury, Q. Carradec, C. da Silva, K. Labadie, J. Poulain, TARA OCEANS COORDINATORS, C. Bowler, D. Eveillard, L. Guidi, E. Karsenti, J. H. Kuhn, H. Ogata, P. Wincker, A. Culley, S. Chaffron, M. B. Sullivan, Diversity and ecological footprint of Global Ocean RNA viruses. *Science* **376**, 1202–1208 (2022).
27. W. Megchelenbrink, M. Huynen, E. Marchiori, optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks. *PLoS ONE* **9** (2014).
28. J.-M. Schwartz, H. Otokuni, T. Akutsu, J. C. Nacher, Probabilistic controllability approach to metabolic fluxes in normal and cancer tissues. *Nat. Commun.* **10**, 2725 (2019).
29. L. Caputi, Q. Carradec, D. Eveillard, A. Kirilovsky, E. Pelletier, J. J. P. Karlusich, F. R. J. Vieira, E. Villar, S. Chaffron, S. Malviya, E. Scalco, S. G. Acinas, A. Alberti, J.-M. Aury, A.-S. Benoiston, A. Bertrand, T. Biard, L. Bittner, M. Boccara, J. R. Brum, C. Brunet, G. Busseni, A. Carratalà, H. Claustre, L. P. Coelho, S. Colin, S. D'Aniello, C. D. Silva, M. D. Core, H. Doré, S. Gasparini, F. Kokoszka, J.-L. Jamet, C. Lejeusne, C. Lepoivre, M. Lescot, G. Lima-Mendez, F. Lombard, J. Lukeš, N. Maillet, M.-A. Madoui, E. Martinez, M. G. Mazzocchi, M. B. Néou, J. Paz-Yepes, J. Poulain, S. Ramondenc, J.-B. Romagnan, S. Roux, D. S. Manta, R. Sanges, S. Speich, M. Sprovieri, S. Sunagawa, V. Taillandier, A. Tanaka, L. Tirichine, C. Trottier, J. Uitz, A. Veluchamy, J. Veselá, F. Vincent, S. Yau, S. Kandels-Lewis, S. Searson, C. Dimier, M. Picheral, P. Bork, E. Boss, C. de Vargas, M. J. Follows, N. Grimsley, L. Guidi, P. Hingamp, E. Karsenti, P. Sordino, L. Stemmann, M. B. Sullivan, A. Tagliabue, A. Zingone, L. Garczarek, F. d'Ortenzio, P. Testor, F. Not, M. R. d'Alcalà, P. Wincker, C. Bowler, D. Iudicone, Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. *Glob. Biogeochem. Cycles* **33**, 391–419 (2019).
30. J. R. Bernhardt, P. Kratina, A. L. Pereira, M. Tamminen, M. K. Thomas, A. Narwani, The

evolution of competitive ability for essential resources. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20190247 (2020).

31. N. D. Price, J. L. Reed, B. Ø. Palsson, Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897 (2004).
32. A. C. Redfield, *On the Proportions of Organic Derivatives in Sea Water and Their Relation to the Composition of Plankton* (University Press of Liverpool, 1934).
33. K. Fennel, J. P. Mattern, S. C. Doney, L. Bopp, A. M. Moore, B. Wang, L. Yu, Ocean biogeochemical modelling. *Nat. Rev. Methods Primer* **2**, 1–21 (2022).
34. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
35. M. A. Moran, E. B. Kujawinski, W. F. Schroer, S. A. Amin, N. R. Bates, E. M. Bertrand, R. Braakman, C. T. Brown, M. W. Covert, S. C. Doney, S. T. Dyhrman, A. S. Edison, A. M. Eren, N. M. Levine, L. Li, A. C. Ross, M. A. Saito, A. E. Santoro, D. Segrè, A. Shade, M. B. Sullivan, A. Vardi, Microbial metabolites in the marine carbon cycle. *Nat. Microbiol.* **7**, 508–523 (2022).
36. F. Tian, J. W. Wainaina, C. Howard-Varona, B. Bolduc, M. C. Gazitúa, G. Smith, M. R. Gittrich, O. Zablocki, D. Cronin, D. Evelliard, S. J. Hallam, M. B. Sullivan, Virus-encoded auxiliary metabolic genes throughout the global oceans. *Submitted*.
37. J. C. Ignacio-Espinoza, M. B. Sullivan, Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. *Environ. Microbiol.* **14**, 2113–2126 (2012).
38. D. Vik, M. C. Gazitúa, C. L. Sun, A. A. Zayed, M. Aldunate, M. R. Mulholland, O. Ulloa, M. B. Sullivan, Genome-resolved viral ecology in a marine oxygen minimum zone. *Environ. Microbiol.* **23**, 2858–2874 (2021).
39. H. Enav, Y. Mandel-Gutfreund, O. Béjà, Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome* **2**, 9 (2014).
40. C. Howard-Varona, M. M. Lindback, G. E. Bastien, N. Solonenko, A. A. Zayed, H. Jang, B. Andreopoulos, H. M. Brewer, T. Glavina del Rio, J. N. Adkins, S. Paul, M. B. Sullivan, M. B. Duhaime, Phage-specific metabolic reprogramming of virocells. *ISME J.* **14**, 881–895 (2020).
41. G. Mahmoudabadi, R. Milo, R. Phillips, Energetic cost of building a virus. *Proc. Natl. Acad. Sci.* **114**, E4324–E4333 (2017).
42. C. F. Kranzler, M. A. Brzezinski, N. R. Cohen, R. H. Lampe, M. Maniscalco, C. P. Till, J. Mack, J. R. Latham, K. W. Bruland, B. S. Twining, A. Marchetti, K. Thamatrakoln, Impaired viral infection and reduced mortality of diatoms in iron-limited oceanic regions. *Nat. Geosci.* **14**, 231–237 (2021).
43. C. Howard-Varona, M. M. Lindback, J. D. Fudyma, A. Krongauz, N. E. Solonenko, A. A. Zayed, W. B. Andreopoulos, H. M. Olson, Y.-M. Kim, J. E. Kyle, T. Glavina del Rio, J. N. Adkins, M. M. Tfaily, S. Paul, M. B. Sullivan, M. B. Duhaime, Environment-specific virocell metabolic reprogramming. *ISME J.* **18**, wrae055 (2024).
44. L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. R. Brum, L. P. Coelho, J. C. I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, S. Kandels-Lewis, M. Picheral, J. Poulain, S. Searson, L. Stemmann, F. Not, P. Hingamp, S. Speich, M. Follows, L. Karp-Boss, E. Boss, H. Ogata, S. Pesant, J. Weissenbach, P. Wincker, S. G. Acinas, P. Bork, C. de Vargas, D. Iudicone, M. B. Sullivan, J. Raes, E. Karsenti, C. Bowler, G. Gorsky, Plankton networks driving carbon export in the

oligotrophic ocean. *Nature* **532**, 465–470 (2016).
45. A. J. Poulton, Shunt or shuttle. *Nat. Geosci.* **14**, 181–183 (2021).
46. M. G. Weinbauer, Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
47. B. L. Hurwitz, J. R. Brum, M. B. Sullivan, Depth-stratified functional and taxonomic niche specialization in the "core" and "flexible" Pacific Ocean Virome. *ISME J.* **9**, 472–484 (2015).
48. A. E. Zimmerman, C. Howard-Varona, D. M. Needham, S. G. John, A. Z. Worden, M. B. Sullivan, J. R. Waldbauer, M. L. Coleman, Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat. Rev. Microbiol.* **18**, 21–34 (2020).
49. J. S. Weitz, S. W. Wilhelm, Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**, 17 (2012).
50. M. B. Sullivan, J. S. Weitz, S. Wilhelm, Viral ecology comes of age. *Environ. Microbiol. Rep.* **9**, 33–35 (2017).
51. A. B. Burd, G. A. Jackson, Particle Aggregation. *Annu. Rev. Mar. Sci.* **1**, 65–90 (2009).
52. E. Karsenti, Self-organization in cell biology: a brief history. *Nat. Rev. Mol. Cell Biol.* **9**, 255–262 (2008).
53. G. A. Pavlopoulos, F. A. Baltoumas, S. Liu, O. Selvitopi, A. P. Camargo, S. Nayfach, A. Azad, S. Roux, L. Call, N. N. Ivanova, I. M. Chen, D. Paez-Espino, E. Karatzas, I. Iliopoulos, K. Konstantinidis, J. M. Tiedje, J. Pett-Ridge, D. Baker, A. Visel, C. A. Ouzounis, S. Ovchinnikov, A. Buluç, N. C. Kyrpides, Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
54. T. S. B. Schmidt, A. Fullam, P. Ferretti, A. Orakov, O. M. Maistrenko, H.-J. Ruscheweyh, I. Letunic, Y. Duan, T. Van Rossum, S. Sunagawa, D. R. Mende, R. D. Finn, M. Kuhn, L. Pedro Coelho, P. Bork, SPIRE: a Searchable, Planetary-scale mIcrobiome REsource. *Nucleic Acids Res.* **52**, D777–D783 (2024).

Acknowledgments: 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/).

Funding: We also acknowledge the commitment of the CNRS to support A.R. Ph.D. funds (CNRS 80 Prime), and the Simons Postdoctoral Fellowships in Marine Microbial Ecology. Some computations were performed thanks to France's BiRD and LIGER computing facilities. This

study was partly supported by ATLANTECO and Horizon Europe project BIOcean5D (award number 101059915). AM is funded by Projects ANID–MILENIO–ICN2021044, Exploración-13220002 and Basal-FB210005. E.S-G. acknowledges support from grant PID2021-123352OB-C32 funded by MICIU/AEI/10.13039/501100011033 and FEDER, 'Una manera de hacer Europa'. A

Author contributions: D.E designed the experiment. E.SG computed the Lagrangian data. F.T reconstructed the models. A.R simulated the models. A.R, F.T and G.S analyzed the data. A.R, F.T, G.S, MB.S and D.E wrote the initial manuscript. J.R, C.A, P.B, M.P, C.HV, A.L, E.SG, C.T, L.G, SJ.H, D.I, E.K, and A.M help writing the final manuscript.

Competing interests: SJH has a conflict of interest; he is a co-founder of Koonkie Inc., a bioinformatics consulting company that designs and provides scalable algorithmic and data analytics solutions in the cloud.

This article is contribution number XXX of *Tara* Oceans.

Data and materials availability: All data used in this paper is already available (-omics) and the models used will be released upon acceptance of the manuscript.

List of Supplementary Materials:

- Materials and Methods
- Supplementary Text 1- 7
- Supplementary Fig 1-15 (Supplementary Fig. 14 is in a separate PDF)

Materials and Methods

The global ocean dataset

The dataset for metabolic network reconstruction consists of paired prokaryote- (n = 127) and virus-enriched (n = 127) metagenomes and the corresponding prokaryotic metatranscriptomes (n = 94) from previously published datasets from *Tara* Oceans (TO) and *Tara* Oceans Polar Circle (TOPC) expeditions (*1*, *2*). These samples were collected from 62 sampling stations, spanning the global ocean, including the South Pacific Ocean, North Pacific Ocean, South Atlantic Ocean, Indian Ocean, Red Sea, Mediterranean Sea, Southern Ocean, North Atlantic, and Arctic Ocean.

Generation of the reference gene catalog

Representative coding sequence (cds) or gene features were selected for prokaryotic microbial and viral fractions. In the case of prokaryotic cds features, the Ocean Microbial Reference Gene Catalog (OM-RGC.v2)(*1*), which included 46,775,174 features, was used as the reference catalog. In the case of viral cds features, genes were called from assemblies in virus-enriched metagenomes (n = 127) using Prodigal (v2.6.1) in the metagenomic mode (*3*). The resulting genes were grouped into clusters using mmseqs2 (*4*) (easy-cluster, -s 7.5, e-value 0.001) at two sets of parameters (30% ANI across 70% of the length and 60% ANI across 80% of the length). To establish functional relevance, 1000 gene clusters, with more than 20 genes per cluster, were randomly selected for each set of parameters. For functional annotation, the cds features within each cluster were searched against the RefSeq database (v208). Ambiguous clusters were further assessed using the Pfam database (*5*) (v33.1). Collectively, this analysis identified representative

cds features (n = 70,347,973) clustered at 30% ANI across 70% of the length as the reference catalog for subsequent modeling efforts.

Calculation of metagenomic and metatranscriptomic profiles

We profiled metagenomic and metatranscriptomic cds features for use in constructing TOGSMMs. The gene and transcript abundance calculations of OM-RGC.v2 from the prokaryotic metagenomes (n = 127) were previously reported (*1*) and used in this study. To calculate gene abundance from viral metagenomes, amino acid sequences of gene representatives were searched against the BiGG database (*6*) (v1.6) using DIAMOND(*7*) (v0.9.24.125) (--more-sensitive, e-value 0.001). This resulted in 4,939,042 genes matched to reactions in the BiGG database. Quality-controlled reads of viral metagenomes (*2*) (n = 127) were non-deterministically recruited to the 5M reference genes using CoverM (v.0.2.0) (https://github.com/wwood/CoverM) with the trimmed mean method. Reads were kept if they were with ≥ 95% identity and ≥75% read coverage. Genes with < 70% coverage within a sample were not considered. Similar strategies were used to calculate the metatranscriptomic composition and map prokaryotic metatranscriptomes (n = 94) to the reference genes in the viral metagenomes.

Building a *Tara* Ocean Genome-Scale Metabolic Model

Metabolic networks were generated per paired samples. For each paired sample, genes with an abundance > 0 calculated from metagenomes were piped through CarveME (*8*) (v1.5.1) using a universal model of bacterial metabolism. The resulting metabolic networks (n = 127) were gap-filled using the marine media (**Table S1**). Similar strategies were applied to genes with

transcriptomic abundance > 0. These identified 94 networks derived from the transcriptomic abundance profile.

We used the classical constraints used in metabolic modeling to obtain the model. To each network, we associated a set of constraints:

$$\begin{cases} \mathbf{S}\mathbf{v} = \mathbf{0} \\ \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} \end{cases},$$

where $\mathbf{S} \in \mathbb{R}^{m,n}$ is the stoichiometric matrix of the network composed of m metabolites and n reactions, $\mathbf{v} \in \mathbb{R}^n$ is the flux vector that represents the flux that goes into each reaction of the network, and $\mathbf{lb}, \mathbf{ub} \in \mathbb{R}^n$ are the lower and upper bounds of the flux vector (default values from the CarveMe algorithm). TOGSMMs are available in here: https://uncloud.univ-nantes.fr/index.php/s/9nXG6kRz66CwDAW (a zenodo link and a DOI will be added upon acceptance of the manuscript).

Auxiliary metabolic genes and their mapping in TOGSMMs

Systematically cataloged AMGs from the global ocean were used to assess the role of viruses in community-level metabolic phenotype. The Global Oceans AMG catalog was previously established. Briefly, AMGs were called on viral contigs with additional curation steps to exclude non-viral regions, mobile genetic elements, and other genes that may function in the random integration of microbial metabolic genes (*9*). To map the Global Oceans AMG catalog (n = 22,779) in TOGSMMs, we searched them against the BiGG database (1.6) present in CarveMe universe using DIAMOND (v2.0.14.152, –more-sensitive). We then applied the logical rules present in CarveMe's GPR to obtain the reactions AMG supports.

Sampling of the *Tara* Ocean Genome-Scale Models

Each metabolic model defines a solution space $\mathcal{F}$, that is, the ensemble of solutions that satisfies the constraint:

$$\mathcal{F} := \{\mathbf{v} \in \mathbb{R}^n, \mathbf{S}\mathbf{v} = 0, \quad \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub}\}$$

Describing the solution space of a metabolic model is equivalent to assessing its metabolic phenotype. To do so, we used optGpSampler (*10*), which allows us to draw random flux distributions from the solution space. We constrained the biomass reaction to produce at least 10% of its maximum. This subjective threshold allows us to discard states with low biomass production, focusing our attention on states that assure the ecosystem's survival.

We sampled 10^5 points with a thinning of 10^4. From this sampling procedure, we computed a pairwise Pearson correlation between each network reaction resulting in the correlation matrix $\mathbf{C} \in R^{(n,n)}$.

Importance of reactions and cycles

Reaction importance is the sum of its pairwise Pearson correlation in absolute value over the entire network. Formally we write the reaction's j importance as $I_j = \sum_{k=1}^{n} \|C_{j,k}\|$, where the sum is done over the set of reactions.

A reaction of high importance is significantly correlated with many other reactions in the network; thus, a change in its flux value will change the fluxes of many other reactions. On the other hand, a reaction with low importance can see its flux vary without changing the overall flux distribution.

Cycle importance is the sum of the reactions' importance composing the cycle. We used KEGG identifiers available in CarveMe reaction database to obtain the composition of cycles. To avoid subjective thresholds on correlation values, we used all correlations. The low correlations due to random effects appear in all the reactions and are uniformly distributed. As such they do not impact our results.

Viruses impact the system through AMGs.

AMGs encoded by viruses have been demonstrated to modify host metabolism sufficiently to alter host biogeochemical/metabolite flux profile (*11*). To incorporate viruses in our metabolic network, we considered AMG-targeted reactions that are in the established TOGSMMs. We estimated the impact of viruses through AMGs using the previously computed reactions' summed importance scores. To compute the impact of viruses on the production of metabolites, we computed the contribution of each reaction in the production or consumption of the considered metabolites. We then looked at the impact of an increase in each reaction flux from the reactions supported by AMGs. This impact was computed by multiplying the participation of the reactions to the dynamic of the metabolites by their correlations to the reactions supported by AMGs.

Formally for a given TOGSMM, this gives:

$$V_k = \sum_{i \in \mathcal{V}} \sum_{j \in [\![1..n]\!]} S_{k,j} C_{i,j} \quad ,$$

where $\mathcal{V}$ is the set of metabolic reactions supported by AMGs, $S_{k,j}$ is the stoichiometric coefficient of metabolite k in reaction j, and $C_{i,j}$ is the pairwise correlation between reaction i and reaction j.

To estimate random viromes for a given sample, we took the total number of reactions supported by the computed virome and randomly chose that many reactions in the network. This resulted in a random distribution of importance and a random importance score of reactions targeted by AMGs.

Stability of the ecosystem

In our framework, stability represents how sensible the system is regarding a change of flux values. The correlation value from the sampling work allows us to assess this stability. A network with many reactions' correlations near zero is considered to represent a stable ecosystem because a change in a single reaction's flux will not significantly change the overall flux distribution and, thus, the estimated behavior of the ecosystem. However, a network with many highly correlated reactions will be more sensitive to changes in a given reaction's flux, and thus will represent a less stable ecosystem.

We used a Kolmogorov-Smirnov test (KS-test) on the absolute pairwise correlation values distribution to compute this metric. We computed the KS distance and its associated p-value. We then created the directed graph, where edges were considered if the test p-value was not significant (>5*1e-2), meaning that we could not reject the null hypothesis. An edge from station A to station B means that the ecosystem captured in station A is more stable than the one modeled in station B.

From this graph, one can compute each node's out-degree, which is the number of edges that start with the considered node. Once normalized, this out degree corresponds to the sample's synergy score, as represented in **Fig. 2**.

Endo and exo-metabolite identification and linking to reactions

Endo and exo-metabolites were determined using Table 1 of Moran et al 2022 (*12*). The metabolites table was joined to the reagents and products of each reaction used in the models to determine which reactions were directly associated with the transformation of endo and exometabolites.

1. G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).
2. A. C. Gregory, A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-É. Tremblay, D. Vik, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, M. B. Sullivan, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
3. D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
4. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
5. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
6. J. Schellenberger, J. O. Park, T. M. Conrad, B. Ø. Palsson, BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11**, 213 (2010).
7. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
8. D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* **46**, 7542–7553 (2018).
9. F. Tian, J. M. Wainaina, C. Howard-Varona, G. Domínguez-Huerta, B. Bolduc, M. C.

Gazitúa, G. Smith, M. R. Gittrich, O. Zablocki, D. R. Cronin, D. Eveillard, S. J. Hallam, M. B. Sullivan, Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. *Microbiome* **12**, 159 (2024).

10. W. Megchelenbrink, M. Huynen, E. Marchiori, optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks. *PLoS ONE* **9** (2014).

11. B. L. Hurwitz, S. J. Hallam, M. B. Sullivan, Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol.* **14**, R123 (2013).

12. M. A. Moran, E. B. Kujawinski, W. F. Schroer, S. A. Amin, N. R. Bates, E. M. Bertrand, R. Braakman, C. T. Brown, M. W. Covert, S. C. Doney, S. T. Dyhrman, A. S. Edison, A. M. Eren, N. M. Levine, L. Li, A. C. Ross, M. A. Saito, A. E. Santoro, D. Segrè, A. Shade, M. B. Sullivan, A. Vardi, Microbial metabolites in the marine carbon cycle. *Nat. Microbiol.* **7**, 508–523 (2022).