Integrating biogeochemistry with multomic sequence information in a model oxygen minimum zone

Stilianos Louca,a Alyse K. Hawley,b Sergei Katsev,c–d Monica Torres-Beltran,a Maya P. Bhatia,a,b,c Sam Kheirandish,a,b Céline C. Michiels,a David Capelle,a Gaute Lavikc Michael Doebeli,d,e,f Sean A. Croweb,f,g,h, and Steven J. Hallamb,e,f,g,h,i

aInstitute of Applied Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2; bDepartment of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada V6T 1Z2; cLarge Lakes Observatory, University of Minnesota Duluth, Duluth, MN 55812; dDepartment of Physics and Astronomy, University of Minnesota Duluth, Duluth, MN 55812; eCanadian Institute for Advanced Research Program in Integrated Microbial Biodiversity, Canadian Institute for Advanced Research, Toronto, ON, Canada M5G 1Z8; fDepartment of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; gBiogeochemistry Group, Max Planck Institute for Marine Microbiology, Bremen D-28359, Germany; hDepartment of Zoology, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; iDepartment of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; jEcosystem Services, Commercialization Platforms, and Entrepreneurship (ECOSCOPE) Training Program, University of British Columbia, Vancouver, BC, Canada V6T 1Z2; and kPeter Wall Institute for Advanced Studies, University of British Columbia, Vancouver, BC, Canada V6T 1Z2

www.pnas.org/cgi/doi/10.1073/pnas.1602897113 PNAS

Published online September 21, 2016 | E5925–E5933

Microorganisms are the most abundant lifeform on Earth, mediating global fluxes of matter and energy. Over the past decade, high-throughput molecular techniques generating multomic sequence information (DNA, mRNA, and protein) have transformed our perception of this microcosmos, conceptually linking microorganisms at the individual, population, and community levels to a wide range of ecosystem functions and services. Here, we develop a biogeochemical model that describes metabolic coupling along the redox gradient in Saanich Inlet—a seasonally anoxic fjord with biogeochemistry analogous to oxygen minimum zones (OMZs). The model reproduces measured biogeochemical process rates as well as DNA, mRNA, and protein concentration profiles across the redox gradient. Simulations make predictions about the role of ubiquitous OMZ microorganisms in mediating carbon, nitrogen, and sulfur cycling. For example, nitrite “leakage” during incomplete sulfide-driven denitrification by SUP05 Gammaproteobacteria is predicted to support inorganic carbon fixation and intense nitrogen loss via anaerobic ammonium oxidation. This coupling creates a metabolic niche for nitrous oxide reduction that completes denitrification by currently unidentified community members. These results quantitatively improve previous conceptual models describing microbial metabolic networks in OMZs. Beyond OMZ-specific predictions, model results indicate that geochemical fluxes are robust indicators of microbial community structure and reciprocity, that gene abundances and geochemical conditions largely determine gene expression patterns. The integration of real observational data, including geochemical profiles and process rate measurements as well as metagenomic, metatranscriptomic and metaproteomic sequence data, into a biogeochemical model, as shown here, enables holistic insight into the microbial metabolic network driving nutrient and energy flow at ecosystem scales.

Significance

Modern molecular sequencing is beginning to provide great insight into microbial community structure and function at ecosystem scales. However, the quantitative integration of multomic sequence information (DNA, mRNA, and protein) and geochemical processes has so far been elusive. Here, we develop a biogeochemical model that integrates geochemistry and multomic sequence information to explain key metabolic processes in the oxygen-starved waters of Saanich Inlet, a model ecosystem for studying microbial community responses to oxygen minimum zone expansion. Our model largely explains DNA, mRNA, and protein distributions and sheds light on the metabolic networks coupling carbon, sulfur, and nitrogen transformations across a redox gradient. Our approach is extensible to other biogeochemical models incorporating feedbacks of global change on ecosystem functions.


This article is a PNAS Direct Submission.

Data deposition: All sequencing data have been deposited in publicly accessible databases, as follows. The raw metagenomic sequence data reported in this paper have been deposited in the Joint Genome Institute (JGI) Genomes Online Database (GOLD) (Project ID nos. Gp0052414 and Gp0052384–Gp0052387). The metagenome assemblies reported in this paper have been deposited in the JGI Integrated Microbial Genomes and Microbiomes Samples Database (Taxon Object ID nos. 3300000198, 3300000213, 3300000255, 3300000256, and 3300000257). The raw metatranscriptomic sequence data reported in this paper have been deposited in the JGI GOLD (GOLD Project ID nos. Gp005274–Gp0052297 and Gp005281). The protein sequences reported in this paper have been deposited in the Mass Spectrometry Interactive Virtual Environment (MassIVE) accession no. MSV000097878; proteome exchange ID no. PXD004493). The KEGG (Kyoto Encyclopedia of Genes and Genomes) annotations of metagenomic contigs as well as detailed mass spectrometry run information are available as Dataset S1 and Dataset S2, respectively.

1To whom correspondence may be addressed. Email: sean.crowe@ubc.ca or shallam@mail.ubc.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1602897113/-/DCSupplemental.
release the catalyzed reactions. Although compelling, these models stop short of incorporating the entire central dogma of molecular biology, and therefore do not achieve a truly quantitative integration of multiomic information and geochemical processes. For example, although such models allowed for a qualitative comparison between modeled gene production rates and selected transcript profiles, they do not provide any explicit mechanistic links (5). The absence of a quantitative validation against process rate measurements or other proxies for activity (e.g., proteins) leaves the adequacy of gene-centric models as a description for microbial activity unsubstantiated.

Here, we construct a gene-centric model for Saanich Inlet, a seasonally anoxic fjord on the coast of Vancouver Island, Canada (7) and a tractable analog for the biogeochemistry of oxygen minimum zones (OMZs). OMZs are widespread and expanding regions of the ocean, in which microbial community metabolism drives coupled carbon, nitrogen, and sulfur cycling (8, 9). These processes exert a disproportionately large influence on global N budgets, with resulting feedbacks on marine primary productivity and climate (10, 11). Extensive time-series data collection in Saanich Inlet provides an opportunity to interrogate biogeochemical processes across defined redox gradients extensible to coastal and open ocean OMZs (12, 13). Our model explicitly describes DNA, mRNA, and protein dynamics as well as biogeochemical reaction rates at ecosystem scales. To calibrate and validate our model, we use geochemical depth profiles; rate measurements for N cycling processes; and metagenomic, metatranscriptomic, and metaproteomic sequence data as well as qPCR-based abundance estimates for SUP05 Gammaproteobacteria—the dominant denitrifiers in Saanich Inlet (14), all obtained from a single location in Saanich Inlet in early or mid-2010.

Construction and Calibration of a Gene-centric Model

A recurring cycle of deep water renewal and stratification in Saanich Inlet results in an annual formation of an oxygen-depleted (O\textsubscript{2}) zone in deep basin waters (Fig. 1B). As this OMZ slowly expands upward during spring, anaerobic carbon remineralization in the underlying sediments leads to an accumulation of ammonium (NH\textsubscript{4}\textsuperscript{+}) and hydrogen sulfide (H\textsubscript{2}S) at depth (15). The oxidation of sulfide using nitrate produced through nitrification in the upper water column fuels chemosynthetic activity and promotes the formation of a narrow sulfide–nitrate transition zone (SNTZ) at intermediate depths (12, 16, 17).

Our model describes the dynamics of several genes involved in carbon, nitrogen, and sulfur cycling across the Saanich Inlet redox gradient. Each gene is a proxy for a particular redox pathway that couples the oxidation of an external electron donor to the reduction of an external electron acceptor. The model builds on a large reservoir of previous work that provides conceptual insight into the microbial metabolic network in Saanich Inlet (12–14, 16–18) and includes aerobic remineralization of organic matter (ROM), nitrification, anaerobic ammonium oxidation (anammox), and denitrification coupled to sulfide oxidation (Fig. 1A). Certain pathways found in other OMZs, such as aerobic sulfide oxidation (19), dissimilatory nitrate reduction to ammonium (DNRA), (20) and sulfate reduction (5), were excluded from our model based on information from previous studies (12–14, 16–18) as well as preliminary tests with model variants (as explained below). Reaction rates (per gene) depend on the concentrations of utilized metabolites according to first- or second-order (Michaelis–Menten) kinetics (21). In turn, the production or depletion of metabolites at any depth is determined by the reaction rates at that depth. The production of genes is driven by the release of energy from their catalyzed reactions and proportional to the Gibbs free energy multiplied by the reaction rate (22).

The model was evaluated at steady state between 100- and 200-m depth. Accordingly, free parameters were calibrated using geochemical profiles obtained in early 2010 during a period of intense water column stratification, which resulted in an extensive anoxic zone that approached a steady state (Fig. 2 and SI Appendix, Section S2.7). After calibration, most residuals to the data are associated with an upward offset of the predicted SNTZ (Fig. 2 B, C, and F) and an underestimation of nitrous oxide (N\textsubscript{2}O) concentrations in deep basin waters (Fig. 2E). These residuals can be explained by subtle deviations from a steady state. Such deviations were revealed in subsequent time-series measurements,

![Fig. 1. Metabolic network and selected time-series data. (A) Coupled carbon, nitrogen, and sulfur redox pathways considered in the model: Aerobic remineralization of organic matter (ROM), aerobic ammonia oxidation (amo), aerobic nitrite oxidation (nxr), and anaerobic ammonia oxidation (hzo), as well as sulfide-driven partial denitrification to nitrous oxide (PDNO) and reduction of nitrous oxide (nosZ) coupled to hydrogen sulfide oxidation. Depth profiles of all shown substrates, except for particulate organic matter (POM), were explicitly modeled; POM concentrations were fixed (more details are in Materials and Methods). Major taxonomic groups encoding specific pathways are indicated. (B) Water column oxygen (O\textsubscript{2}), nitrate (NO\textsubscript{3}-), and hydrogen sulfide (H\textsubscript{2}S) concentrations measured at Saanich Inlet station SI03 from January 2008 to December 2011. The shaded interval and the dates at the bottom indicate the chemical measurements that were used for model calibration. The vertical white line marks the time of molecular sampling. The model considers depths between 100 and 200 m.](image-url)
during which the SNTZ continued to migrate upward in the water column (Fig. 2).

**DNA Profiles and Process Rates**

The calibrated model makes predictions about gene abundance and process rates, which can be validated using metagenomic sequence data and N process rate measurements from the same location and period as the geochemical calibration data (Fig. 3). Consistent with metagenomic depth profiles, the model predicts a redox-driven partitioning of pathways across the water column. Genes associated with ROM (ABC transporters mapped to the SNTZ (Fig. 3)). Consistent with metagenomic depth profiles, the model predicts a redox-driven partitioning of pathways across the water column. Genes associated with ROM (ABC transporters mapped to the SNTZ (Fig. 3)).

The similarity of the PDNO, nosZ, and hzo gene profiles is indicative of their strong metabolic interaction (Fig. 3A). In particular, the co-occurring peaks of PDNO and nosZ gene abundances in the absence of N$_2$O accumulation (Fig. 3E) reflect a quantitative coupling between the two denitrification steps and imply that both steps support extensive microbial growth at the SNTZ. This coupling is intriguing, because genomic reconstructions from both uncultivated and cultivated SUP05, the dominant denitrifier in Saanich Inlet, have not identified the nosZ gene (14, 17, 23). The absence of nosZ from SUP05 suggests that incomplete nitrate reduction by SUP05 and reduction of nitrous oxide by unidentified community members constitute separate and complementary metabolic niches in Saanich Inlet under low-oxygen and anoxic conditions (24).

The superposition of electron donor–acceptor pairs in redox transition zones supports chemical energy transfer in stratified water columns (25, 26), and previous studies have revealed relatively high cell abundances and chemosynthetic activity in such zones (12, 27, 28). At the SNTZ in Saanich Inlet, the simultaneous availability of NO$_3^-$ and H$_2$S fuels chemosynthetic nitrate reduction coupled to sulfide oxidation, in turn supplying anammox with NO$_3^-$ via “leaky” denitrification (up to 88% of NO$_3^-$ supplied by PDNO) (SI Appendix, Section S2.11). Most of the NH$_4^+$ used by anammox (0.3 mmol·m$^{-2}$·d$^{-1}$), however, is predicted to originate from the underlying sediments and reach the SNTZ via eddy diffusion. Accordingly, both anammox and denitrification rates are predicted to peak around the SNTZ and lead to production of N$_2$. This prediction is consistent with process rate measurements from discrete depth intervals during subsequent cruises in 2010 (Fig. 3B) as well as elevated SUP05 abundances at the SNTZ (Fig. 3A) (estimated via qPCR). In fact, the good agreement between predicted PDNO gene counts and observed SUP05 abundances suggests that energy fluxes associated with denitrification can be accurately translated to denitrifier growth rates. Predicted peak sulfide-driven denitrification rates are somewhat higher than peak anammox rates, although depth-integrated nitrogen loss rates are comparable for both pathways (∼0.3 mmol·N$_2$·m$^{-2}$·d$^{-1}$). These predictions are partly consistent with rate estimates derived directly from the geochemical profiles using inverse linear transport modeling (ILTM) (details in Materials and Methods and Fig. 3B). Hence, near steady-state conditions, coupled sulfide-driven denitrification and anammox can concurrently drive significant nitrogen loss in the water column.

The fraction of NO$_2^-$ leaked during denitrification, compared with the total NO$_3^-$ consumed (L$_{PDNO} = 0.352$) (SI Appendix, Section S2.3), was calibrated as a free model parameter based on the observed geochemical profiles. Such a high NO$_2^-$ leakage may result from an optimization of energy yield under electron donor limitation. Additional experimental work is needed to determine the mechanisms controlling this leakage by SUP05. Heterotrophic denitrification and nitrification are conventionally thought of as the primary sources of both nitrite and ammonium for anammox in OMZs (20, 29), and so far evidence for a direct coupling between sulfide-driven denitrification and anammox has been scarce (28). Our results indicate that incomplete sulfide-driven denitrification can be an important precursor for anammox, particularly under conditions of organic carbon limitation (30). This coupling and the benthic supply of ammonium lead to a substantial departure of the fraction of total nitrogen

---

**Fig. 2.** Measured and predicted geochemical profiles. (A) Oxygen, (B) ammonium, (C) nitrate, (D) nitrite, (E) nitrous oxide, and (F) hydrogen sulfide concentration as predicted by the calibrated model at steady state (thick blue curves). Dots indicate data used for the calibration measured during cruise 41 (cr. 41) on January 13, 2010 (SI044_01/13/10; ♦), cr. 42 (SI042_02/10/10; □), and cr. 43 (SI043_03/10/10; ●). Oxygen profiles were not available for cr. 41 and cr. 43; hence, data from cr. 44 (SI044_04/07/10; * ) were used instead. Thin black curves indicate data measured during cr. 47 (SI047_07/07/10), shortly before deep water renewal. Details on data acquisition are in SI Appendix, Section S1.
loss via anammox in Saanich Inlet (∼50%; predicted at steady state) from previous predictions based on labile organic matter stoichiometry (∼28%) (31).

Steady-state gene production rates for chemoautotrophic pathways are predicted to peak around the SNTZ, reaching ∼3.2 × 10^6 genes L^-1 d^-1. This gene production rate corresponds to a dark carbon assimilation (DCA) rate of ∼60 nmol C L^-1 d^-1, assuming a carbon:dry weight ratio of 0.45 (32) and a dry cell mass of m = 5 × 10^-13 g (22). Previously measured peak DCA rates in the Saanich Inlet OMZ reached 2 μmol C L^-1 d^-1 (16), which is significantly higher than the values predicted here. The potential activity of pathways not considered here, such as heterotrophic or mixotrophic inorganic carbon assimilation (33), may explain some of these differences. Moreover, the model assumes steady-state conditions, whereas redox conditions were far from steady state during previous DCA measurements (16). Transient dynamics in Saanich Inlet can exhibit significantly higher nitrogen fluxes and chemoautotrophic activity (34), which might further explain discrepancies between the model and measured DCA rates. Accounting for chemoautotrophic productivity based on oxidant and reductant supply in redox transition zones is generally difficult because of limited knowledge on active pathways, the possibility of cryptic nutrient cycling, and potential lateral substrate intrusions, and discrepancies similar to our study are frequently reported for other OMZs (35–38). Hence, fully accounting for measured DCA rates remains an unresolved problem with important implications for carbon cycling in OMZs.

Previously detected amino acid motifs similar to those found in proteins catalyzing DNRA suggested that SUP05 may also be providing NH\textsubscript{4}^+ to anammox through DNRA (14). DNRA, not included in the model, is known to fuel anammox in anoxic sediments and water columns (20, 39). So far, incubation experiments have not revealed any DNRA activity in Saanich Inlet, and measured ammonium profiles do not indicate a significant ammonium source at or below the SNTZ (Fig. 2B). Nevertheless, DNRA could be active in Saanich Inlet and remain undetected because of rapid ammonium consumption by anammox (39). An extension of the model that included DNRA as an additional pathway, which we calibrated to the same geochemical data (January to March 2010), predicted negligible DNRA rates compared with denitrification and anammox and consistently converged in redox transition zones.

![Molecular and rate profiles](Fig. 3). Molecular and rate profiles. (A) Predicted DNA, mRNA, and protein concentrations for ROM, amo, nxr, norBC, hzo, and nosZ genes (thick curves) compared with corresponding metagenomic, metatranscriptomic, and metaproteomic data (circles; February 10, 2010). The dashed curve under PDNO genes (row 1, column 4) shows concurrent qPCR-based abundance estimates for SUP05, the dominant denitrifier in Saanich Inlet. (B) Denitrification and anammox rates predicted by the model (thick blue curves) compared with rate measurements (circles) during cruise (cr.) 47 (SI047_07/07/10) and cr. 48 (SI048_08/11/10) as well as rates estimated from geochemical concentration profiles using ILTM (SI Appendix, Section S5). The ILTM estimates calibration (calibr.) in columns 3 and 6 is based on the same geochemical data as used for model calibration (Fig. 2).
to the same predictions as the simpler model. These results suggest that DNRA may be absent from the Saanich Inlet water column—at least near steady-state conditions in late spring—and that the hydroxylamine-oxidoreductase homolog encoded by SUP05 plays an alternative role in energy metabolism (17, 24).

DNA concentration profiles of anammox and denitrification genes appear more diffuse and are skewed toward deep basin waters compared with their corresponding rate profiles and the SNTZ (Fig. 3). The model explains this apparent discrepancy based on turbulent diffusion and sinking, which transport genes away from their replication origin. Hence, community composition at any depth is the combined result of local as well as nonlocal population dynamics. Metabolic flexibility encoded in the genomes of microorganisms mediating these processes may also contribute to broader distributions of individual genes than their predicted activity range (30). This disconnect between local metabolic potential and activity needs to be considered when interpreting metagenomic profiles in a functional context, especially in environments with strong redox gradients, such as OMZs (9) or hydrothermal vents (6).

The concentration maxima of anammox and denitrification genes are predicted at shallower depths than measured (Fig. 3A). This observation is consistent with the upward offset of the predicted SNTZ and highlights an important limitation of steady-state models applied to dynamic ecosystems. Indeed, process rate maxima predicted via ILTM at multiple time points continue to move upward beyond the time interval used for model calibration (Fig. 3B). In reality, an electron donor/electron acceptor interface as narrow as predicted by the model would only develop after sufficient time for transport processes and microbial activity to reach a true steady state. Such narrow interfaces do appear in permanently stratified meromictic lakes (40) or the Baltic Sea, where stagnation periods can persist for many years (30).

**mRNA and Protein Profiles**

Metatranscriptomics and metaproteomics present powerful means to assess community metabolic activity—rather than just metabolic potential—and each method comes with its own set of advantages (14, 41). For example, although transcripts represent immediate proxies for gene up-regulation (e.g., in response to changing redox conditions), proteins reflect the immediate catalytic potential of a community, and in vitro characterization of enzyme kinetics can facilitate the projection of protein abundances to in situ process kinetics (42). Transcript abundances need not always correlate strongly with protein abundances (for example, in cases of translational control or protein instability) (42, 43), and hence, metatranscriptomics and metaproteomics provide different perspectives on community activity. Hence, a systematic evaluation of the consistency between these alternative layers of information in real ecosystems is warranted. In fact, a unifying mechanistic model describing the processes that control environmental mRNA and protein distributions is crucial for the correct interpretation of multiomic data in relation to biogeochemical processes (41).

Although DNA replication and process rates are predicted by our gene-centric model merely based on environmental redox conditions, it is uncertain to what extent intermediate stages of gene expression (transcription and translation) can be explained based on such a paradigm. For example, environmental mRNA concentrations measured via qPCR have previously been directly compared with predicted reaction rates (5), but such a heuristic comparison ignores other mechanisms controlling environmental biomolecule distributions, such as physical transport processes. Here, in an attempt to mechanistically describe mRNA and protein dynamics at ecosystem scales, we hypothesized that both mRNA and protein production rates at a particular depth are proportional to the total reaction rate at that depth (calculated using the calibrated model). This premise is motivated by observations of elevated transcription and translation rates during high metabolic activity or growth (44–46). Furthermore, we assumed that mRNA and protein molecules are subject to the same hydrodynamic dispersal processes as DNA, while decaying exponentially with time postsynthesis. The decay time of each molecule as well as the proportionality factor between the reaction rate and synthesis were estimated statistically using metatranscriptomic and metaproteomic data (SI Appendix: Section S2.10).

The general agreement between this model and the molecular data (Fig. 3A) suggests that the production–degradation dynamics of several of these molecules is, at the ecosystem level, dominated by the mechanisms described above. The best fit (in terms of the coefficient of determination) (SI Appendix, Table S5) is achieved for nosZ and nxr mRNA as well as amo, norBC, ROM, and nxr proteins. The greater number of protein over mRNA profiles that can be explained by the model suggests that the proteins considered here are, indeed, simply produced on demand and slowly degrade over time, whereas mRNA dynamics are subject to more complex regulatory mechanisms (41, 47). In particular, the decay times of some transcripts and proteins were estimated to be as high as several weeks (SI Appendix, Table S5). For proteins, these estimates fall within known ranges (47); however, for transcripts, these estimates are much higher than decay times determined experimentally in cells (48). One reason for this discrepancy seems to be the underestimation of the SNTZ depth range by the model, which in turn leads to longer estimates for mRNA decay times needed to explain the detection of these molecules outside of the SNTZ. Alternatively, transcripts and proteins might persist in the cells in inactive states for a significant period, even after dispersal into areas with low substrate concentrations. For example, stable but silent transcripts have been found in bacteria after several days of starvation (49, 50). Furthermore, gene expression may not change immediately in response to external stimuli (14). For example, for some prokaryotic transcription cascades, the basic time unit may be the cell doubling time (which can reach several weeks in anoxic environments) (51) because of regulation by long-lived transcription factors (52). Hence, the decay times estimated here may reflect a hysteresis in gene down-regulation after nutrient depletion, perhaps in anticipation of potential future opportunities for growth (53, 54). Overall, these observations suggest that multiomic sequence data and models for environmental mRNA dynamics would benefit from a consideration of additional control mechanisms (for example, derived from cell-centric transcription models) (55, 56).

**Consequences for Geobiology**

Gene-centric models have the potential to integrate biogeochemical processes with microbial population dynamics (5, 6). According to the central dogma of molecular biology (4), gene transcription and translation are intermediate steps that regulate metabolic processes in individual cells, but the appropriate projection of the central dogma to ecosystem scales remained unresolved, because transcription and translation were not explicitly considered in previous models (5, 6). We have developed a biogeochemical model that explicitly incorporates multiomic sequence information and predicts pathway expression and growth in relation to geochemical conditions. In particular, when mRNA and protein dynamics are omitted, the gene-centric model only includes four calibrated parameters and yet is able to largely reproduce geochemical profiles (Fig. 2), relative genomic profiles (Fig. 3A), and SUP05 cell abundances, indicating that the good agreement between the model and the data is unlikely caused by overfitting. In fact, as we refined and calibrated our model to the geochemical profiles, we observed that the metagenomic profiles were well-reproduced as soon as the model’s geochemical predictions roughly aligned with the data, even if the parameters being calibrated had not converged.
yet. This observation reinforces the interpretation that fluxes of matter and energy are robust predictors of microbial productivity and functional community structure.

Our model successfully explains a large fraction of environmental mRNA and protein distributions based on DNA concentration profiles and biochemical reaction rates in the OMZ. These results are consistent with the idea that DNA is a robust descriptor of an ecosystem’s biotic component (57, 58), which in conjunction with the geochemical background, determines pathway expression and process rates (59). This idea implies that the co-occurrence of a metabolic niche with cells able to exploit it is sufficient to predict microbial activity. Under this paradigm, DNA may be regarded as directly interacting with the geochemical background, whereas the production of mRNA and proteins is an inevitable consequence of the biologically catalyzed flow of matter and energy. This interpretation is supported by the overall consistency between the metatranscriptomic and metaproteomic profiles for N and S cycling genes (Fig. 34). Hence, mRNA and proteins may each be adequate proxies for pathway activity in Saanich Inlet, although as discussed above, questions regarding the relevant time scales remain. Additional work is needed to test this paradigm in other ecosystems, especially under non-steady-state conditions. Nevertheless, many aquatic ecosystems are permanently or almost permanently anoxic (8, 40, 60), and hence, our approach and conclusions are expected to be particularly applicable to these systems.

In addition to providing a systematic calibration and validation of the model, we identified processes that need to be considered when interpreting multiomic data. Conventional proxies for activity, such as mRNA and proteins, are themselves subject to complex population dynamics that include production and active or passive degradation as well as physical transport processes. Consequently, the close association between process rates and biomolecule production suggested above does not imply that biomolecule distributions, per se, are equivalent to local microbial activity rates. In Saanich Inlet, for example, the widespread distribution of DNA, mRNA, and proteins across the OMZ, in contrast to a relatively narrow metabolic “hot zone” at the SNTZ, is predicated on a balance between spatially confined production and dispersal across the water column. This “mass effect” (61) means that geochemical or biochemical information is needed to assign actual activity to genes or pathways identified in multiomic data, especially for components mediating energy metabolism. This conclusion is generalizable and should be applied to other ecosystems exhibiting dispersal across strong environmental gradients, such as estuaries (62) or hydrothermal vents (6). Moreover, in dynamic ecosystems with rapidly changing geochemical conditions, past population growth rates can influence future community structure and biomolecular patterns, and hence cross-sectional community profiles may not reflect current dynamics (63). In such systems, an incorporation of multiple layers of geochemical and biological information into a mechanistic model—as shown here—will be crucial for disentangling the multitude of processes underlying experimental observations.

The gene-centric model constructed here, although evaluated at steady state, is in fact a spatiotemporal model that could, in principle, predict gene population dynamics and process rates over time. A spatiotemporal analysis of the model would require multiomic time series coverage and knowledge of nonstationary physical processes, such as seasonal patterns in surface productivity and hydrodynamics during deep water renewal events. Multiomic time series would be especially useful for improving the mRNA and protein models introduced here because of the high number of currently unknown parameters. For example, integrating metatranscriptomic, metaproteomic, and geochemical time series during rapid environmental changes into our model would allow for a more direct determination of in situ transcriptional and translational responses and biomolecule decay times.

The multiomic profiles that we used to validate our model are only given in terms of relative—rather than absolute—biomolecule concentrations. Hence, the observed abundance of each biomolecule may be influenced by the abundances of other biomolecules, which could explain some of the discrepancies between the model and the multiomic data. Unfortunately, this limitation is currently pervasive across environmental shotgun sequencing studies, largely because of technical challenges in estimating in situ DNA, mRNA, and protein concentrations (64). These challenges will likely be overcome in the future (65, 66). Given this current caveat, the general agreement of the model with the shape of the multiomic profiles (Fig. 34) is remarkable and suggests that the spatial structuring of the metabolic network is well-captured by the model. In fact, our qPCR-based estimates for absolute SUP05 abundances are consistent with absolute PDNO gene concentrations predicted by the model as well as the shape of the PDNO abundance profiles in the metagenomes (Fig. 34). This double agreement suggests that—at least for PDNO—both our model and our metagenomic datasets (Datasets S1 and S2) reflect the actual gene distributions.

Conclusions
Most major metabolic pathways driving global biogeochemical cycles are encoded by a core set of genes, many of which are distributed across distant microbial clades (1). These genes are expressed and proliferate in response to specific redox conditions and in turn, shape Earth’s surface chemistry. Here, we have shown that the population dynamics of genes representative of specific metabolic pathways, their expression, and their catalytic activity at ecosystem scales can all be integrated into a mechanistic framework for understanding coupled carbon, nitrogen, and sulfur cycling in OMZs. This framework largely explained DNA, mRNA, and protein concentration profiles and resolved several previous uncertainties in metabolic network structure in Saanich Inlet, including a direct coupling of sulfide-driven denitrification and anammox through leaky nitrite production by SUP05 as well as the presence of a metabolic niche for nitrous oxide reduction contributing to nitrogen loss. Beyond OMZ-specific predictions, model results indicate that geochemical fluxes are robust indicators of microbial community structure and reciprocally, that gene abundances and geochemical conditions largely determine gene expression patterns. Such integrated modeling approaches offer insight into microbial community metabolic networks and allow prediction of elemental cycling in a changing world.

Materials and Methods

Model Overview. The core model is a set of differential equations for the concentrations of eight metabolites and six proxy genes (DNA) across depth (100–200 m) and time. Each gene is a proxy for a particular energy-yielding pathway, which couples the oxidation of an external electron donor to the reduction of an external electron acceptor. Each gene is considered as a replicating unit that is independent of other genes. This simplifying assumption corresponds to the case where each cell occupies a single metabolic niche associated with one of the modeled pathways (14, 67). Gene-specific reaction rates depend on the concentrations of metabolites according to first- or second-order (Michaelis– Menten) kinetics (5, 21) (SI Appendix, Section S2.4). In turn, the production or depletion of metabolites at any depth is determined by the reaction rates at that depth, taking into account reaction stoichiometry (SI Appendix, Section S2.3) and diffusive transport across the water column. The production of genes at any depth is driven by the release of energy from their catalyzed reactions and is proportional to the Gibbs free energy multiplied by the reaction rate (22) (SI Appendix, Section S2.5). In addition, gene populations are subject to exponential decay rates, diffusive transport, and sinking.
Mathematical Model Structure. The DNA concentration for gene \( r \) (\( Z_r \) copies per volume) exhibits the dynamics

\[
\frac{dZ_r}{dt} = -\frac{q_r}{\tau_r} Z_r + \frac{1}{\tau_c} \left( L_r C_{H_2} dR_r - \frac{\partial}{\partial z} \left( \frac{\partial Z_r}{\partial z} \right) \right). \tag{1}
\]

whereas the concentration of the mth metabolite (\( C_m \) mole per volume) follows

\[
\frac{dC_m}{dt} = \sum S_m H_r \frac{\partial}{\partial z} + \frac{\partial}{\partial z} \left( K C_m \frac{\partial C_m}{\partial z} \right). \tag{2}
\]

Both the gene concentrations \( \Gamma_r \) and metabolite concentrations \( C_m \) depend on time \( t \) and depth \( z \). The first term on the right-hand side of Eq. 1 corresponds to cell death, with \( q_r \) being the exponential death rate in the absence of any metabolic activity for pathway \( r \). The second term corresponds to gene production, with \( H_r \) being the per-gene reaction rate as a function of metabolite concentrations (SI Appendix, Section S2.4). The biomass production coefficient \( Z_r \) is a linear function of the Gibb's free energy of reaction \( r \) (SI Appendix, Section S2.5). \( \alpha \) is the average dry cell mass, which is used to convert biomass production into cell production. The third term corresponds to cell sinking at speed \( v \). The last term in Eqs. 1 and 2 corresponds to diffusive transport, with \( K \) being the vertical eddy diffusion coefficient. In Eq. 2, \( S_m \) is the stoichiometric coefficient of metabolite \( m \) in reaction \( r \) (SI Appendix, Section S2.3). The sum on the right-hand side of Eq. 2 iterates through all reactions and accounts for microbial metabolic fluxes. Eqs. 1 and 2 specify the rates of change for the DNA and chemical concentration profiles. Steady-state profiles were obtained after long simulations when all profiles had eventually stabilized.

Considered Pathways. Redox pathways occurring in a single cell require at least two enzymes: one involved in the oxidation of the initial electron donor and one involved in the reduction of the final electron acceptor. In the model, such pathways are represented by single proxy genes, chosen such that their functional role is minimized. For example, nitrous oxide reduction using nosZ coupled to sulfide oxidation is identified with nosZ, and one involved in the reduction of the final electron acceptor. In the model, the first reaction (2) iterates through all reactions and accounts for microbial metabolic fluxes. Eqs. 1 and 2 specify the rates of change for the DNA and chemical concentration profiles. Steady-state profiles were obtained after long simulations when all profiles had eventually stabilized.

Model Calibration and Data. Unknown parameters of the basic gene-centric model (Eqs. 1 and 2) (ignoring mRNA and protein dynamics) were calibrated by comparing steady-state predictions with measured depth profiles of oxygen, ammonium, nitrate, nitrite, hydrogen sulfide, and nitrous oxide. Chemical calibration data were acquired on January 13, February 10, and March 10, 2010 (or February 10 and April 7 for oxygen) from a single location in Saanich Inlet (123° 30.30’ W, 48° 35.50’ N) (SI Appendix, Section S1.2). The calibrated parameters were the maximum cell-specific reaction rate \( V_{P_{rmax}} \) and the first-order rate constants \( A_{rmax} \) as well as the PDNO leakage fraction \( \lambda_{PDNO} \) (SI Appendix, Table S3). Calibration was performed by maximizing the likelihood of a statistical model, in which the deterministic part (i.e., expectation) is given by the predictions of the gene-centric model and the stochastic part (i.e., error) is normally distributed (SI Appendix, Section S2.8). This calibration method is known as maximum likelihood estimation and is widespread in statistical regression and physics (70). Maximization of the likelihood was performed using the interior point method using its open-source and parallelized implementation. The sensitivity of the model to parameter variation was assessed via local sensitivity analysis (72) as described in SI Appendix, Section S2.12. An overview of our workflow is shown in SI Appendix, Fig. S2. Samples for molecular sequencing were collected on February 10, 2010 from the same geographical location as the geochemical data (SI Appendix, Sections S1.3 and S1.4). Because metaproteomes were missing at depth 100 m (the upper bound of our simulation domain) and to increase statistical power when evaluating our protein models, we used linear interpolation between depths 97 and 120 m (where metaproteomes were available) to estimate protein normalized spectral abundance factor (NSAF) values at 100-m depth (“unit imputation”). Metagenomic profiles (a priori in relative units) were rescaled to match the model scales using maximum likelihood estimated factors (SI Appendix, Section S2.9). SUP05 abundances for February 10, 2010 were estimated via qPCR using SUP05-specific primers targeting the S19–1,048 region of the SUP05 16S rRNA gene following the protocol in ref. 73; 16S gene counts were corrected for the number of 16S RNA gene copies per cell estimated using the Tax4Fun pipeline (74) (SI Appendix, Section S1.6). Denitrification and anammox rates were measured on cruises 47 (SI047_0707/10) and 48 (SI048_0811/10) via ex situ incubation experiments and subsequently corrected for differences between in situ and incubated substrate concentrations (SI Appendix, Section S1.5).

mRNA and Protein Models. As mentioned previously, following calibration of the gene-centric model to the geochemical profiles, we extended the model to describe mRNA (and similarly, protein) dynamics in the water column. Specifically, the production rate of an mRNA (transcripts produced per time and volume of seawater) at a particular depth was assumed to be proportional to the total reaction rate (molecules per time and volume of seawater) at that depth. A linear relation, although only an approximation, can be justified by the fact that increased enzyme dilution rates at elevated cell division rates must be balanced (at the population level) by correspondingly increased translation—and hence, transcription—rates (55). We also assumed that mRNA molecules disperse via diffusion and sinking similar to genes (because they are hosted by the same cells) and decay exponentially with time. Thus, environmental mRNA concentrations satisfy the partial differential equation

\[
\frac{dT_r}{dt} + \frac{\partial T_r}{\partial z} + \frac{\partial}{\partial z} \left( \frac{\partial T_r}{\partial z} \right) = \frac{\partial}{\partial z} \left( K \frac{\partial T_r}{\partial z} \right) \tag{3}
\]

where \( T_r \) is the mRNA concentration corresponding to the rth reaction, \( \tau_r \) is the decay time of the mRNA molecule, \( R_r = H_r \Gamma_r \) is the total reaction rate, and \( \alpha_r \) is an unknown proportionality constant. We considered \( T_r \) in the same units as the multicomponent data, i.e., reads per kilobase per million mapped reads (RPKM) for metatranscriptomes and NSAF for metaproteomes. Consequently, \( \alpha_r \) is the ratio between the rth reaction rate (molecules·day\(^{-1}\)·liter\(^{-1}\)) and the corresponding RPKM (or NSAF) “production rate” (RPKM·day\(^{-1}\)). HSD thus depends closely not only on the particular reaction but also on our sampling protocol and sequencing pipeline. The above model was evaluated at steady state, when mRNA production, dispersal, and decay are balanced at each depth (\( \partial T_r/\partial t = 0 \)). The parameters of the mRNA and protein models


21. van de Graaf P, et al. (2007) Determination of the intramural research and development program of the W. R. Wiley Environmental Molecular Sciences Laboratory (EMSL). The EMSL is a national scientific user facility sponsored by the US DOE Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory operated by Battelle for the US DOE. S.L. was supported by the Pacific Institute for the Mathematical Sciences (International Graduate Training Centre for Mathematical Biology), as well as the Department of Mathematics, University of British Columbia. M.P.B. received support from the Canadian Institute for Advanced Research Global Fellowship in the Integrated Microbial Biodiversity Program. S.L., M.P.B., and M.D. also received support from NSERC. G.L. was supported by the Max Planck Society.

22. Perez, Ariah N., and Natasha Shiota for comments on the manuscript. We also thank the Joint Genome Institute, including Susannah Tringe, Stephanie Wasilenko, and Tajana Glaser for technical assistance. This work was performed under the auspices of the US Department of Energy (DOE) Joint Genome Institute, supported by US DOE Office of Science Contract DE-AC02-05CH11231; the G. Unger Vetlesen and Ambrose Monell Foundations; the Tula Foundation-funded Centre for Microbial Diversity and Evolution; the Natural Sciences and Engineering Research Council of Canada (NSERC); Genome British Columbia; the Canada Foundation for Innovation; and grants from the Canadian Institute for Advanced Research (to S.A.C. and S.J.H.). Metaproteomics support came from the intramural research and development program of the W. R. Wiley Environmental Molecular Sciences Laboratory (EMSL). The EMSL is a national scientific user facility sponsored by the US DOE Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory operated by Battelle for the US DOE. S.L. was supported by the Pacific Institute for the Mathematical Sciences (International Graduate Training Centre for Mathematical Biology), as well as the Department of Mathematics, University of British Columbia. M.P.B. received support from the Canadian Institute for Advanced Research Global Fellowship in the Integrated Microbial Biodiversity Program. S.L., M.P.B., and M.D. also received support from NSERC. G.L. was supported by the Max Planck Society.
Integrating biogeochemistry with multi-omic sequence information in a model oxygen minimum zone
- SI Appendix -

Stilianos Louca, Alyse K. Hawley, Sergei Katsev, Monica Torres-Beltran, Maya P. Bhatia, Sam Kheirandish, Céline C. Michiels, David Capelle, Gaute Lavik, Michael Doebeli, Sean A. Crowe, Steven J. Hallam

S1 Data acquisition

S1.1 Sampling site and time

Saanich Inlet (SI) a seasonally anoxic fjord on the coast of Vancouver Island, British Columbia Canada has been the site of intensive study for many decades (1, 2). The presence of a shallow glacial entrance sill at 75 m depth limits mixing and ventilation of basin waters below approximately 100 m, resulting in stratification and oxygen depletion during spring and summer (Fig. 1a). Shifts in coastal currents in late summer and fall lead to an influx of denser, oxygenated and nutrient-rich water into the Inlet shoaling anoxic basin waters upward in a process known as deep water renewal (2, 3). Consistent partitioning of the microbial community along the redox cline and similarity to other OMZ microbial communities make Saanich Inlet a model ecosystem for studying the intersection between environmental sequence information and biogeochemical activity along defined redox gradients (3–5).

The fjord has a maximal depth of 232 m at the sampling site SI03 (123° 30.300’ W, 48° 35.500’ N).

Sampling is conducted monthly during daylight hours using a combination of 5 and 8 L Niskin bottles and 12 L Go-Flo bottles attached to a nonconducting wire. A Sea-Bird CTD (conductivity, temperature and depth) sensor attached to the bottom of the wire provides depth profiles for temperature, salinity, PAR/Irradiance, conductivity, density, and dissolved oxygen (Sea-Bird Electronics™). Water sampling for multiple chemical and microbial parameters proceeds directly from the bottles in the following order: First, samples are taken for dissolved O₂ measurements via Winkler titration, followed by sampling of dissolved gases. Next, samples are taken for RNA, then protein followed by ammonium, hydrogen sulfide and nitrite. Finally, salinity is measured for a subset of depths for CTD calibration, and samples are taken for DNA.

Chemical data were acquired on January 13, 2010 (cruise SI041_01/13/10), February 10 (SI042_02/10/10), March 10 (SI043_03/10/10), April 7 (SI044_04/07/10), July 7 (SI047_07/07/10)
and August 11 (SI048_08/11/10). All molecular sequencing was performed using samples collected on February 10, 2010 (SI040_02/08/10) at depths 100 m, 120 m, 135 m, 150 m, 200 m for metagenomes and metatranscriptomes and at depths 97 m, 100 m, 120 m, 150 m, 165 m, 200 m for metaproteomes.

S1.2 Chemical and physical depth profiles

Temperature, salinity and depth were measured using the CTD sensor described above. The Winkler titration method was used to measure dissolved oxygen (O$_2$) concentrations (6) and calibrate CTD measurements. Samples were collected into Winkler glass Erlenmeyer flasks using latex tubing, overflowing three times to ensure no air contamination, manganese (III) sulphate and potassium iodide were added in succession, inverted to mix and stored at room temperature. Samples were titrated using an automatic titrator. CTD data were processed and manually curated using the Sea-Bird Seasoft™ software.

Samples for dissolved nutrient (nitrate, nitrite, sulphate and silicate) analyses were collected into 60 mL syringes and filtered through a 0.22 μm Millipore Acrodisc™ into 15 mL falcon tubes. Prior to analysis all samples were stored on ice. Nitrate (NO$_3^-$) samples were stored at -20°C in the laboratory, and later analyses carried out using a Bran Luebbe autoanalyser using standard colorimetric methods. For nitrite (NO$_2^-$) analysis, 2 mL of sample water were supplemented with 100 μL sulfanilamide and 100 μL nicotinamide adenine dinucleotide in 4 mL plastic cuvettes. Prepared standards were supplemented with reagents at the same time. Cuvettes were inverted for mixing, and temporarily stored on ice for not more than 4 hrs. Concentration was measured using a Cary60® spectrometer, based on absorbance at 452 nm.

Samples for ammonium (NH$_4^+$) and hydrogen sulphide (H$_2$S) were collected directly from Niskin and GoFlo bottles into 15 mL amber scintillation vials and 15 mL falcon tubes aliquoted with 200 μL 20% zinc acetate respectively. Samples were stored on ice prior to analysis. For NH$_4^+$ analysis, amber vials for standard curve and samples were pre-aliquoted with 7.5 mL O-phthaldialdehyde (OPA) reagent respectively. 5 mL of sample water in triplicate and standard solutions were transferred into OPA pre-aliquoted amber vials. Vials were inverted and stored up to 4 hours. From each standard solution and sample water vial, 300 μL were transferred into a 96 well round bottom plate. Fluorescence at 380 ex/420 emm was read using a Varioskan plate reader. For H$_2$S analysis, 300 μL samples were transferred in triplicate to a 96 well plate, and finally Hach Reagent 1 and 2 (6 μL per well) were added. Absorbance at 670 nm was read after 5 min incubation using a Varioskan™ plate reader.

Water for dissolved nitrous oxide (N$_2$O) analysis was collected using Go-flo or Niskin bottles, and was transferred via a Teflon tube into 30 mL or 60 mL borosilicate glass serum vials. Vials were overflown three times their volume in order to remove any bubbles from the vial or tubing. Vials were subsequently spiked with 50 μL saturated mercuric chloride using a pipette. Vials were then crimp-sealed with a butyl-rubber stopper and aluminum cap, and stored in the dark at 4°C until they were analyzed. Dissolved nitrous-oxide concentrations were measured using a purge-and-trap
auto-sampler coupled with a gas-chromatography mass-spectrometer (7).

S1.3 Metagenomics, metatranscriptomics and metaproteomics

Metagenome and metaproteome datasets were generated using the same methods as described in Hawley et al. (8). Metaproteome sequence coverage was quantified using normalized spectral abundance factors (NSAF) (9). Metatranscriptome samples were filtered in the field onto 0.2 \( \mu \text{m} \) Sterivex filter with inline pre-filter of 2.7 \( \mu \text{m} \) pre-filter, adding 1.8 mL of RNAlater\textsuperscript{®} (Qiagen) and freezing on dry ice before transferring to \(-80^\circ\text{C}\). RNA later was removed by washing Sterivex filter with Ringer’s solution before proceeding with cell lysis in the filter cartridge. Total RNA was extracted using the mirVana\textsuperscript{TM} miRNA extraction kit (Ambion), DNA was removed using the TURBO DNA-free\textsuperscript{TM} kit (Ambion) and total RNA was purified using RNeasy\textsuperscript{TM} MiniElute Cleanup Kit (Qiagen). RNA concentration and quality was determined using a Bioanalyzer. Production of cDNA libraries and sequencing was carried out at the Joint Genome Institute using the TruSeq\textsuperscript{®} Stranded Total RNA Sample preparation Guide, including depletion of ribosomal RNA using Ribo-Zero. Assembled metagenomic sequences (contigs) were run through Metapathways (10) for annotation using a combination of RefSeq (11), KEGG (12), COG (13) and MetaCyc (14) databases. KEGG annotations of metagenomic contigs are provided as Dataset S1. Mass-spectrometry (metaproteomics) run information are provided as Dataset S2. Contig coverage by metagenome or metatranscriptome reads was quantified using RPKM values (Appendix S1.4). KEGG-annotated contigs were assigned to the selected process proxy genes of the model (Table 1.3 in the Appendix); gene coverage at each depth was then quantified by summing all assigned contig RPKM values. Because metaproteomes were missing at depth 100 m (the upper bound of our simulation domain), and in order to increase statistical power when evaluating our protein models, we used linear interpolation between depths 97 m and 120 m to estimate protein NSAF values at 100 m depth (“unit imputation”).

Nitrate reductase (narGHIJ) assigned to planctomycetes showed a decline with depth, suggesting that it may be acting in reverse as a nitrite oxidase (15). In fact, narGHIJ counts affiliated with planctomycetes (narGHIJ-P) dominated all other nxr-associated counts in the metagenomes, metatranscriptomes and metaproteomes. We thus associated nxr with narGHIJ-P. However, because planctomycetes perform anammox in deeper depths (16), we observed a secondary peak in the narGHIJ-P DNA closer to the SNTZ that did not dissipate completely in bottom waters. Given this ambiguity in the interpretation of detected narGHIJ-P genes, we omitted the narGHIJ-P metagenomes and only used the narGHIJ-P metatranscriptomes and metaproteomes. For more details see Appendix S3.4.

All nosZ-related protein sequences mapped to a nosZ homolog found in the strictly aerobic Roseobacter Maritimibacter alkaliphilus HTCC2654 (17, 18) and showed strong inconsistencies with nosZ metagenomic and metatranscriptomic profiles. nosZ genes have been found to be enriched on particles, likely because they constitute a more anaerobic niche (19). Our metaproteomes were pre-filtered to remove eukaryotes and particles and are expected to be impoverished in nosZ proteins, facilitating a potential masking by related but functionally different proteins. We thus omitted the nosZ metaproteomic data from our analysis.
Table S1: KEGG orthologous groups (KOG) identified with each gene in the metagenomes or metatranscriptomes. The abundance of each gene was the sum of RPKM values (Appendix S1.4) assigned to all included KOGs.

<table>
<thead>
<tr>
<th>gene or pathway</th>
<th>KOGs</th>
<th>restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>K12536, K05648</td>
<td>ABC transporters in <em>Pelagibacter</em> and <em>Roseobacter</em></td>
</tr>
<tr>
<td>amo</td>
<td>K10945, K10946</td>
<td></td>
</tr>
<tr>
<td>nxr</td>
<td>K00370, K00371, K00374, K00373</td>
<td>narGHIJ in <em>Planctomycetacea</em></td>
</tr>
<tr>
<td>hzo</td>
<td>K10535</td>
<td>hao in <em>Planctomycetacea</em></td>
</tr>
<tr>
<td>PDNO (norBC)</td>
<td>K04561, K02305</td>
<td></td>
</tr>
<tr>
<td>nosZ</td>
<td>K00376</td>
<td></td>
</tr>
<tr>
<td>sat</td>
<td>K00958</td>
<td></td>
</tr>
<tr>
<td>aprAB</td>
<td>K00394, K00395</td>
<td></td>
</tr>
<tr>
<td>dsrAB</td>
<td>K11180, K11181</td>
<td></td>
</tr>
<tr>
<td>nirKS</td>
<td>K00368, K15864</td>
<td></td>
</tr>
<tr>
<td>napAB</td>
<td>K02567, K02568</td>
<td></td>
</tr>
<tr>
<td>narGHIJ</td>
<td>K00370, K00371, K00374, K00373</td>
<td></td>
</tr>
</tbody>
</table>

S1.4 Quantifying metagenomic and metatranscriptomic data using RPKM

Relative open reading frame (ORF) abundance in the metagenomic and metatranscriptomic datasets was determined for quantitative assessment of pathway coverage. This was achieved by adapting the reads per kilobase per million mapped (RPKM) coverage measure as described by Konwar *et al.* (20). Briefly, unassembled Illumina reads were mapped to assembled contigs using the short-read aligner BWA-MEM. The resulting SAM file is then inputed into the MetaPathways v2.5 software (20), which generates an RPKM value per ORF that is extended to an RPKM per pathway via summation. For the case of determining the abundance of pathways expressed in the metatranscriptome relative to those present in the metagenome, the unassembled metatranscriptome reads were mapped back to the assembled metagenome contigs. The RPKM calculation is a simple proportion of the number of reads mapped to a particular section of sequence normalized for ORF length and sequencing depth.

S1.5 Process rate measurements

Rate measurements for anammox and denitrification were carried out as follows: Sample water from each depth was collected anaerobically with sterile nitrile tubing directly into 200 mL glass serum bottles, six per depth, and capped with butyl-rubber stopper and aluminum cap and stored at 10°C for approximately 1 hr while collection was completed. The protocol described by Holtappels *et al.* (21), and briefly outlined here, was then followed. One sample from each depth was bubbled with
He for 30 min to decrease concentration of N₂. The following substrates were then added: \(^{15}\text{NH}_4^+\) alone, \(^{15}\text{NH}_4^+\) and \(^{14}\text{NO}_2^-\) combined, \(^{15}\text{NO}_2^-\) alone, \(^{15}\text{NO}_2^-\) and \(^{14}\text{NH}_4^+\) combined or \(^{15}\text{NO}_2^-\) alone. A blank for each depth was also bubbled with He. Sample water was then transferred from the serum bottle into a 12 mL exetainer, capped and stored upside down. Samples in exetainers were then killed with 50 µL saturated HgCl₂ at time intervals of 0 min, 6 hr, 12 hr, 24 hr, 48 hr and 72 hrs. Partial pressures of \(^{29}\text{N}_2\) and \(^{30}\text{N}_2\) evolved during the incubations were measured by gas chromatography coupled to isotope ratio mass spectroscopy. Rates of anammox and denitrification were calculated as described by Holtappels et al. (21).

Rate measurements using N isotope methods require a compromise between ensuring detection of labeled tracer elements and avoiding excessive perturbation of ambient substrate concentrations (22, §2.1). Due to the extremely low in-situ substrate levels in some of our samples (Fig. 2 in the main article), tracer substrate concentrations in the ex-situ incubator (25 µM \(^{14}\text{NH}_4^+\), 2 µM \(^{15}\text{NO}_2^-\) and 5 µM \(^{15}\text{NO}_2^-\)) significantly exceeded in-situ concentrations. On the other hand, denitrification and anammox-related genes were found throughout the OMZ water column (Fig. 3a in the main article). Hence, rates measured in the incubator are only potential rates that likely overestimate actual in-situ rates, especially in substrate-depleted regions far from the SNTZ. For example, Dalsgaard et al. (23) reports a 2–4 fold increase of anammox rates following the addition of 10 µM \(^{14}\text{NH}_4^+\) in anoxic water column experiments. Similarly, Wenk et al. (24) found high potential denitrification rates in nitrate-depleted regions of a meromictic lake. We thus corrected our rate measurements for differences between in-situ and incubator substrate concentrations, as described below.

The simplest approach would be to multiply measured rates with the ratios of in-situ over ex-situ substrate concentrations, as has been done in previous ex-situ incubation experiments (25). However, such a linear rescaling implicitly assumes that substrate half-saturation constants are much higher than both the in-situ as well as ex-situ concentrations, an assumption that may not be justifiable in regularly substrate-depleted natural environments. For example, members of the Scalindua candidate clade, which is well represented in Saanich Inlet (16), exhibit nitrite half-saturation constants as low as 0.45 µM (26). To avoid an implicit assumption of 1st order kinetics, and for consistency with the assumptions of our model, we corrected our rates using Michaelis-Menten kinetic curves (Appendix S2.4) with the same half-saturation constants as used in our model (Appendix S2.7). Specifically, if \(R_{\text{hzo}}^*(z)\) is the measured ex-situ (i.e. potential) anammox rate at some particular depth, then the corrected in-situ rate was assumed to be

\[
R_{\text{hzo}} = R_{\text{hzo}}^*(z) \cdot \frac{[\text{NH}_4^+]}{[\text{NH}_4^+]^*} \frac{[\text{NO}_2^-]}{[\text{NO}_2^-]^*} \frac{K_{\text{NH}_4^+}[\text{NH}_4^+] + K_{\text{NO}_2^-}[\text{NO}_2^-]}{K_{\text{NH}_4^+}[\text{NH}_4^+]^* + K_{\text{NO}_2^-}[\text{NO}_2^-]^*}.
\]

Here, \(K_{\text{NH}_4^+}\) and \(K_{\text{NO}_2^-}\) are anammox half-saturation constants for \(^{14}\text{NH}_4^+\) and \(^{15}\text{NO}_2^-\), respectively (Appendix S2.7), \([\text{NH}_4^+]\) and \([\text{NO}_2^-]\) are the corresponding measured in-situ concentrations and \([\text{NH}_4^+]^*\) and \([\text{NO}_2^-]^*\) are the concentrations in the incubator at the beginning of the experiment, i.e. \([\text{NH}_4^+]^* = [\text{NH}_4^+] + 25 \mu\text{M}\) and \([\text{NO}_2^-]^* = [\text{NO}_2^-] + 2 \mu\text{M}\). Measured denitrification rates were corrected in a similar way to account for differences in \(^{15}\text{NO}_3^-\) concentrations.
S1.6 qPCR quantification of SUP05 abundances

All metagenomic, metatranscriptomic and metaproteomic profiles presented here only provide relative — rather than absolute — biomolecule abundances. This remains the de facto standard for multi-omic data sets, owing largely due to methodological challenges involved in absolute DNA, mRNA and protein quantification (but see Smets et al. (27) for recent advancements). As we explain below (section 2.9), multi-omic depth profiles were linearly rescaled to facilitate comparison with our model predictions — expressed in absolute gene counts, however this comes at the cost of additional rescaling parameters.

In order to perform an independent validation of modeled gene concentrations, we compared the predicted PDNO gene concentrations to independent cell-count estimates for SUP05 (the dominant nitrate reducer in Saanich Inlet; 16), obtained through quantitative polymerase chain reaction (qPCR). qPCR quantification of SUP05 abundances was performed for water samples collected at 8 distinct depths from the same location and time as for multi-omic sequencing (Fig. 3a in the main article). Water samples (volume ∼ 1 L) were filtered in the field onto 0.2 µm sterilvex filters. Samples were not pre-filtered in order to obtain an accurate estimate of total in-situ SUP05 abundances. We used a custom SUP05-specific primer set (Ba519F–1048R) to amplify the 519–1048 region of the SUP05 16S rRNA gene, and followed the protocol described by Hawley et al. (8) to estimate the starting template concentration. qPCR was performed in triplicate for each sample. We multiplied the average template concentration for each sample by the volume of extracted fluid (∼ 200 – 400 µL), divided by the volume of filtered seawater, to obtain an estimate for the concentration of SUP05 16S gene copies in seawater. To correct for multiple 16S gene copies in single cells, we divided this concentration by the 16S gene copy number (3.767), estimated for members of the SUP05 clade based on closely related fully sequenced reference genomes. Specifically, we used the 16S gene copy number assigned by the Tax4Fun pipeline (28) to the clade “Oceanospirillales;SUP05 cluster;uncultured gamma proteobacterium” in the SILVA 123 database (29). Note that Tax4Fun (28) uses a probabilistic model to assign multiple reference genomes with varying weights to each clade in the SILVA database. Hence, the effective 16S gene copy number assigned by Tax4Fun to each clade is the weighted harmonic mean of the 16S gene copy numbers in each reference genome assigned to that clade.

S2 Mathematical model

S2.1 Overview

The gene-centric model describes the spatiotemporal dynamics of 8 metabolite concentrations and 6 gene (DNA) concentrations along the Saanich Inlet water column between depths 100–200 m. Each gene is a proxy for a particular redox pathway that couples the oxidation of an external electron donor to the reduction of an external electron acceptor (Appendix S2.2). The model assumes that each cell occupies a single metabolic niche, associated with one of the modeled pathways and thus
one of the considered proxy genes. Reaction rates (per gene) depend on the concentrations of all used metabolites according to 1st order or 2nd order (Michaelis-Menten) kinetics (30, 31) (Appendix S2.4). In turn, the production or depletion of metabolites at any depth is determined by the reaction rates at that depth, taking into account reaction stoichiometry (Appendix S2.3). The production of genes (or more precisely, their host cells) at any depth is driven by the release of energy from their catalyzed reactions, and is proportional to the Gibbs free energy multiplied by the reaction rate (Appendix S2.5) (32). In addition, genes are subject to exponential decay as well as eddy-diffusion and sinking. Metabolites are also subject to eddy-diffusion.

Mathematically, the model is defined as a set of partial differential equations (PDE) for the gene and metabolite concentrations across time and depth. More precisely, the DNA concentration of the r-th gene \( \Gamma_r \), copies per volume) at any a given depth \( z \) changes according to

\[
\frac{\partial \Gamma_r}{\partial t} = - q_r \Gamma_r + \frac{1}{c} Z_r H_r \Gamma_r - v \frac{\partial \Gamma_r}{\partial z} + \frac{\partial}{\partial z} \left( K(z) \frac{\partial}{\partial z} \Gamma_r \right),
\]

(2)

and the concentration of the m-th metabolite (\( C_m \), mole per volume) changes according to

\[
\frac{\partial C_m}{\partial t} = \sum_r m_r H_r \Gamma_r + \frac{\partial}{\partial z} \left( K(z) \frac{\partial C_m}{\partial z} \right).
\]

(3)

Both the DNA concentrations \( \Gamma_r \) and metabolite concentrations \( C_m \) depend on time \( t \) and depth \( z \). The first term in equation (2) corresponds to cell death, with \( q_r \) being the exponential death rate for cells hosting gene \( r \) in the absence of any metabolites. The 2nd term in (2) corresponds to gene production proportional to the per-gene reaction rate \( H_r \) (which in turn depends on metabolite concentrations, see Appendix S2.4). The biomass production coefficient \( Z_r \) is a linear function of the Gibbs free energy of the reaction catalyzed by gene \( r \) and depends on the reaction quotient (Appendix S2.5). In particular, \( Z_r \) increases when product concentrations are low and decreases when substrate concentrations are low. \( c \) is the average dry cell mass, which is used to convert biomass production into cell production. The 3rd term in equation (2) corresponds to cell sinking at a constant speed \( v \). The last term in equation (2) and equation (3) corresponds to diffusive transport (33), with \( K \) being the vertical eddy-diffusion coefficient. The 1st term in equation (3) corresponds to production or depletion of metabolites due to microbial metabolism. Reaction rates are transformed into metabolite fluxes via the stoichiometric matrix \( S \), with entry \( S_{mr} \) corresponding to the stoichiometric coefficient of metabolite \( m \) in reaction \( r \) (Appendix S2.3).

The differential equations (2) and (3) give the rate of change of each metabolite and gene profile, if the profiles are known at a given moment in time. Once all boundary conditions (Appendix S2.6), model parameters (Appendix S2.7) and initial profiles are specified, the model predicts the profiles at any future time point. Steady state profiles were obtained by running simulations of the model until convergence to equilibrium. Because the predicted profiles depend on model parameters, parameters can be calibrated such that the predicted steady state profiles best reproduce the measured data: We used chemical depth profiles to fit poorly known model parameters, thus obtaining a model calibrated to Saanich Inlet’s OMZ (Appendix S2.8). This calibrated model was then used to make predictions about steady state DNA profiles, which were compared to measured metagenomic profiles (sections S1.3 and S2.9). This comparison, described in the main article, serves as a test of the model’s ability to explain metagenomic profiles in Saanich Inlet’s OMZ. Reaction rates at each depth are automatically calculated using the kinetics described in Appendix S2.4.
S2.2 Considered pathways

The model considers key dissimilatory redox pathways involved in nitrogen and sulfur cycling. When comparing model predictions to molecular data, each pathway was represented by a single gene. For example, nitrous oxide reduction (nosZ gene) coupled to hydrogen sulfide oxidation (dsr, apr and sat genes) is formally represented by nosZ. Other pathways considered by the model were aerobic ammonium oxidation (amo), aerobic nitrite oxidation (nxr), partial denitrification of nitrate to nitrous oxide (PDNO) coupled to sulfide oxidation, anammox (hzo) and remineralization of organic matter via aerobic respiration (ROM). PDNO comprises 3 denitrification steps: Reduction of nitrate to nitrite (narGHIJ or napAB genes), reduction of nitrite to nitric oxide (nirKS genes) and reduction of nitric oxide to nitrous oxide (norBC genes), all three of which are suspected to be predominantly performed by SUP05 γ-proteobacteria (16, 34). The first denitrification step was assumed to be leaky, so that a small fraction of nitrite was released into the extracellular environment (35). PDNO was represented by norBC genes when comparing the model to molecular data (Fig. 3a in the main article, but see Figures S4d,e,f for narGHIJ, napAB and nirKS multimolecular data). Aerobic ammonium oxidation included a weak production of nitrous oxide (nitrifier denitrification (36)), although the inclusion of this process did not noticeably affect model predictions because most of the nitrous oxide was produced by PDNO. Aerobic respiration of organic matter included the release of ammonium and sulfate at ratios adjusted to measured C:N:S ratios for marine bacterial biomass (37). The choice of pathways follows the hypotheses made by Hawley et al. (16) based on metagenomic and metaproteomic depth profiles, as well as reports of nitrous oxide reduction in Saanich Inlet’s OMZ (38). Hydrogen sulfide is assumed to originate from the sediments via diffusion, where high rates of sulfate reduction have been observed (39, 40) (Appendix S3.1 for a discussion of this assumption). Figure 1a in the main article gives an overview of the described reaction network. The detailed reaction stoichiometry is given in section S2.3.

S2.3 Pathway stoichiometry

We list the stoichiometry of the dissimilatory redox pathways considered by the model:

- Remineralization of organic matter through aerobic respiration:

\[
\frac{1}{6}\text{POM} + \text{O}_2 \xrightarrow{\text{ROM}} \text{CO}_2 + \text{H}_2\text{O} + \nu \text{NH}_4^+ + \sigma \text{SO}_4^{2-}
\] (4)

where POM corresponds to

\[
(C_6H_{12}O_6)(\text{NH}_4^+)^{6\nu}(\text{SO}_4^{2-})^{6\sigma}
\] (5)

and

\[
1 : \nu : \sigma = 1 : 0.184 : 0.0113
\] (6)

correspond to typical molar C : N : S ratios in marine bacterial biomass (37).
• Aerobic ammonium oxidation:

\[
\text{NH}_4^+ + \frac{1}{2}(3 - L_{\text{amo}}) \times \text{O}_2 \xrightarrow{\text{amo}} (1 + L_{\text{amo}}/2) \times \text{H}_2\text{O} + (1 - L_{\text{amo}}) \times \text{NO}_2^- + (L_{\text{amo}}/2) \times \text{N}_2\text{O} + (2 - L_{\text{amo}}) \times \text{H}^+ .
\]  

(7)

Here, \(L_{\text{amo}}\) is a parameter representing the fraction of N released as \(\text{N}_2\text{O}\) via nitrifier denitrification, compared to the total \(\text{NH}_4^+\) consumed (36). For example, if \(L_{\text{amo}} = 0\), then ammonium is completely oxidized and released as nitrite.

• Aerobic nitrite oxidation:

\[
2\text{NO}_2^- + \text{O}_2 \xrightarrow{\text{nxt}} 2\text{NO}_3^- .
\]  

(9)

• Anaerobic ammonium oxidation (anammox):

\[
\text{NH}_4^+ + \text{NO}_2^- \xrightarrow{\text{hzo}} \text{N}_2 + 2\text{H}_2\text{O} .
\]  

(10)

• Partial denitrification to nitrous oxide (PDNO) coupled to hydrogen sulfide oxidation:

\[
(1 - L_{\text{PDNO}}/2) \times \text{H}_2\text{S} + 2\text{NO}_3^- \xrightarrow{\text{PDNO}} (1 - L_{\text{PDNO}}) \times \text{N}_2\text{O} + 2L_{\text{PDNO}} \times \text{NO}_2^- + (1 - L_{\text{PDNO}}/2) \times \text{SO}_4^{2-} + (1 - L_{\text{PDNO}}) \times \text{H}_2\text{O} + L_{\text{PDNO}} \times \text{H}^+ .
\]  

(11)

(12)

(13)

Here, \(L_{\text{PDNO}}\) is a parameter representing the fraction of \(\text{NO}_2^-\) leaked to the extracellular medium during PDNO, compared to the total \(\text{NO}_3^-\) consumed (35).

• Nitrous oxide reduction coupled to hydrogen sulfide oxidation:

\[
\text{H}_2\text{S} + 4\text{N}_2\text{O} \xrightarrow{\text{hoxZ}} 4\text{N}_2 + \text{SO}_4^{2-} + 2\text{H}^+ .
\]  

(14)

• Nitrate reduction to ammonium (DNRA, identified with the nirBD gene):

\[
\text{H}_2\text{S} + \text{NO}_3^- + \text{H}_2\text{O} \xrightarrow{\text{DNRA}} \text{NH}_4^+ + \text{SO}_4^{2-} .
\]  

(15)

DNRA was eventually omitted from the model for reasons described in Appendix S3.2.

\section*{S2.4 Reaction kinetics}

Respiration of organic matter involves the hydrolysis of particulate organic matter (POM) to dissolved organic matter (DOM), which is subsequently broken down to simpler organic molecules by fermenters that provide non-fermenting organotrophs with a reactive DOM pool. However, reactive DOM rarely accumulates and most of the DOM pool is expected to be refractory (41). Furthermore, POM degradation has been found to be strongly correlated to bacterial growth in subeuphotic zones,
likely due to limiting POM hydrolysis rates (42). We thus modeled organic matter respiration rates as a first-order function of POM concentrations (43). More precisely, the gene-specific ROM reaction rate, \( H_{\text{ROM}} \), is a function of metabolite concentrations \( C \) given by

\[
H_{\text{ROM}}(C) = A_{\text{ROM}}F_T \times \frac{C_{\text{POM}}C_{\text{O}_2}}{C_{\text{O}_2} + K_{\text{ROM,O}_2}},
\]

where \( K_{\text{ROM,O}_2} \) is the oxygen half-saturation constant, \( A_{\text{ROM}} \) is a first-order rate constant (“affinity”) and \( F_T \) is the unitless thermodynamic potential factor given by Reed et al. (31) (equation S1).

Half-saturation constants reported for nitrous oxide oxidation are typically in the order of \( 0.37 - 2.5 \ \mu\text{M} \text{N}_2\text{O} \) (44, 45) and \( 40 \ \mu\text{M} \text{H}_2\text{S} \) (46), which are well above the typical \( \text{N}_2\text{O} \) and \( \text{H}_2\text{S} \) concentrations in the Saanich Inlet OMZ (Fig. 2 in the main article). Sulfide-driven nitrous oxide reduction in Saanich Inlet is therefore likely limited both by electron donor as well as electron acceptor availability. We thus modeled nitrous oxide reduction using first order substrate kinetics with oxygen inhibition:

\[
H_{\text{nosZ}}(C) = A_{\text{nosZ}}F_T \times \frac{C_{\text{N}_2\text{O}}C_{\text{H}_2\text{S}}K_{\text{nosZ,O}_2}}{C_{\text{O}_2} + K_{\text{nosZ,O}_2}},
\]

where \( K_{\text{nosZ,O}_2} \) is the oxygen half-inhibition constant and \( A_{\text{nosZ}} \) is a first-order rate constant.

All other gene-specific reaction rates \( (H_r) \) are modeled using Michaelis-Menten kinetics with possible inhibition (30, 31):

\[
H_r(C) = V_rF_T \times \prod_{m \text{ reactant of reaction } r} \frac{C_m}{K_r + C_m} \times \prod_{n \text{ inhibitor of reaction } r} \frac{K_r^*}{K_r^* + C_n}. \tag{18}
\]

Here, \( V_r \) is the maximum gene-specific reaction rate and \( K_r \) and \( K_r^* \) are half-saturation and half-inhibition constants, respectively, given in Appendix S2.7. The only explicitly modeled inhibition was oxygen inhibition for anammox (hzo), PDNO and nitrous oxide reduction (nosZ).

## S2.5 Gibbs free energy and gene growth

Following Roden et al. (32) and Reed et al. (31), we set

\[
Z_r = 2.08\gamma_r^e - 0.0211\Delta G_r,
\]

(in g biomass per mole reaction flux) where \( \gamma_r^e \) is the negative stoichiometric coefficient of the electron donor in the reaction,

\[
\Delta G_r = \Delta G_r^o + R_gT \ln Q_r
\]

10
is the Gibbs free energy of the reaction (in kJ per mol), $\Delta G_r^\circ$ is the standard Gibbs free energy of the reaction and

$$Q_r = \prod_m C_m^{S_m r}$$  \hspace{1cm} (22)

is the reaction quotient (47). Each $\Delta G_r^\circ$ depends on the local temperature and pressure and was calculated using the CHNOSZ R package (48).

### S2.6 Boundary conditions

Uniquely solving the partial differential equations (2) and (3) requires appropriate boundary conditions (BC) for all genes and metabolites at the top and bottom boundaries (100 m and 200 m, respectively). For all metabolites except $\mathrm{N}_2$, $\mathrm{N}_2\mathrm{O}$, $\mathrm{SO}_4^{2-}$ and $\mathrm{O}_2$, BCs were fixed values set to the average measurements from cruises 41 (SI041_01/13/10), 42 (SI042_02/10/10) and 43 (SI043_03/10/10). For $\mathrm{N}_2$ and $\mathrm{N}_2\mathrm{O}$, lower BCs were set to Neumann (zero flux). For $\mathrm{O}_2$, we used Dirichlet BCs (fixed value) with values equal to the average measurements from cruise 42 and 44 (SI044_04/07/10), because $\mathrm{O}_2$ data were unavailable for cruises 41 and 43. For $\mathrm{SO}_4^{2-}$ we used Dirichlet BCs set to 28 mM on both sides (43). Metabolite boundary conditions are summarized in Table S2. These boundary conditions result in a net oxygen and nitrate influx from the top as well as an ammonium and sulfide influx from the sediments (40, 49, 50).

Gene boundary conditions were either set to fixed zero (hzo and norBC top BCs, ROM, amo and nxr bottom BCs) or to fixed relative gradients (ROM, amo, nxr, nirBD and nosZ top BCs, hzo, nirBD, norBC and nosZ bottom BCs), with the relative gradient inferred from the metagenomic profiles.

Table S2: Top (100 m) and bottom (200 m) boundary conditions for metabolites in the gene-centric partial differential equation model. Numerical values denote Dirichlet boundary conditions. ‘N’ denotes zero-flux Neumann conditions.

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Top (µM)</th>
<th>Bottom (µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathrm{NH}_4^+$</td>
<td>0</td>
<td>8.67</td>
</tr>
<tr>
<td>$\mathrm{O}_2$</td>
<td>77.23</td>
<td>0</td>
</tr>
<tr>
<td>$\mathrm{NO}_3^-$</td>
<td>27.59</td>
<td>0</td>
</tr>
<tr>
<td>$\mathrm{NO}_2^-$</td>
<td>0.045</td>
<td>0</td>
</tr>
<tr>
<td>$\mathrm{N}_2$</td>
<td>$4.8 \times 10^{-4}$</td>
<td>N</td>
</tr>
<tr>
<td>$\mathrm{SO}_4^{2-}$</td>
<td>$28 \times 10^3$</td>
<td>$28 \times 10^3$</td>
</tr>
<tr>
<td>$\mathrm{H}_2\mathrm{S}$</td>
<td>0</td>
<td>14.07</td>
</tr>
<tr>
<td>$\mathrm{N}_2\mathrm{O}$</td>
<td>$24.49 \times 10^{-3}$</td>
<td>N</td>
</tr>
</tbody>
</table>
### S2.7 Model parameterization

Half-saturation and half-inhibition constants for all involved pathways are listed in Table S3. Maximum cell-specific reaction rates were set to $V_{amo} = 1.23 \times 10^{-13}$ mol/(cell·d) (51), $V_{nxr} = 3.26 \times 10^{-13}$ mol/(cell·d) (51) and $V_{hzo} = 2 \times 10^{-14}$ mol/(cell·d) (52, 53). The nitrifier denitrification fraction $L_{amo}$ was set to $10^{-4}$, according to nitrifier denitrification fractions of marine ammonium oxidizing archaea measured by Santoro et al. (36, Fig 2) over varying NO$_2$ concentrations, and the fact that in Saanich Inlet NO$_2$ concentrations are typically below 2 µM (Fig. 2 in the main article). Because of a lack of reliable information, the rate constants $A_{ROM}$, $V_{PDNO}$ and $A_{nosZ}$, as well as the PDNO leakage fraction $L_{PDNO}$, were calibrated to chemical profiles as described in Appendix S2.8 and in the main article. Calibration yielded $A_{ROM} = 5.11 \times 10^{-9}$ L/(cell·d), $V_{PDNO} = 2.18 \times 10^{-14}$ mol/(cell·d), $A_{nosZ} = 0.098$ L/(cell·d) and $L_{PDNO} = 0.352$. An overview of fixed and calibrated reaction-kinetic parameters is provided in Table S3. The sensitivity of the model to parameter variation is illustrated in Appendix S2.12.

The dry cell mass was assumed to be $c = 5 \times 10^{-13}$ g, for consistency with the mass used by Roden et al. (32) to obtain the regression formula (20). Cell death rates were set to $q_{ROM} = 0.063$ d$^{-1}$ in accordance with turnover times estimated by Whitman et al. (54) for marine prokaryotic heterotrophs above 200 m; to $q_{amo} = 0.024$ d$^{-1}$ in accordance with average values reported for ammonium oxidizing bacteria (55); to $q_{nxr} = 0.054$ d$^{-1}$ corresponding to values estimated for nitrite oxidizers (56) and to 0.0033 d$^{-1}$ for all other genes, in accordance with turnover times estimated by Whitman et al. (54) for marine prokaryotes below 200 m.

The concentration of H$^+$ was fixed to 8.5 nM, corresponding to pH = 8.07 (57). The total dissolved inorganic carbon (DIC) was fixed to 2141 µmol/kg (58) and a surface water density of 1018 kg/m$^3$. Accordingly, the dissolved CO$_2$ concentration was fixed at 28 µM according to aquatic carbonate equilibrium at the given pH and DIC (59). The particulate organic carbon (POC) profile was calculated from data reported for February 2011 by Luo et al. (60) and POM was set to $(1/6) \times$ POC (Fig. S1c in the Appendix). Fixing the POM profile circumvents poorly understood physical processes contributing to organic matter fluxes in Saanich Inlet (60). CO$_2$, H$^+$ and POM concentrations, while fixed, were still included in the reaction quotients (Appendix S2.5) as well as the reaction-kinetics (Appendix S2.4).

The diapycnal eddy diffusion coefficient $K$ was set to $N^{-2} \times 3.7 \times 10^{-10}$ W·kg$^{-1}$, where $N$ is the buoyancy frequency (61, 62). The latter was calculated using temperature and salinity profiles from January 13, 2010, using the oce R package (63) (Fig. S1 in the supplement) after loess-smoothing temperature at degree 2 and salinity at degree 1, with a span of 75%. We chose this time point because the two subsequent temperature and salinity measurements (February 10th and March 10th) were unreliable due to technical problems with our CTD. The cell sinking speed $v$ was set to 0.1 m/day, in accordance with previous marine microbial ecological models (64, 65).
Table S3: Reaction-kinetic parameters used in the gene-centric model, either calibrated or taken from the literature: Half-saturation substrate concentrations ($K$), half-inhibition concentrations ($K^*$), cell-specific maximum rates for 2nd order kinetics ($V$), 1st order kinetic constants ($A$, “affinities”), nitrifier denitrification fraction ($L_{amo}$) and PDNO leakage fraction ($L_{PDNO}$). The exact role of each parameter is explained in Appendix S2.4. Additional (non-kinetic) fixed model parameters are provided in Appendix 2.7. Clades with members that have been found active in the Saanich Inlet OMZ (16) are marked with a “†”.

<table>
<thead>
<tr>
<th>reaction</th>
<th>parameter</th>
<th>value</th>
<th>units</th>
<th>organism/region</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>$K_{O_2}$</td>
<td>0.121</td>
<td>µM</td>
<td>$Escherichia$ $coli$</td>
<td>(66)</td>
</tr>
<tr>
<td></td>
<td>$A$</td>
<td>5.11</td>
<td>nL/(cell · d)</td>
<td>calibr.</td>
<td></td>
</tr>
<tr>
<td>amo</td>
<td>$K_{NH_4^+}$</td>
<td>0.133</td>
<td>µM</td>
<td>$Ca$. Nitrosopumilus $maritimus$†</td>
<td>(67)</td>
</tr>
<tr>
<td></td>
<td>$K_{O_2}$</td>
<td>3.91</td>
<td>µM</td>
<td>$Ca$. Nitrosopumilus $maritimus$†</td>
<td>(68)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>123</td>
<td>fmol/(cell · d)</td>
<td>Nitrosomonas $spp.$†</td>
<td>(51)</td>
</tr>
<tr>
<td></td>
<td>$L_{amo}$</td>
<td>$10^{-4}$</td>
<td>–</td>
<td>marine ammonia oxidizing archaea†</td>
<td>(36)</td>
</tr>
<tr>
<td>nxr</td>
<td>$K_{NO_2^-}$</td>
<td>11.7</td>
<td>µM</td>
<td>$Nitrospira$ $spp.$†</td>
<td>(69)</td>
</tr>
<tr>
<td></td>
<td>$K_{O_2}$</td>
<td>0.78</td>
<td>µM</td>
<td>Chilean OMZ</td>
<td>(70)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>326</td>
<td>fmol/(cell · d)</td>
<td>Peruvian OMZ</td>
<td>(31, 71)</td>
</tr>
<tr>
<td>hzo</td>
<td>$K_{NH_4^+}$</td>
<td>3</td>
<td>µM</td>
<td>$Ca$. Scalindua $sp.$†</td>
<td>(26)</td>
</tr>
<tr>
<td></td>
<td>$K_{NO_2^-}$</td>
<td>0.45</td>
<td>µM</td>
<td>$Ca$. Scalindua $sp.$†</td>
<td>(26)</td>
</tr>
<tr>
<td></td>
<td>$K^*$</td>
<td>0.2</td>
<td>µM</td>
<td>marine anoxic basin</td>
<td>(72)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>20</td>
<td>fmol/(cell · d)</td>
<td>Planctomycetales†</td>
<td>(52, 53)</td>
</tr>
<tr>
<td>PDNO</td>
<td>$K_{NO_3^-}$</td>
<td>2.9</td>
<td>µM</td>
<td>marine anoxic basin</td>
<td>(72)</td>
</tr>
<tr>
<td></td>
<td>$K_{H_2S}$</td>
<td>2</td>
<td>µM</td>
<td>Saanich Inlet OMZ</td>
<td>(73)</td>
</tr>
<tr>
<td></td>
<td>$K^*$</td>
<td>0.1</td>
<td>µM</td>
<td>Eastern South Pacific OMZ</td>
<td>(65, 74)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>21.8</td>
<td>fmol/(cell · d)</td>
<td>calibr.¹</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$L_{PDNO}$</td>
<td>35.2</td>
<td>%</td>
<td>calibr.²</td>
<td></td>
</tr>
<tr>
<td>nosZ</td>
<td>$K^*$</td>
<td>0.971</td>
<td>µM</td>
<td>low-oxygen activated sludge</td>
<td>(75)</td>
</tr>
<tr>
<td></td>
<td>$A$</td>
<td>0.098</td>
<td>L/(cell · d)</td>
<td>calibr.</td>
<td></td>
</tr>
</tbody>
</table>

¹ Frey et al. (76) reports cell-specific thiosulphate-driven denitrification rates for $Sulfurimonas$ $gotlandica$ in the range $24.2 – 74.3$ fmol/(cell · d).

² Reported fractions of nitrite leakage during incomplete denitrification ($L_{PDNO}$) range from 0% to 87% (77–79).
S2.8 Calibrating reaction-kinetic parameters to data

As described in the previous section, most model parameters were obtained from the literature, however a subset of reaction-kinetic parameters ($A_{ROM}$, $V_{PDNO}$, $L_{PDNO}$ and $A_{nosZ}$; overview in Table S3) had to be calibrated due to the lack of available information. Here we describe the statistical methods used to calibrate unknown reaction-kinetic model parameters to available chemical depth profile data. The steady state solution of the model defines a mapping from a given choice of parameter values (collectively written as a vector $p$) to predicted depth profiles for metabolite concentrations, $C_1, C_2, \ldots$. We assumed that measured concentrations ($\tilde{C}_1, \tilde{C}_2, \ldots$) are normally distributed:

$$\tilde{C}_i = C_i + \sigma_i \cdot \varepsilon_i.$$  \hspace{1cm} (23)

Here, $\varepsilon_i$ is a standard-normally distributed error and $\sigma_i$ is the (unknown) standard deviation of measurement errors (henceforth referred to as error scale). We allowed for a different $\sigma_i$ for each metabolite to account for variations in the magnitude of measurement errors.

In the context of our spatial model, the concentrations $C_i$ are predicted as functions of depth, $z$, i.e. $C_i = C_i(z; p)$. Calibration data is given as tuples $(z_{ij}, \tilde{C}_{ij})$, where each $\tilde{C}_{ij}$ is a measurement of the $i$-th concentration at some depth $z_{ij}$ and $j$ enumerates all measurements of the $i$-th concentration. The overall log-likelihood function for such a data set is given by

$$l(\sigma, p) = -\sum_{i,j} \ln \left( \sigma_i \sqrt{2\pi} \right)$$

$$-\sum_{i,j} \frac{1}{2\sigma_i^2} \left[ \tilde{C}_{ij} - C_i(z_{ij}; p) \right]^2.$$  \hspace{1cm} (25)

The model was calibrated by maximizing the log-likelihood $l(\sigma, p)$ by choice of the error scales $\sigma_i$ and the parameter values $p$. This calibration method is known as maximum-likelihood (ML) estimation, and is widespread in statistical regression and physics (80). Maximization of the log-likelihood was performed using the MATLAB® function $fmincon$, which uses repeated simulations and gradual exploration of parameter space (81). The following chemical concentration data were used for calibration: $\text{NH}_4^+$, $\text{NO}_3^-$, $\text{NO}_2^-$, $\text{N}_2\text{O}$ and $\text{H}_2\text{S}$ from cruises 41–43, and $\text{O}_2$ from cruises 42 and 44.

S2.9 Calibrating multi-omic data units

Metagenomic, metatranscriptomic and metaproteomic data are given only in relative units. For example, the correspondence between metagenomic RPKM values and actual DNA concentrations in the water column is, a priori, unknown. In fact, RPKM values for different genes may correspond to different gene concentrations due to detection biases (82–84). Furthermore, model predictions regarding RNA and protein abundances are in arbitrary units because the transcriptional, translational and enzymatic efficiency of proteins is unknown and differs between proteins.

In order to compare model predictions to multi-omic sequence data, we assumed that each measured DNA, mRNA and protein abundance profile is related to the corresponding model prediction by
a constant linear conversion factor. Conversion factors were estimated via maximum-likelihood estimation, separately for each molecule to account for detection biases. More precisely, for each data set we assumed a normal error distribution as already described in Appendix S2.8. Hence, measured environmental biomolecule concentrations, for example amo DNA concentrations, are distributed as

\[ \tilde{\Gamma}_i = \Gamma_i / \beta_i + \sigma_i \cdot \varepsilon_i, \]  

where \( \varepsilon_i \) are uncorrelated standard-normally distributed errors, scaled by an unknown factor \( \sigma_i \), and \( \beta_i \) is the unknown proportionality factor between amo metagenomic RPKM values \( \tilde{\Gamma}_i \) and actual DNA concentrations. The log-likelihood of a measured depth profile comprising \( N_i \) data points, \((z_{i1}, \tilde{\Gamma}_{i1}), \ldots, (z_{iN_i}, \tilde{\Gamma}_{iN_i})\), is thus

\[ l_i(\sigma_i; p) = -\sum_{j=1}^{N_i} \ln \left( \sigma_i \sqrt{2\pi} \right) - \sum_{j=1}^{N_i} \frac{1}{2\sigma_i^2} \left[ \tilde{\Gamma}_{ij} - \Gamma_i(z_{ij}; p) / \beta_i \right]^2. \]  

For any fixed model parameter choice \( p \) (and therefore fixed predictions \( \Gamma_i \)), the log-likelihood \( l_i(\sigma_i; p) \) is maximized by choosing

\[ \beta_i = \sqrt[N_i]{\prod_{j=1}^{N_i} \frac{\Gamma_i(z_{ij}; p)}{\tilde{\Gamma}_{ij}}} \]  

(i.e. the geometric mean of model predictions over measurements) and

\[ \sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} \left| \tilde{\Gamma}_{ij} - \Gamma_j(z_{ij}; p) / \beta_i \right|^2. \]  

Choosing \( \beta_i \) as in equation (28) yields maximum-likelihood estimates for the appropriate conversion factors between metagenomic units (RPKM) and actual DNA concentrations (genes/L) (see table S4 in the supplement). Inserting the estimated \( \beta_i \) and \( \sigma_i \) back into equation (27) yields the log-likelihood of the particular metagenomics data set and for a particular choice of model parameters \( p \). A similar approach was used to compare metatranscriptomic and metaproteomic data sets to model predictions (Appendix S2.10).

The estimated proportionality factors \( \beta_i \) are listed in table S4 of the supplement, and range from \( 3.9 \times 10^4 \) genes \cdot L^{-1} \cdot RPKM^{-1} \) for norBC up to \( 3.3 \times 10^7 \) genes \cdot L^{-1} \cdot RPKM^{-1} \) for ROM. These differences may be due to variable DNA extraction efficiencies across cells, uneven community sampling due to filter-size partitioning (19) or differences in gene copy numbers per cell. Additionally, the assumption of a common cell mass for all modeled genes may have resulted in an inaccurate conversion of predicted biomass production to gene production. However, the good overall agreement between predicted functional gene concentrations and SUP05 abundances (Fig. 3 in the main article) suggests that this may only be a minor problem.
Table S4: Proportionality factors ($\beta$) between environmental gene abundances and metagenomic RPKM values (in $\text{genes} \cdot \text{L}^{-1} \cdot \text{RPKM}^{-1}$), as defined in Appendix S2.9. Estimated by comparing the predictions of the calibrated model with metagenomic data from February 10, 2010. Unambiguous metagenomic data was not available for nxr (see Appendix S3.4).

<table>
<thead>
<tr>
<th>gene</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>$4.1 \times 10^4$</td>
</tr>
<tr>
<td>amo</td>
<td>$1.0 \times 10^6$</td>
</tr>
<tr>
<td>nxr</td>
<td>NA</td>
</tr>
<tr>
<td>hzo</td>
<td>$3.2 \times 10^5$</td>
</tr>
<tr>
<td>norBC</td>
<td>$3.4 \times 10^4$</td>
</tr>
<tr>
<td>nosZ</td>
<td>$4.5 \times 10^4$</td>
</tr>
</tbody>
</table>

**S2.10 Predicting metatranscriptomic and metaproteomic profiles**

A priori, the gene-centric model makes no predictions regarding mRNA or protein dynamics; in fact transcription and translation are circumvented by assuming that the release of energy manifests directly as DNA replication. To explore the possibility of explaining mRNA and protein distributions in Saanich Inlet’s OMZ, we extended the model to a set of hypothetical mechanisms driving the production, decay and dispersal of these molecules. More precisely, we assumed that mRNA and protein production rate at a particular depth is proportional to the total reaction rate at that depth ($H_r \Gamma_r$), and that mRNA and proteins disperse similarly to genes (Appendix S2.10). The assumption that mRNA and protein production rates are proportional to reaction rates is motivated by observations of a positive relation between transcription and translation rates and metabolic activity or growth (85–87). A linear relation, in particular, may be justified by the fact that increased enzyme dilution rates at elevated cell growth must be balanced (at the population level) by correspondingly increased translation (and hence transcription) rates (88).

This simple description introduces two unknown parameters per mRNA or protein: The proportionality factor that converts reaction rates to molecule production rates, and the decay time of molecules following production. We calibrated both parameters using metatranscriptomic and metaproteomic depth profiles and then checked how well the latter could be reproduced. Our methodology is described for mRNA in detail below. Protein dynamics were modeled and compared to metaproteomic data in a similar way.

As mentioned, our first assumption was that the mRNA production rate (transcripts produced per time and per volume of seawater) at a particular depth is proportional to the total reaction rate (mol per time and per volume of seawater) at that depth. We also assumed that mRNA molecules disperse via diffusion and sinking similarly to genes, as they are hosted by the same cells. Thus, environmental mRNA concentrations satisfy the partial differential equation

$$
\partial_t T_r = -T_r/\tau_r + R_r/\alpha_r - v \partial_z T_r + \partial_z K \partial_z T_r,$$

where $T_r(t, z)$ is the mRNA concentration corresponding to the $r$-th reaction, $\tau_r$ is the decay time
of the mRNA molecule, $R_r(t, z) = H_r(t, z)\Gamma_r(t, z)$ is the total reaction rate at depth $z$ and $\alpha_r$ is an unknown proportionality constant. We considered $T_r$ in the same units as the multi-omic data (i.e. RPKM for metatranscriptomes and NSAF for metaproteomes). Consequently, $\alpha_r$ is the ratio between the $r$-th reaction rate (mol per time per vole of seawater) and the corresponding RPKM (or NSAF) “production rate” (RPKM per time), and thus not only depends on the particular reaction, but also on our sampling protocol and sequencing pipeline.

For each gene $r$, the transcript profile $T_r$ will satisfy the same boundary conditions as the DNA profile $\Gamma_r$, provided that the latter are either zero value (Dirichlet), zero flux (Neumann) or fixed relative gradient boundary conditions (Appendix S2.6). We calculated the steady state solution of equation (30), $T_r^*$, by solving the time-invariant equation

$$0 = -\frac{T_r^*}{\tau_r} + \frac{R_r}{\alpha_r} - v\partial_z T_r^* + \partial_z K\partial_z T_r^*$$  

(31)

using the MATLAB function bvp4c (81). This was done after the gene-centric model had already reached steady state, at which point the reaction rates $R_r$ are time-independent functions of depth.

Note that the steady state profile $T_r^*(z)$ is proportional to $1/\alpha_r$, all else being equal. Hence, by comparing $T_r^*$ to metatranscriptomic data (for some given $\tau_r$), the constant $\alpha_r$ can be calibrated via maximum-likelihood estimation as described in Appendix S2.9. On the other hand, maximizing the log-likelihood in equation (27) (separately for each gene) by choice of $\alpha_r$, $\tau_r$ and the corresponding error scale, yields an estimate of the decay time $\tau_r$. This was done through repeated solutions of equation (31) with varying $\tau_r$ and using the interior-point optimization algorithm implemented by the MATLAB function fmincon (81). We confined the fitted $\tau_r$ to between $10^{-4}$ and $10^5$ days.

After calibration of the decay time $\tau_r$ and proportionality factor $\alpha_r$, we calculated the coefficients of determination,

$$R^2_r = 1 - \frac{\sum_j \left[ \bar{T}_{r,j} - T_r(z_{r,j}) \right]^2}{\sum_j \left[ \bar{T}_{r,j} - \bar{T}_r \right]^2},$$  

(32)

to evaluate how well the mRNA model explained the metatranscriptomic data. Here, $\bar{T}_{r,1}, \bar{T}_{r,2}, ..$ are measured mRNA abundances at depths $z_{r,1}, z_{r,2}, ..$ and $\bar{T}_r$ is their average. For any given gene $r$, $R^2_r$ is a measure for the goodness of fit of the above model to the multi-omic data. Table S5 in the supplement lists the results for all genes for which $R^2_r \geq 0.5$.

The statistical significance (P-value) of the obtained $R^2$ was defined as the probability of obtaining the same or greater $R^2$ by applying the same procedure to a random data set, with independent normally distributed values with mean and standard deviation set to the original sample mean and standard deviation. We estimated the P-values for cases where $R^2_r \geq 0.9$ using Monte Carlo simulations of 1000 random data sets: all of them were estimated below 0.005.

**S2.11 Calculating metabolic fluxes between pathways**

Dissimilatory metabolic reactions can be interpreted as sources and sinks of metabolites distributed along the water column, producing and consuming metabolites at rates given by the first term in
Table S5: Proportionality factors ($\alpha$) between mRNA or protein production rates and reaction rates (in mol/(L·RPKM) or mol/(L·NSAF), respectively), exponential mRNA or protein decay times ($\tau$) and coefficients of determination ($R^2$), estimated as described in Appendix S2.10. Only cases with $R^2 \geq 0.4$ are shown.

<table>
<thead>
<tr>
<th>molecule</th>
<th>$\alpha$</th>
<th>$\tau$ (days)</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>nxr</td>
<td>mRNA 9.7 x 10^{-10}</td>
<td>52.2</td>
<td>0.93</td>
</tr>
<tr>
<td>nosZ</td>
<td>mRNA 2.5 x 10^{-8}</td>
<td>222</td>
<td>0.95</td>
</tr>
<tr>
<td>ROM</td>
<td>protein 3.9 x 10^{-3}</td>
<td>89.2</td>
<td>0.59</td>
</tr>
<tr>
<td>amo</td>
<td>protein 9.6 x 10^{-6}</td>
<td>16.6</td>
<td>0.88</td>
</tr>
<tr>
<td>nxr</td>
<td>protein 7.5 x 10^{-5}</td>
<td>123</td>
<td>0.42</td>
</tr>
<tr>
<td>norBC</td>
<td>protein 4.9 x 10^{-4}</td>
<td>329</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Due to diffusive transport (2nd term in equation (3)), metabolite fluxes from sources to sinks need not be localized and can span across different depths. Furthermore, some metabolites are partly transported across the OMZ boundaries, towards or from the top layers or the sediments. In the following we describe our approach for calculating steady-state metabolite fluxes across individual reactions.

Let us focus on a particular metabolite and consider a single hypothetical particle created at time 0 at depth $x$. Let $G(t, x, y)$ be the Green’s function of the dispersal-destruction model, so that $G(t, x, \cdot)$ is the distribution density of a particle (created at depth $x$) at depth $y$ and after time $t$. Note that $G(t, x, \cdot)$ may integrate to less than unity if the particle has a positive probability of being consumed anywhere in the water column. The probability rate at which that particle is consumed by any sink $j$ at time $t$ is then

$$\int dy G(t, x, y) \frac{\lambda_j(y)}{C(y)}$$

where $\lambda_j(y)$ gives the rate at which sink $j$ consumes particles at depth $y$ and $C(y)$ is the steady state metabolite concentration at that depth. Since each sink corresponds to a pathway consuming the metabolite, $\lambda_j(y)$ is given by the community-wide reaction rate at $y$ multiplied by the appropriate stoichiometric coefficient. The probability that the particle will eventually be destroyed by sink $j$ is given by

$$\int_0^\infty dt \int dy G(t, x, y) \frac{\lambda_j(y)}{C(y)}$$

The total rate at which particles created by source $i$ are destroyed by sink $j$ across the entire OMZ, denoted $F_{ij}$, is

$$F_{ij} = \int dx b_i(x) \int_0^\infty dt \int dy G(t, x, y) \frac{\lambda_j(y)}{C(y)}$$
where \( b_i(x) \) is the rate at which the metabolite is produced by source \( i \) at depth \( x \). Switching integrals in (35) gives

\[
F_{ij} = \int dy \frac{\lambda_j(y)}{C(y)} \int_0^\infty dt \vartheta_i(t, y),
\]

(36)

where

\[
\vartheta_i(t, y) = \int dx b_i(x) G(t, x, y)
\]

(37)
is the solution to the dispersal-destruction model with initial distribution \( b_i(x) \):

\[
\frac{\partial}{\partial t} \vartheta_i(t, y) = -\frac{\vartheta_i(t, y)}{C(y)} \sum_j \lambda_j(y) + \partial_y [K(y) \partial_y \vartheta_i(t, y)],
\]

(38)

\[
\vartheta_i(0, y) = b_i(y).
\]

(39)

Particles crossing the domain boundary are considered to be lost. Hence, Dirichlet (Neumann) boundary conditions in the original model correspond to zero-value (zero-flux) boundary conditions for \( \vartheta_i \). The total boundary loss rate of particles created by source \( i \) is the remainder

\[
F_{i,o} = \int dx b_i(x) - \sum_j F_{ij}.
\]

(40)

Similarly, the rate at which particles flow in at the boundary and are destroyed by sink \( j \) is given by

\[
F_{o,j} = \int dx \lambda_j(x) - \sum_i F_{ij}.
\]

(41)

We solved equation (38) using the MATLAB® function `pdepe` and evaluated all integrals in equation (36), (40) and (41) using the trapezoid integration scheme (81).

**S2.12 Local sensitivity analysis**

We evaluated the sensitivity of the model predictions to small changes in model parameters using normalized local sensitivity coefficients (NLSC) (89). NLSCs compare the relative changes in model output variables (\( V_j \), integrated over all depths) to the relative changes of model parameters (\( p_i \)) by means of partial derivatives, evaluated at the default (e.g. fitted) parameter values:

\[
\text{NLSC}_{ij} = \left| \frac{p_i \frac{\partial V_j}{\partial p_i}}{V_j \frac{\partial p_i}{\partial p_i}} \right|.
\]

(42)

Hence, \( \text{NLSC}_{ij} \) is a measure for the relative effects that parameter \( i \) has on the output variable \( j \). The partial derivative in equation (42) was approximated numerically by changing \( p_i \) by 1% from its default value. The results are summarized in figure S3 in the supplement.
The sensitivity of the model varied strongly among parameters. For example, the kinetic constants for ROM (aerobic remineralization of organic matter) had a relatively strong effect on chemical as well as gene concentration profiles by modulating the availability of oxygen and ammonium near and above the SNTZ. On the other hand, the kinetic constants for PDNO and nosZ (which constitute the denitrification pathway) had relatively little effects on the predicted chemical profiles, as long as both were increased or decreased in unison. Similar observations were made for amo and nxr, which constitute the nitrification pathway. Moreover, the total predicted gene concentrations (Fig. 3 in the main article) were robust against parameter changes and only varied within an order of magnitude as long as the calibrated geochemical profiles matched the data moderately well. This suggests that geochemical fluxes are good predictors for microbial growth, but less suited for estimating reaction-kinetic parameters, especially when these are correlated (90).

S3  Caveats and special notes

S3.1 The role of sulfate reduction

The choice of pathways included in the model was based on metaproteomics data by Hawley et al. (16). None of the proteins associated with sulfur-metabolism were mapped to known sulfate reducers, suggesting that these proteins may act in sulfur oxidation and that sulfate reduction only played a minor role in Saanich Inlet’s OMZ at the time of sampling. In particular, an NCBI BLASTP search mapped all detected dsrA and aprAB proteins to SUP05 (91). All other taxonomically resolved sulfite reductase proteins were mapped to Candidatus Ruthia magnifica, a sulfur-oxidizing endosymbiont (92). The mRNA depth profiles of sat, aprAB and dsrAB (Figs S4a,b,c in the Appendix), which comprise the dissimilatory sulfide oxidation pathway (or sulfate reduction pathway when reversed), show a clear peak at the SNTZ, consistent with the metatranscriptomic profiles of norBC and nosZ (Fig. 3 in the main article). These multimolecular data suggest that the sat, aprAB and dsrAB enzymes act predominantly in sulfur oxidation. The high sat, aprAB and dsrAB gene concentrations at the bottom might be due to sediment resuspension, cell sinking from the more productive SNTZ or cell diffusion from the sulfate reducing sediments (60, 93).

Due to the much higher organic matter concentrations in the sediments, heterotrophic sulfate reduction and anaerobic remineralization is correspondingly higher in the sediments than in the water column (39, 40). Hence, most of the H$_2$S and NH$_4^+$ in the sulfidic part of the OMZ is expected to originate from the adjacent sediments via diffusion. An influx of H$_2$S and NH$_4^+$ predominantly from the sediments is compatible with the measured steep H$_2$S and NH$_4^+$ gradients (Figs 2 b,f in the main article), as well as the gradual upward progression of the H$_2$S and NH$_4^+$ fronts following annual renewal (Figs 1b and 2 b,f in the main article). Sediments have previously been indicated as the main sulfide sources in other OMZs, such as the the Eastern Boundary upwelling system (94) or the central Namibian coastal upwelling zone (95).

Due to the lack of rate measurements heterotrophic sulfate reduction and cryptic sulfur cycling cannot be completely ruled out. However, calibrating the above model to the chemical data (Fig.
2 in the main article), while including sulfate reduction as an additional pathway, dramatically decreases the goodness of fit. This is because an additional H$_2$S source in the OMZ shifts the SNTZ further up, thereby increasing the main discrepancy between the model and the data. Hence, on grounds of parsimony, we eventually omitted sulfate reduction from the model and assumed that H$_2$S originates from the sediments via diffusion.

We note that similar theoretical work by Reed et al. (31) did suggest the existence of a cryptic sulfur cycle in the Arabian Sea OMZ. However, the latter is located more than 1 km above the sediments and hydrogen sulfide influx from the sediments into the OMZ is not possible due to elevated oxygen levels below the OMZ (96).

S3.2 The role of DNRA

It has been previously hypothesized that dissimilatory nitrate reduction to ammonium (DNRA) might be active in Saanich Inlet’s OMZ, possibly providing ammonium to anammox bacteria (16, 35, 97). So far DNRA was not detected in any of our incubation experiments, although we cannot rule out cryptic DNRA due to rapid ammonium consumption by anammox (97). Measured ammonium profiles in Spring 2010 did not indicate a significant ammonium source at or below the SNTZ (Fig 2 b in the main article). Similarly, Schunck et al. (94) reports negligible DNRA for a sulfidic OMZ off the coast of Peru.

Nevertheless, we tested an extension of our model with DNRA as an additional pathway. Calibrating the model to the same data (January–March 2010) consistently predicted negligible DNRA rates, and the goodness of fit (in terms of the log-likelihood) did not significantly improve with the inclusion of DNRA. On grounds of parsimony we thus eventually omitted DNRA from the model. We mention that calibrating the model to chemical data from September 2009 (16) indicated significant DNRA as well as anammox rates (both in the order of 1 mmol N/(m$^2$$\cdot$ d)), suggesting that DNRA-fed anammox activity fluctuates strongly throughout the year. High spatiotemporal variability of N-loss activities are known for other OMZs and may be associated with fluctuations in surface primary production, as well as fluctuations in electron acceptor availability driven by annual deep water renewal (98–100).

S3.3 The role of aerobic sulfide oxidation

Extensive previous work points towards NO$_3^-$ and other nitrogen compounds as dominant electron acceptors for H$_2$S oxidation in Saanich Inlet during periods of strong stratification (3, 5, 16, 101, 102). For example, as shown in Fig. 1b in the main article, the upper boundary of H$_2$S concentrations closely follows the lower boundary of NO$_3^-$ — rather than O$_2$ — over time, especially during the period considered in this study (early 2010). The strong similarity between sulfur cycling gene profiles and denitrification gene profiles (February 10, 2010; Fig. S4) provides further evidence for the tight coupling between denitrification and sulfide oxidation at that time. Similarly, nitrogen
compounds have been shown to be the dominant electron acceptors for sulfide oxidation in the Peruvian OMZ (94), and Canfield et al. (103) established a strong link between sulfide oxidation and nitrate reduction in the Chilean OMZ. Note that during renewal events in Fall, O$_2$ can indeed become an important electron acceptor for H$_2$S oxidation in Saanich Inlet (3). This does not, however, affect this study, which focuses on periods of intense stratification near steady state conditions.

We note that we had initially considered aerobic sulfide oxidation as an additional reaction in our model. Preliminary calibrations to geochemical data showed that the model’s explanatory power was significantly compromised by this reaction, because diffusive O$_2$ fluxes into the sulfidic zone could not account for the O$_2$ needed for sulfide oxidation (in addition to O$_2$ needed for nitrification). In fact, in our simulations ammonium ended up competing with H$_2$S for O$_2$, which in turn negatively affected the accuracy of the predicted NO$_3^-$ profile. While lateral intrusions of oxygenated water could in principle account for the additional O$_2$ needed for sulfide oxidation, spatiotemporal O$_2$ profiles do not provide any indication of such intrusions during this period of intense stagnation (Fig. 1b in the main article). We thus omitted aerobic sulfide oxidation from our final model.

**S3.4 Planctomycetes and nxr**

Our molecular data suggest that the anammox bacteria planctomycetes (23) are also aerobically oxidizing nitrite to nitrate in the oxycline (16) using the nitrate oxidoreductase narGHIJ (15). Meta-transcriptomic and metaproteomic profiles of planctomycete-associated narGHIJ sequences peak at about 120 m depth and decrease rapidly below that (Fig. 3 in the main article), while planctomycete-associated HAO (anammox-associated hydroxylamine-oxidoreductase (15)) sequences are most abundant at 150 m depth and at appreciable levels all the way down to 200 m. As a consequence, narGHIJ is expected to also proliferate in regions where it is not actually being transcribed. Indeed, metagenomic data show a bimodal profile of Planctomycete-associated narGHIJ sequences, with local maxima at 120 m and 150 m depths, corresponding to the putative maxima of nitrite oxidation and anammox activity. Due to this bimodality we did not include narGHIJ nor nxr metagenomic profiles in our analysis.

**S4 Simulation code**

All simulations, model calibration and sensitivity analysis were performed with MATLAB® (81). The complete code is available upon request from Stilianos Louca. In the code, the biochemical model is defined as a list of genes, a list of metabolites and a stoichiometric matrix for all involved pathways. In addition, the user can specify optional depth profile data sets for chemical concentrations as well as metagenomics, metatranscriptomics and metaproteomics. These are then automatically compared to the model predictions, or used for model calibration. The diffusion coefficient can be provided as an external data set (e.g. calculated from standard CTD data), or internally as a mathematical function. Boundary conditions for the partial differential equations can be specified as Dirichlet (i.e. fixed value) or Neumann (i.e. fixed derivative), independently for
each metabolite or gene. The set of parameters to be calibrated or perturbed for sensitivity analysis can be customized in the code.

S5 Inverse linear transport modeling (ILTM)

Chemical concentration profiles were used to estimate denitrification and anammox rates across the water column, independently of the gene-centric model and the rate measurements described in Appendix S1.5. In short, a steady state diffusion model was used to estimate the net metabolite production (or consumption) rates that “best” explained the observed depth profiles. This so called inverse linear transport modeling (ILTM) approach is widespread in oceanography and atmospheric sciences, were known global distributions of compounds such as trace gases are used to estimate unknown sources and sinks (104, 105).

In the following, we explain our procedure for estimating the net production profile, \( \rho(z) \), for a particular metabolite with a given concentration profile, \( \hat{C}(z) \). All calculations were performed in Mathworks MATLAB®. Each profile \( \hat{C}(z) \) was obtained through Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation of the actual measured concentrations. ILTM was applied separately to concentration profiles from cruises 47 and 48, as well as to the chemical profiles used for model calibration (cruises 41–44, Appendix S2.8) after averaging across replicates at each depth.

Our starting point is the diffusive transport model

\[
0 = \rho + \frac{\partial}{\partial z} \left[ K(z) \cdot \frac{\partial C}{\partial z} \right],
\]

which describes the steady-state distribution \( C(z) \) across depth \( z \), given a particular net production profile \( \rho(z) \) and eddy diffusion coefficient \( K(z) \). The eddy diffusion coefficient was calculated as described in Appendix S2.7. Our goal is to determine the appropriate \( \rho(z) \) that “best” explains the observed steady state profile \( \hat{C}(z) \), through the following steps:

1. Calculate the discretized Green’s function (106) of the above partial differential equation (PDE) with zero Dirichlet boundary conditions: Let \( G_{nm} \) be an approximation for \( G(z_n, z_m) \), where \( G \) solves the time-independent PDE

\[
0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial G(x, y)}{\partial x} \right] + \delta(x - y)
\]

on the domain \( \Omega := [\text{top}, \text{bottom}] \), with boundary conditions

\[
G(x, y) \big|_{x \in \partial \Omega} = 0.
\]

In practice, \( G_{nm} \) can be set to \( dz_m \cdot G(z_n, z_m) \), where \( G \) is the solution to the PDE system

\[
0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial G(x, z_m)}{\partial x} \right] + H(x - z_m + dz_m/2)H(z_m + dz_m/2 - x)/dz_m,
\]

\[
G(x, z_m) \big|_{x \in \partial \Omega} = 0.
\]
Here, \( H \) is the Heaviside step function and \( dz_m \) is the grid’s step at \( z_m \), assumed to be chosen small enough (\( dz = 2 \text{ m} \) in our case).

2. Note that for any candidate net production profile \( \rho(x) \), the sum

\[
\sum_m G_{nm} \cdot \rho(z_m)
\]

becomes an approximation for \( C^o(z_n) \), where \( C^o \) is a solution to the following steady-state transport problem with zero Dirichlet boundary conditions:

\[
0 = \frac{\partial}{\partial x} \left[ D(x) \frac{\partial C^o}{\partial x} \right] + \rho(x), \quad C^o(x) \big|_{x \in \partial \Omega} = 0. \tag{49}
\]

3. For the given measured concentrations \( \hat{C}(x) \) at the domain boundary \( x \in \{ \text{top, bottom} \} \), calculate the particular solution \( C^p \) to the transport problem with given boundary values but no sources:

\[
0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial C^p}{\partial x} \right], \quad C^p(x) \big|_{x \in \partial \Omega} = \hat{C}(x). \tag{50}
\]

After solving for \( C^p \), evaluate \( C^p \) on the grid, i.e. set \( C^p_n = C^p(z_n) \).

4. Note that for any candidate net production profile \( \rho(x) \), the sum \( C := C^o + C^p \) is a solution to the full PDE problem

\[
0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial C}{\partial x} \right] + \rho(x), \quad C(x) \big|_{x \in \partial \Omega} = \hat{C}(x). \tag{51}
\]

Similarly, the sum

\[
C^p_n + \sum_m G_{nm} \cdot \rho(z_m)
\]

is an approximation for \( C(z_n) \).

5. Note that \( C^p \) corresponds to the hypothetical steady-state profile that would result purely from transport across the domain boundary, in the absence of any sources or sinks in its interior. Similarly, the difference \( B = \hat{C} - C^p \) is the part that cannot be explained by transport across boundaries, but must rather be attributed to production and consumption inside \( \Omega \). Hence, using the particular discretized solution \( C^p_n \), the discretized profile \( \hat{C}_n = \hat{C}(z_n) \) and the discretized steady-state transport kernel \( G_{nm} \), one could in principle estimate \( \rho_m = \rho(z_m) \) by minimizing the sum of squared residuals (SSR)

\[
\text{SSR} = \sum_n \left| \sum_m G_{nm} \cdot \rho_m - B_n \right|^2, \tag{53}
\]

where \( B_n = \hat{C}_n - C^p_n \). The above problem is a classical linear least-squares problem if one considers \( G_{nm} \) as a matrix (\( G \)) and \( \rho_m, \hat{C}_n, C^p_n \) as vectors (\( \rho \in \mathbb{R}^M, \hat{C} \in \mathbb{R}^N \) and \( C^p \in \mathbb{R}^N \)):

\[
\text{SSR} = \|G \cdot \rho - B\|^2. \tag{54}
\]
The minimum SSR is then obtained for

\[ \rho = \hat{G} \cdot (\hat{C} - C^p), \]

(55)

where \( \hat{G} \) is the Moore-Penrose pseudoinverse of \( G \). Put simply, the so estimated \( \rho \) is the net production profile that “best” explains the observed steady-state concentration profile \( \hat{C} \), after subtracting the part \( C^p \) explained by transport across the domain boundaries.

6. The least-squares estimator in Eq. (55) becomes unstable if the reference profile \( \hat{C} \) stretches linearly (or almost linearly) across large depth intervals, leading to spurious oscillations in the estimated profile \( \rho \). To address this problem, we “penalized” strong oscillations in the estimated net production profile by instead minimizing the modified SSR

\[ SSR^* = \| G \cdot \rho - B \|^2 + M^{-2} \| \xi \rho \|^2, \]

(56)

where \( \xi \) is an appropriately chosen regularization parameter that quantifies the penalty imposed on large \( |\rho| \). The above regularization method is known as Tikhonov regularization. A larger Tikhonov factor \( \xi \) will typically result in a smoother \( \rho \) but also a poorer overall fit, since goodness of fit is sacrificed in favor of small \( \rho \). We manually chose \( \xi \) as large as possible but still small enough such that the residual \( \| G \cdot \rho - B \| \) remained much smaller than \( \| B \| \).

7. Assuming that \( \text{H}_2\text{S} \) is mostly consumed by denitrification (PDNO and nosZ) according to the stoichiometry given in Appendix S2.3, one mol of consumed \( \text{H}_2\text{S} \) corresponds to \( 8 \cdot (1 - L_{\text{PDNO}})/(5 - 3L_{\text{PDNO}}) \) mol \( \text{N} \) released as \( \text{N}_2 \). Similarly, one mol of consumed \( \text{NH}_4^+ \) by anammox corresponds to 2 mol \( \text{N} \) released as \( \text{N}_2 \), however nitrification likely also contributes to \( \text{NH}_4^+ \) consumption in the more oxygenated layers. Hence, whenever the net \( \text{NO}_3^- \) production was positive, the net \( \text{NO}_3^- \) production rate was subtracted from the net \( \text{NH}_4^+ \) consumption rate, yielding an estimate for \( \text{NH}_4^+ \) consumption purely by anammox.

Figure S1: (a) Temperature and salinity profiles at Saanich Inlet main station, January 13, 2010. (b) Corresponding smoothened eddy diffusivity profile, as used in the simulations. (c) Fixed POM profile used in the simulations.
Figure S2: **Overview of this study.** Previous geochemical and multi-omic investigations provide conceptual information on the metabolic network in the Saanich Inlet OMZ (3, 5, 16, 34, 101, 102). This information was used to construct a gene-centric biogeochemical mathematical model, which describes the population dynamics of individual genes and metabolic process rates. Unknown reaction-kinetic parameters of the model were calibrated using geochemical depth profiles. The predictions of the calibrated gene-centric model were then validated using independent metagenomic sequence data, qPCR-based abundance estimates for SUP05 as well as process rate measurements. A subsequent extension of the model describes the production, dispersal and decay of mRNA and protein molecules based on the reaction rates predicted by the calibrated gene-centric model. Unknown parameters for the mRNA and protein dynamics are estimated using metatranscriptomic and metaproteomic data. The “goodness of fit” to these multi-omic data is used to further evaluate the gene-centric model, to assess the adequacy of the postulated mRNA and protein dynamics and to gain insight into potentially important but omitted mechanisms of mRNA and protein regulation at ecosystem scales.
Figure S3: Local sensitivity heatmap of the calibrated model by means of normalized local sensitivity coefficients. A brighter color corresponds to a higher sensitivity. “Khalf” stands for half-saturation constants, “Kinh” for half-inhibition constants, “V” for maximum cell-specific reaction rates and “q” for cell death rates. The heatmap is hierarchically clustered using UPGMA linkage and Euclidean metric. Methodological details are given in Appendix S2.12.
Figure S4: Metagenomic, metatranscriptomic and metaproteomic depth profiles of (a) sulfate adenylyltransferase (sat), (b) adenylylsulfate reductase (aprAB) and (c) sulfite reductase (dsrAB) genes, which together comprise the sulfide oxidation pathway (or sulfate reduction pathway, if reversed), as well as (d) periplasmic nitrate reductase napAB, (e) nitrate reductase narGHIJ and (f) NO-forming nitrite reductase nirKS. Data taken on February 10, 2014. All of the dsrAB, aprAB and most of the napAB protein sequences were mapped to the γ-proteobacterial SUP05 clade (34). All detected narGHIJ protein sequences were either mapped to SUP05 or to the anammox planctomycete bacteria Candidatus Scalindua profunda and KSU-1 (108) (only SUP05 proteins are shown). Similarly, only non-planctomycete-annotated narGHIJ and nirKS DNA abundances are shown.
References


