Pacific Northwest
NATIONAL LABORATORY

PNNL-38789

A Computational Workflow of Elucidating Viral Impact on Mediating Microbial Response to In-situ Experimental Warming

Bridging microbial modeling to carbon and mineral modeling

September 2025

Ruonan Wu (PI)
Jianqiu Zheng
Winston Anthony
Yuliya Farris

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-38789

A Computational Workflow of Elucidating Viral Impact on Mediating Microbial Response to In-situ Experimental Warming

Bridging microbial modeling to carbon and mineral modeling

September 2025

Ruonan Wu (PI)
Jianqiu Zheng
Winston Anthony
Yuliya Farris

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Viruses are abundant in soils and shape microbial communities in ways that can potentially influence ecosystem processes, yet their contributions to carbon cycling and mineral transformations remain poorly understood. Here we present a multi-phase framework that links virus-host interactions to soil biogeochemistry by combining ecological simulations, genome- and community-scale metabolic modeling, and statistical and machine-learning analyses. We first calibrated microbial abundance profiles under explicit infection scenarios to capture how viral pressure alters community structure, then explored alternative interaction strategies, including kill-the-winner, piggyback-the-winner, and mixed lytic-lysogenic modes, through forward simulations. These ecological shifts were translated into metabolic consequences using exchange fluxes summarized into biologically meaningful categories, while integrated statistical and machine-learning screens elevated subtle but consistent signals. Application of this framework revealed that viral infections shift the balance between organic and inorganic fluxes, redirecting metabolism from diffuse organic transformations toward inorganic pools such as protons and CO_2, directly linking viral regulation to respiration and soil carbon balance. The roll-up analysis also isolated perturbations in critical mineral ions, including magnesium, manganese, zinc, and copper, which serve as essential enzymatic cofactors. In piggyback-the-winner scenarios, uptake of these ions was strongly suppressed. Contrasting viral strategies produced distinct community structures and metabolic outcomes, from broad suppression under kill-the-winner dynamics to dramatic redistributions under high-lytic and high-gain lysogenic regimes that collapsed vulnerable microbial populations while promoting opportunists. Together, these results provide a tractable path to trace viral perturbations from host abundance shifts to metabolic flux adjustments and ecosystem-scale processes, offering a practical way to include viruses in earth system models.

Summary

Viruses are everywhere in soil microbiomes and act as important regulators of microbial populations, but their influence on ecosystem functions, such as carbon cycling and microbially mediated mineral transformations, remains poorly understood. In this work, we introduce a multi-phase framework that connects virus-host interactions to soil biogeochemistry by bringing together ecological simulations, genome/community-scale metabolic modeling, and statistical and machine-learning analyses. We began by calibrating microbial abundance profiles under explicit viral infection scenarios to establish baselines for how infections reshape community structure. We then ran forward simulations of different interaction strategies, including kill-the-winner, piggyback-the-winner, and mixed lytic-lysogenic modes, which revealed that viruses not only reduce the dominance of fast growers but also redirect the trajectories of whole microbial guilds. These ecological changes were then translated into metabolic outcomes using genome- and community-scale models, parameterizing growth rates, biomass yields, substrate stoichiometry, and carbon use efficiency from exchange fluxes summarized into biologically meaningful categories. Finally, by combining statistical methods with machine-learning feature selection, we could detect weak and sparse but consistent patterns and highlight the flux categories most strongly perturbed by viral infection and further trace down the microbial populations that are responsible for such changes.

Applying this framework uncovered several important insights. Viral activity shifted the balance between organic and inorganic fluxes, moving community metabolism away from broad organic transformations toward more concentrated inorganic pools such as protons and CO_2. This directly links viral regulation to respiration and soil carbon balance. The category-based roll-up analyses also made it possible to isolate perturbations in specific resource classes, such as critical mineral ions that serve as essential cofactors in microbial enzymes. In the piggyback-the-winner scenario (S3), viral pressure strongly suppressed uptake of these ions, a result confirmed by multiple machine-learning screens. Linking growth and flux analyses provided a short list of the microbial populations most responsible for driving mineral-linked fluxes under viral influence. Across the scenarios, different viral strategies structured microbial communities in contrasting ways, which in turn produced divergent metabolic outcomes. Kill-the-winner dynamics (S1-S2) broadly suppressed dominant taxa, leading to diffuse and modest metabolic shifts. Piggyback-the-winner (S3-S4) created a more heterogeneous restructuring, selectively stabilizing some host lineages while disadvantaging others, with clear metabolic signatures in amino acid and carbohydrate fluxes. The most dramatic outcomes occurred in the high-lytic (S5) and high-gain lysogenic (S6) cases, which collapsed vulnerable groups such as *Bathyarchaeia*, *Phycisphaerae*, and *Desulfobacterota*, while promoting opportunists like *Acidobacteriota* and *Acidimicrobiia* of the studied system. These community-level changes mapped directly onto metabolism, from carbohydrate uptake and organic acid overflow under S5 to shifts in ion and nitrogen cycling under S6.

Together, these results show how viral perturbations can be traced step by step from host abundance shifts to metabolic adjustments to ecosystem-scale processes. The framework not only demonstrates feasibility but also underscores the ecological importance of explicitly including viruses in biogeochemical models. By integrating virus-host dynamics into carbon and mineral cycling, this approach offers both mechanistic depth and interpretability, creating a practical way to bring viral effects into earth system models. While we focused here on carbon and critical minerals, the same framework can be applied to other elemental cycles, opening new opportunities for predictive and integrative modeling of microbial-ecosystem interactions.

Acknowledgments

This research was supported by the Earth & Biological Sciences Mission Seed Investment, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Contents

Figures

1.0 Introduction

Soils store more than three times the carbon contained in the atmosphere [1], yet the fate of this reservoir under environmental perturbations remains uncertain. Microorganisms are central to regulating soil carbon dynamics, and viruses can modulate these processes in both direct and indirect ways [2, 3]. Soil viruses infect all domains of life, with up to 40% of soil bacteria estimated to harbor inducible temperate viruses [2]. By reprogramming host metabolism during infection and facilitating horizontal gene transfer, viruses alter the functional potential of microbial communities across both short and long timescales. Viral infections also reshape carbon cycling through the release of cellular materials upon host lysis, the so-called viral shunt, which fuels rapid remineralization by other microbes [2, 4]. Alternatively, host-derived macromolecules can aggregate with organic particles and be stabilized in soil, a process referred to as the viral shuttle [5]. Together, these processes influence microbial population dynamics, metabolic capabilities, and ultimately the magnitude and direction of soil carbon fluxes.

Many of these mechanisms have been extensively investigated in marine systems, where viral abundance is among the strongest predictors of global ocean carbon flux, explaining up to 89% of its variance [6]. In contrast, the role of viruses in soil carbon cycling remains poorly understood, largely due to the technical challenges of recovering viruses and predicting viral hosts in the complex soil matrices. Environmental drivers such as soil warming are known to reshape microbial processes and can also modulate virus-host interactions, for example, by increasing infection frequencies [2, 7] or reducing viral persistence [8]. Yet the quantitative impacts of viruses on microbial abundance patterns and downstream carbon metabolism under such perturbations remain unresolved.

To address this gap, we established a multi-phase framework linking virus-host interactions with soil carbon cycling. In Phase 1, we calibrated microbial abundance profiles under virus-infection scenarios, generating baseline estimates of virus-mediated community shifts. In Phase 2, we performed forward simulations of alternative virus-host interaction hypotheses to explore how such dynamics reshape microbial abundance profiles under perturbations. In Phase 3, we applied genome- and community-scale metabolic modeling to evaluate the consequences of these abundance shifts for microbial metabolism and carbon fluxes. In Phase 4, we applied complementary statistical modeling and machine-learning methods to screen biologically meaningful features either at the reaction level or the automated category level. We also demonstrate practical paths forward for connecting microbial metabolic modeling to geochemical modeling and for resolving the community context on which populations are responsible for specific metabolic shifts. Using microbially mediated critical mineral ion import as a use case, we show how viral perturbations can be linked to changes in resource cycling and attributed back to microbial guilds. This approach highlights both the mechanistic depth and ecological relevance of our framework.

2.0 Method

2.1 Metagenome curation and bin analyses

To establish a computational workflow for evaluating viral impacts on soil carbon cycling, we leveraged soil metagenomic data generated by the Spruce and Peatland Responses Under Changing Environments (SPRUCE) experiment. This site has been under long-term (since 2014) in situ deep belowground and aboveground warming treatments, creating a whole-ecosystem gradient of environmental perturbations that capture shifts in virus-host interactions. We compiled a working set of metagenomes from the JGI processing pipeline by excluding samples that failed quality control (i.e., datasets missing a complete folder structure with QC_and_Genome_Assembly, Raw_Data, Sequencing_QC_Reports, IMG_Data, Filtered_Raw_Data, Binning_Data) or that did not yield metagenome-assembled genomes (MAGs) or bins. After filtering, 73 metagenomes were retained for downstream analyses.

Because viruses exert their effects through specific microbial hosts, we began with bin-based analyses to define microbial populations as the fundamental host units for viral interaction. To quantify the abundance of each genomic bin per sample, we integrated JGI bin membership files with scaffold-level depth profiles, mapping scaffolds to bins and joining these mappings to coverage values, normalizing scaffold depths by scaffold length. We then summed the length-weighted scaffold coverages within each bin. To enable cross-sample comparison, we generated a non-redundant set of representative bins using dRep (v3.6.2), with each representing a genomically unique microbial population. We then post-processed clustering output to assign each bin to its cluster identifier and identify the chosen representative. Dereplication reduces statistical non-independence among closely related genomes and focuses downstream analyses on genome-level lineages rather than sample-specific redundancies. Community composition was then summarized using the dereplicated representative bins, producing a matrix of representative bin coverages across samples.

2.2 Viral analyses

We identified viral and proviral sequences using the MVP workflow (v1.1.5), incorporating quality control from CheckV (v0.6) and geNomad (v1.2). Viral sequences were clustered according to community standards (85% aligned fraction and 95% average nucleotide identity). Viral hosts were inferred using CRISPR spacers. Spacers and repeats were extracted from the de novo assemblies, and non-viral spacers were matched to representative viral cluster sequences using short-read BLAST (v2.16.0+) with stringent criteria optimized for CRISPR targeting (≥97% identity and ≤1 mismatch). This threshold balances natural spacer divergence with specificity and minimizes spurious cross-matches. Infection types were assigned with VIBRANT (v1.2.1) applied to viral contigs >10 kb. We then consolidated multiple inputs, including viral infection type, spacer-informed virus-host pairs, infected host bins, representative bin assignments, and combined depth profiles, to resolve virus-host associations. Each host contig was assigned to its corresponding bin and representative bin, and per-phage coverage estimates were appended.

2.3 Calibration of virus-mediated microbial abundance profile

A proxy for viral pressure (VH) was computed for each microbial population (dereplicated bin, rep_bin) in each sample as the mean phage coverage divided by the observed host representative bin coverage (Observed_Total), with a small constant added for numerical

stability (1×10^{-6}). Because per-sample phage coverage can be sparse, we first calculated a sample-specific mean for each rep_bin and, if unavailable, substituted the rep_bin-wide average. VH values were then scaled to the range [0,1] within each rep_bin. The scaled VH was mapped to three quantities that parameterize infection outcomes, the lytic fraction (p), a lytic loss (l), and a lysogenic gain (g). We used VH as a common driver to provide a tractable proxy for the aggregate viral impact on host abundance. We set $p = VH_{\text{scaled}}$ and $l = p$ so that infection probability and lytic loss both scale directly with viral pressure. The gain factor was defined as a bounded transformation $g = \min\{1 + \gamma(1-p), \text{max\_gain}\}$, with γ fixed at 0.50 (gamma_gain) and max_gain set to 2.0. The moderate but bounded lysogeny benefits capture the concept that prophages can enhance host fitness and these effects are typically modest and do not exceed a doubling of host potential. These terms together define the per rep_bin, per sample denominator $K_{\text{den}} = p(1-l) + (1-p)g$, which governs how infection redistributes host abundance between infected and uninfected states.

To estimate the baseline host abundance in the absence of viral effects for each rep_bin, S_{tot}, we take the median Observed_Total among low-VH samples within that rep_bin, because samples with weak viral pressure are most likely to reflect host populations close to their intrinsic carrying capacity. The low-VH is defined as $VH_{\text{scaled}} \leq 0.10$ and falling back to the lower quantile if needed to ensure that baseline estimates are anchored in empirically observed states and avoid bias from viral suppression. We then estimate a rep_bin-specific infected fraction parameter (q). Because $\text{Pred\_Total} = S_{\text{tot}}\left(qK_{\text{den}} + (1-q)\right)$, then $\frac{\text{Pred\_Total}}{S_{\text{tot}}} - 1 = q\,(K_{\text{den}} - 1)$, and assuming that observed totals approximate the model at the rep_bin level, q was calculated by regressing $y_s = \frac{\text{Observed\_Total}_s}{S_{\text{tot}}} - 1$ on $x_s = K_{\text{den},s} - 1$ across samples of each rep_bin, using a slope-through-origin fit. The resulting slope is truncated to [0,1] and given an infection type-specific floor using the global infection class (q_floor_map = {lytic-lysogenic: 0.10, lytic-only: 0.05, lysogenic-only: 0.05}), producing a global q_{hat} per rep_bin. For each per rep_bin per sample, we also compute a sample-level $\hat{q}$ from the same relation to minimize the noise from single-sample estimates and then partially pool toward the global estimate with λ_q of 0.30, so that $q_{\text{eff}} = (1-\lambda_q)\,q_{\text{sample}} + \lambda_q\,q_{\text{hat}}$. This partial pooling is to balance sample-specific and rep_bin-level estimates of the infected fraction, giving greater weight to local variation while still stabilizing estimates with a global anchor that reflects the assumption that infection fractions vary across samples but also share a consistent baseline within specific microbial taxon represented by each rep_bin.

Predictions follow directly from these parameters. The total, infected, and uninfected abundances of each of the microbial populations represented by rep_bins are $\text{Pred\_Total} = S_{\text{tot}}\left(q_{\text{eff}}K_{\text{den}} + (1-q_{\text{eff}})\right)$, $\text{Pred\_Infected} = S_{\text{tot}}\,(q_{\text{eff}}K_{\text{den}})$, $\text{Pred\_Uninfected} = S_{\text{tot}}\,(1-q_{\text{eff}})$. These were computed for all infected rep_bins and used to construct a baseline $S0$, the abundance expected in the absence of viral effects. When the model produced a valid prediction of the uninfected state (Pred_Uninfected), that value was taken directly as the S0 baseline. If no Pred_Uninfected could be calculated for a given rep_bin, the baseline abundance estimate (S_{tot}) for that rep_bin was instead replicated across all of its samples to provide a consistent no viral impact reference. For rep_bin*sample combinations labeled as uninfected, the S0 entry was simply set to the observed abundance. This ensured that every possible rep_bin*sample pair had a corresponding baseline value, giving a complete reference matrix for downstream comparisons.

2.4 Simulations of the six virus-host interaction hypotheses

We define a general simulator that takes a vector of target lytic fractions (p_{sim}) for each rep_bin*sample combination, applies the same loss mapping, and computes a lysogenic gain. The gain is calculated as $g_{\text{raw}} = 1 + \gamma(1 - p_{\text{sim}})$, capped by $g_{\max}$, and then adjusted with the infected fraction q. For each scenario simulation, q was explicitly set, otherwise the pooled q_{eff} was used. Predicted totals are then recomputed using the same structural formula as in calibration but substituting this scenario-specific denominator.

Four structured hypotheses (S1-4) and two extreme cases (S5-6) are implemented. Scenarios S1 and S2 (kill-the-winner) increase lytic pressure on winners, defined either by abundance within each sample (S1) or by cross-sample occupancy (S2). For winners, p is set high (mean of p_winner = 0.85-0.95). For others, p is set low (p_other = 0.02-0.10). Viral gain is muted (gamma = 0.20, g_max = 1.2). In both, the infected fraction q is floored higher for winners (q_winner_floor = 0.40) than for others (q_other_floor = 0.10). Scenarios S3 and S4 (piggyback-the-winner) invert the pressure so that winners have low lytic fractions (p_winner = 0.10-0.20) and others are more lytic (p_other = 0.50-0.70) with higher lysogenic gain (gamma = 1.20, g_max = 2.0) and the same q floors (0.50 vs 0.10). S5 and S6 are two extreme cases where S5 places the system into a mostly lytic, low-gain state (p_fix = 0.99, gamma = 0.01, g_max = 1.05, q_floor_uniform = 0.50), and S6 into a mostly lysogenic, high-gain state (p_fix = 0.01, gamma = 1.50, g_max = 3.00, q_floor_uniform = 0.50). To preserve overall per-sample abundance budget while emphasizing the intended contrasts, S1-S4 apply a one-sided renormalization. In brief, after computing Pred_Infected and Pred_Uninfected, the uninfected mass of non-winners is scaled so that total predicted abundance per sample matches the S0 target, leaving the winner totals and infected mass unchanged to ensure robust contrasts without budget violations.

2.5 Genome and community metabolic modeling

Genome-scale metabolic models (GEMs) were reconstructed for each representative bin (rep_bin) by executing CarveMe (v1.5.1) in parallel across bins, taking protein FASTA files as input and generating SBML models. Models were produced both with and without gapfilling, using the default M9 minimal medium (-g M9 -i M9) and flux balance consistency (--fbc2) options to ensure compatibility with downstream community metabolic modeling. Following reconstruction, models were balanced elemental and charge stoichiometry, harmonized reaction and metabolite identifiers with ModelSEED and BiGG namespaces, and appended standardized annotations.

To evaluate how shifts in microbial composition propagate to ecosystem-scale metabolic fluxes under different virus-host infection scenarios, we integrated the rep_bin abundance profiles from the S0-S6 simulations with the curated GEMs to build sample-specific community models for each scenario using a customized MICOM workflow (v0.37.1). All communities were simulated under a shared global medium to isolate the effect of community composition from environmental variability. Exchange reactions were enumerated from the cached models, an initial permissive base medium was assigned by setting a maximum uptake bound of 10 mmol gDW^{-1} h^{-1} for each reaction, and the complete_community_medium function was used to calculate a feasible, community-consistent medium vector. Community growth was then optimized with MICOM parsimonious flux balance analysis (grow() with pFBA), applying a tradeoff parameter of 0.5 to balance individual taxon growth with community-level feasibility. For each sample, growth rates of the rep_bins were exported alongside community exchange fluxes, following the MICOM sign convention.

2.6 Statistical modeling and machine learning for feature screening

Community model outputs from MICOM simulations across scenarios (S0-6) were grouped and compared to examine how virus-host interactions could influence microbial growth and metabolism with community context. Baseline comparisons were made by computing log2 fold-change values relative to scenario S0 for both growth and import/export exchange fluxes.

To identify differential responses, we applied complementary statistical approaches. Empirical Bayes moderated t-tests implemented in the limma framework were used to compare each perturbation scenario (S1-S6) against the baseline S0, with false discovery rates controlled at $q \leq 0.20$. In parallel, binomial directionality tests were used to evaluate whether increases or decreases beyond a specified effect threshold occurred more frequently than expected by chance across samples. These tests were corrected for multiple hypothesis testing using FDR adjustment, and features were only retained if they passed both statistical significance ($q \leq 0.20$) and a minimum absolute effect size of $\tau = 0.30$ on the log2FC scale. Features meeting either criterion were considered statistically responsive.

To further prioritize multivariate signals, three machine learning-based screening strategies were applied. The first approach used stability-selected LASSO via the glmnet framework, where scenario versus baseline labels were predicted from reaction-level features. Selection probabilities were recorded for each feature, and those with probabilities above 0.70 were retained. The second approach applied gradient boosting classifiers with XGBoost and computed Shapley additive explanations (SHAP) to quantify feature importance across samples. Mean absolute SHAP values were tested against null distributions, and features were considered significant at $q \leq 0.20$. The third approach examined principal component loadings from PCA applied to centered and standardized feature matrices, identifying reactions with significant contributions to components that separated baseline and perturbation scenarios. Features were retained at $q \leq 0.20$. The outputs of these three analyses were stored as independent screening sets that could be intersected or combined with the limma and binomial frameworks mentioned above to construct consensus responsive feature sets. To summarize the response counts and effect size across scenarios, exchange reactions and growth features were mapped and automatically rolled up to 23 functional categories, including Critical minerals, Amino acid, Gas, Carbohydrate, Lipid/fatty acid.

3.0 Uneven host dominance with lytic-skewed viral linkages

Microbial populations (rep_bins) are strongly uneven across the metagenomes, with a few dominant bins consistently enriched while most remain moderate to low abundance. Across 73 samples, we observed 72 dereplicated rep_bins or bin clusters forming 317 rep_bin*sample combinations (**Supplementary Data 1**). Weighted read coverage of the rep_bins spanned from 6.8 to 238.1 (median 15.8, mean 27.5), and several bin clusters were repeatedly high, for example, rep_bin 3300017941_22 (mean 69.4 among 23 samples) and rep_bin 3300009630_7 (mean 53.5 among 25 samples). Viral contigs were present at overlapping scales (mean 16.3, median 12.6, range from 5.4 to 74.0, **Supplementary Data 2**).

Viral infections are abundant and lytic-skewed by count, but coverage-weighted analyses reveal a more balanced contribution from lysogenic viral infections. Virus-host associations were extensive, totaling 1,249,222 linkages that involved 96 viral contigs and 31,537 unique host contigs (**Supplementary Data 2**). Nearly 75% of the host contigs were binned or assigned to 27 rep_bins (with rep_bin assignments available for 935,766 links). By count, infection-type calls revealed a strong contrast between lytic and lysogenic associations. Of the 1,249,222 virus-host linkages, 730,088 (58.4%) were classified as lytic, compared to only 197,364 (15.8%) lysogenic, with 321,770 (25.8%) unresolved. This near 4-fold difference between lytic and lysogenic calls underscores that the infection landscape is lytic-skewed by count, even though a substantial fraction remains unclassified. When viral coverage is considered, however, the balance shifts. Coverage-weighted totals show lytic infections accounting for 47.8%, lysogenic infections rising to 23.2%, with the rest remaining unresolved. This shift indicates that while lysogenic associations are fewer in number, they often involve phages with higher coverage, suggesting their ecological weight and persistence may be disproportionately important compared to counts alone.

Viral pressure was highly uneven across hosts, with a small number of rep_bins carrying the majority of associations, while most rep_bins accrued fewer links. Notable examples include 3300009640_8 (143,272 links; mean 14.42 among 18 samples), 3300009618_8 (116,150 links; coverage 11.5), 3300009618_9 (73,616 links; mean 23.7 among 16 samples), 3300009641_8 (60,861 links; mean 20.5 among 21 samples), 3300009617_15 (49,308 links; mean 24.1 among 22 samples) and 3300018005_9 (23,089 links; mean 35.8 among 8 samples). These bins correspond to distinct microbial lineages, including Bacteria; Desulfobacterota; Syntrophia; Syntrophales; Smithellaceae, Archaea; Crenarchaeota; Bathyarchaeia, Bacteria; Desulfobacterota; Desulfobaccia; Desulfobaccales, and Archaea; Thermoplasmatota; Thermoplasmata_A (**Supplementary Data 3**), suggesting the central players in peatland carbon cycling serve as hotspots for viral infections and underscoring the importance of explicitly incorporating viral impacts into carbon cycle models.

4.0 Baseline integration and calibration link viral processes to host abundance dynamics

To examine how viral processes redistribute microbial abundances and metabolic capacity at the taxon-explicit and community levels, we built a quantitative baseline and viral pressure-aware calibration framework based on the 317 rep_bin*sample coverage estimates and parameterized by over 1.24 million virus-host linkages with 27 infected rep_bins. The calibration framework reproduces observed host abundances with high fidelity across infected rep_bins. When all infected rep_bins*sample were evaluated together (n=48), predicted totals closely matched observed abundances with a slope of 0.98, an intercept of 6.54, and an overall R^2=0.77 (Spearman's ρ=0.75, $p < 2.0 \times 10^{-16}$, **Fig. 1**). Breaking down by infection strategy revealed strong performance for lysogenic-only bins (n=14 rep_bin*sample), where the slope was 1.00, intercept 1.68, and R^2=0.99 (ρ=0.97). Similarly, lytic-only bins (n=12 rep_bin*sample) were well fit with slope 0.94, intercept 16.9, and R^2=0.70 (ρ=0.98). In contrast, the bins under both lytic and lysogenic viral infections (lytic-lysogenic, n=22 rep_bin*sample) were captured relatively less accurately, with a slope of 0.76, an intercept of 8.26, and a lower R^2=0.48 (ρ=0.17), suggesting that combined viral strategies introduce more variability into the calibration. This clear difference across infection types validates the need to model viral impacts in a strategy-specific manner, since lytic and lysogenic modes impose distinct and sometimes opposing pressures on their hosts. By explicitly separating these strategies within our framework, we ensure that viral effects are represented in a biologically realistic and analytically robust way, laying a stronger foundation for downstream simulations of community and ecosystem functions.

Estimated infected fractions across rep_bins are modest, but predictions consistently capture the expected decline in host abundance as lytic fraction increases. Infected fraction parameters spanned from 0.05 to 1.0, with the majority of bins clustering between 0.05 and 0.15, indicating a generally low but measurable viral burden across the communities (**Supplementary Data 4**). Only a few rep_bins carried higher fractions, most notably 3300009640_8, where the global estimate was $q_{\text{hat}} = 0.61$, but pooled values reached $q_{\text{eff}} = 0.88$ in samples with strong infection signatures, and 3300018005_9, which reached the upper bound of the calibration scale at $q_{\text{hat}} = 1.0$. The latter case is supported by very strong viral connectivity mentioned above, consistent with a lineage whose abundance is dominated by viral processes. Partial pooling is critical in these cases. For 3300009640_8, it tempers episodic sample-level estimates of $q_{\text{sample}} = 1.0$ toward the more moderate global fit, while for 3300018005_9 it anchors the estimate at the upper bound but still integrates across samples, preventing runaways. Sensitivity checks on baseline selection and regression stability further confirmed these outcomes. Together, these results indicate that for a small subset of bins with exceptionally high viral signal, the fitted parameters provide biologically meaningful upper-bound estimates of infected fraction, capturing both lineage-specific viral dominance and episodic infection dynamics rather than artefactual overfitting.

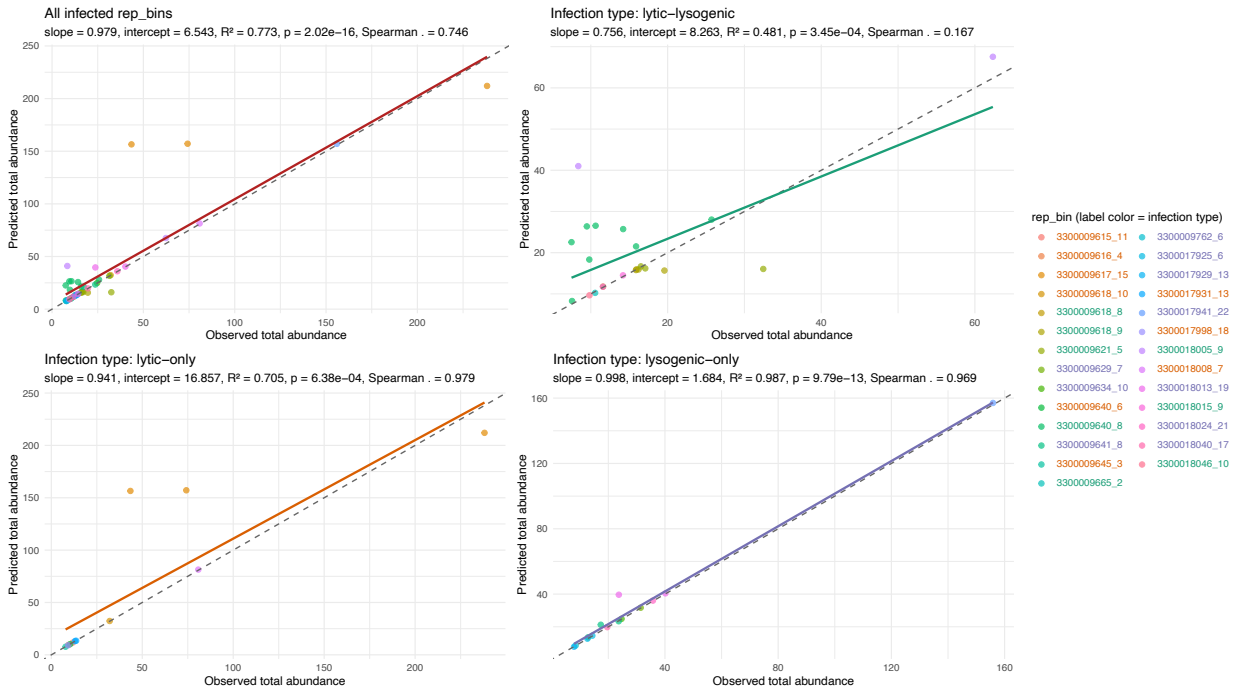

Figure 1. Observed-predicted plots of host abundance calibration.

5.0 Structured virus-host interaction scenarios show lysis drives turnover while lysogeny buffers persistence

The six simulation scenarios (S0-S6) span from baseline no-infection states to extremes of lytic and lysogenic dominance. S0 defines the null community without viral influence (p = 0, q = 0, g = 1.0), providing a stable reference. Lytic-skewed scenarios mimic kill-the-winner dynamics. S1 targets the most abundant bins within each sample, whereas S2 targets the most widespread bins across samples. Both impose high median lytic fractions (p of 0.61-0.65) with moderate infected fractions (q of 0.40) and minimal gain (g of 1.0-1.07). As a result, both scenarios predict biomass depletion and increased evenness, though through different ecological pathways. These outcomes are supported by depletions in Δ(S1/S2-S0) (**Fig. 2**).

Lysogeny-skewed scenarios retain host biomass and confer modest amplification, consistent with piggyback-the-winner dynamics. S3 reinforces the most abundant hosts, while S4 stabilizes the most prevalent microbial populations across samples. Both increase gain substantially (g of 1.49-1.80), operate under lower lytic pressure (p of 0.15-0.47), and elevate infected fractions (q of 0.50). These parameter regimes yield modest positive Δ values concentrated in already successful rep_bins (**Fig. 2**). Ecologically, these scenarios capture lysogenic benefits, where prophage-encoded functions (e.g., carrying auxiliary metabolic genes or AMGs, defense, stress responses) can modestly increase host fitness and persistence.

The extreme scenarios bound the space. S5 (maximal lytic dominance: p of 1.0, g of 1.0) maximizes top-down losses, producing the strongest biomass depletion across bins (**Fig. 2**). These outcomes align with strong viral shunt conditions, where winner depletion releases labile substrates, elevates turnover, and accelerates short-cycle respiration. In contrast, S6 represents maximal lysogenic gain with p of 0.2, g of 2.0, mirroring a lysogenic buffering extreme, where prophage integration and functional carryover stabilize host biomass, extend lineage persistence, and channel resources into long-term retention.

The divergent outcomes of lytic- versus lysogeny-dominant scenarios highlight how viral strategies redistribute microbial abundances in fundamentally different ways. Because K_{sim} couples lytic loss and lysogenic gain, lytic-skewed cases (S1/S2) depress K and predict stronger biomass depletion, particularly among dominant bins, yielding larger negative $\log_2$(Scenario/S0) shifts (**Fig. 2**). In contrast, lysogeny-skewed cases (S3/S4) modestly raise K and concentrate positive shifts in already successful lineages (**Fig. 2**). Together, these contrasts indicate that lysis predominantly trims dominant lineages and enhances turnover, while lysogeny reinforces successful hosts and promotes persistence, generating distinct and testable signatures in microbial community composition and ecosystem fluxes.

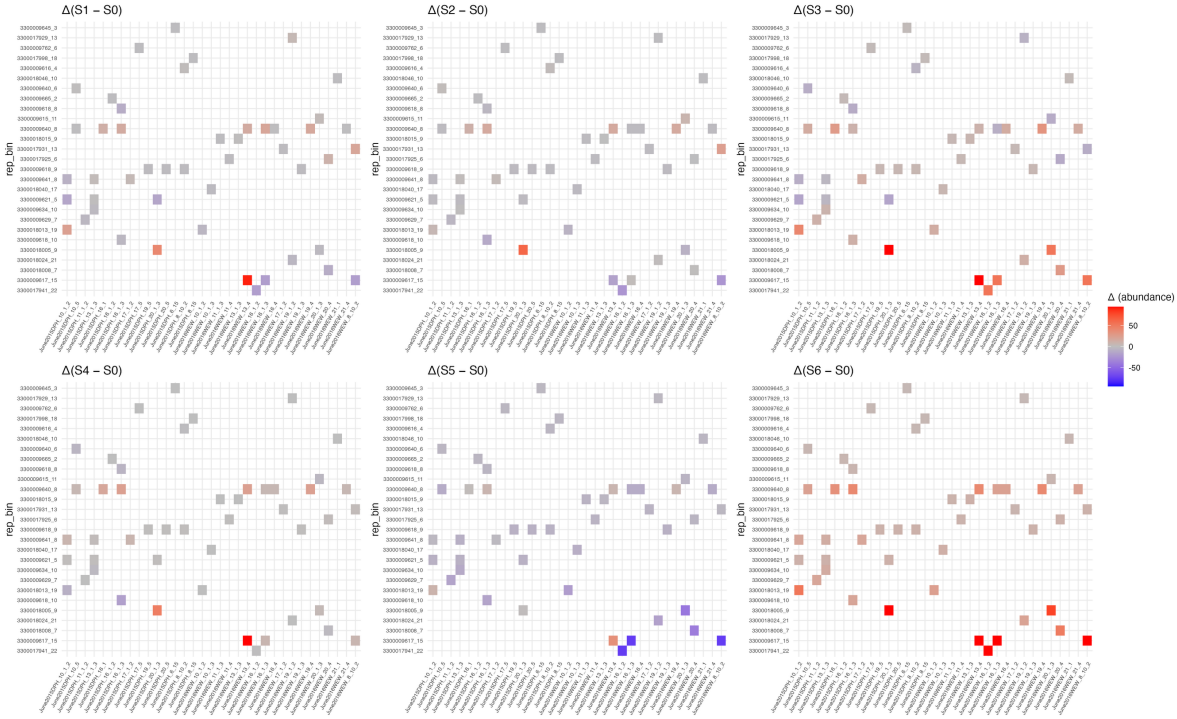

Figure 2. Relative heatmaps showing log2 ratios of scenario predictions versus the S0 baseline.

6.0 Viral strategies shift community metabolism from organic processes to inorganic fluxes

Virus-host interaction scenario perturbations induced measurable shifts in both growth and metabolite exchanges, with magnitude and direction varying by hypothesis. Community growth rates showed scenario-dependent changes relative to S0 (**Fig. 3**). Under Kill-the-Winner (S1-S2), the distributions were tightly centered (S1 median log2FC of -0.0037, interquartile range or IQR of 0.13; S2 median of -0.009, IQR of 0.081; **Supplementary Data 3 and 5**) indicating broad but modest effects, while some examples of suppression including the *Phycisphaerae* bin 3300009552_13 (S1 median log2FC of -0.69; S2 of -1.24). At the same time, *Acidimicrobiia* showed small gains rather than losses (e.g., 3300009616_4 in S1 with a median log2FC of +0.20, S2 with a median log2FC of +0.2625), underscoring that Kill-the-Winner does not uniformly depress all competitive lineages. By contrast, Piggyback-the-Winner scenarios broadened the responses. In S3, the spread increased (median log2FC of -0.082, IQR of 0.55), with strong winners such as the *Acidobacteriota* represented by 3300018015_9 (median log2FC of +0.49) and losers including *Planctomycetota* (3300009618_8 with median log2FC of -3.57; 3300009621_5 with median log2FC of -2.70; 3300009552_13 with median log2FC of -2.08). S4 remained centered (median with median log2FC of -0.0005, IQR of 0.0806) but exhibited strong declines in specific taxa (e.g., *Bathyarchaeia* 3300018002_8 with median log2FC of -1.88, *Phycisphaerae* 3300009552_13 with median log2FC of -0.86), consistent with occupancy-driven heterogeneity. Under high-lytic S5, responses broadened (median with median log2FC of -0.0052, IQR of 0.35) with marked decreases in *Actinobacteriota* and others (e.g., 3300009616_4 with median log2FC of -1.1930; 3300018015_9 with median log2FC of -0.7948). The high-gain S6 scenario produced the widest restructuring (log2FC with a median of -0.0488, ranging from -2.53 to 9.67; IQR of 0.4846), featuring strong increases in *Actinobacteriota* and *Acidobacteriota* (e.g., 3300018046_10 with median log2FC of +0.73; 3300018015_9 with median log2FC of +0.68; 3300009616_4 with median log2FC of +0.78) and pronounced decreases in *Phycisphaerae* and *Bathyarchaeia* (3300009552_13 with median log2FC of -2.36; 3300018002_8 with median log2FC of -2.53). Together, these scenario-specific patterns support a gradient from modest, diffuse adjustments (S1-S2) to heterogeneous stabilization (S3-S4) and, ultimately, extreme restructuring (S5-S6), priming downstream shifts in exchange fluxes and elemental cycling.

Exchange fluxes showed structured and scenario-dependent responses that scaled by the type of perturbation. Under Kill-the-Winner dynamics (S1 and S2), flux changes were relatively modest as well (**Supplementary Data 8**). S1 featured 1,828 imports and 726 exports above |log2FC| ≥ 0.30, while S2 had 1,544 imports and 638 exports, reflecting broad but shallow suppression of dominant competitors. By contrast, Piggyback-the-Winner scenarios (S3 and S4) produced more heterogeneous patterns (**Supplementary Data 8**). In S3, imports broadened to 2,399 reactions from non-infected taxa and 1,365 from virus-infected taxa, including large shifts in EX_4abut_e (4-aminobutanoate; log2FC of +18.2) and EX_fru_e (D-fructose; +21.4). Exports under S3 upregulated EX_4abut_e (4-aminobutanoate; +15.2) and EX_pacald_e (propionaldehyde; +18.7). S4 retained a narrower overall distribution but still produced high-magnitude responses, reflecting occupancy-based stabilization of some host lineages while disadvantaging others. Under extreme perturbations, the contrasts were more obvious. In the high-lytic S5 scenario, net exchange shifts increased in breadth, and in the high-gain S6 scenario, the system reached its maximum response with 4,810 imports and 1,732 exports above threshold (**Supplementary Data 8**). High-magnitude S6 imports included EX_g3pe_e (sn-glycerol-3-phosphate; +22.3), EX_asp_D_e (D-aspartate; +21.5), and EX_lyx_L_m (L-lyxose; +18.0), while dominant exports were EX_pacald_e (propionaldehyde; +20.9),

EX_pacald_m (propionaldehyde; +16.7), and EX_urea_m (urea; +16.1). Limma confirmed 60 significant import reactions in S6 ($q \leq 0.20$), and SHAP-based screening flagged additional high-impact features in S1 and S3. SHAP identified imports not detected by limma in S1 (n = 3) and S3 (n = 3), including EX_h2o_e (water) and EX_man_e (D-mannose) on the import side and, for exports in S1 (n = 5), EX_etoh_e (ethanol), EX_meoh_e (methanol), and EX_glc_D_e (D-glucose). This illustrates how ML surfaces sparse, influential reactions that pass FDR when modeled nonlinearly but not in other tests. PCA module analysis further captured coherent multivariate structure. For example, 109 import reactions in S3 and 113/122/115 export reactions in S2/S4/S5, respectively, were significant at $q \leq 0.20$, while consensus filters (≥2 screens) retained only the most reproducible effects, improving precision when prioritizing mechanisms for follow-up. Collectively, these patterns confirm that viral strategies scale from modest impacts under Kill-the-Winner, to heterogeneous but selective shifts under Piggyback-the-Winner, and more restructuring under extreme lytic or gain scenarios.

When aggregated to metabolite categories, scenario-specific responses converged on a small set of functional groups (**Supplementary Data 8, Fig. 4**). We automated the characterization of metabolites into biologically meaningful categories, leveraging KEGG, PubChem, and other identifiers for standard mapping while incorporating customized categories (e.g., Critical minerals, Ion, Gas, Organic acid) to capture processes not well represented in canonical ontologies. Under Kill-the-Winner (S1-S2), shifts were dispersed across many categories, consistent with the broad suppression of dominant taxa without strong functional specialization. In Piggyback-the-Winner (S3-S4), amino-acid and carbohydrate imports emerged as recurrent signals, e.g., EX_fru_e (D-fructose; import) and EX_4abut_e (4-aminobutanoate/GABA; import), while exports concentrated in organic acids and nitrogenous metabolites, reflecting viral stabilization of selected host metabolisms. High-lytic S5 shifted the balance toward broad carbohydrate uptake with organic acid overflow, consistent with host collapse and compensatory community metabolism. The most coherent roll-up appeared under High-gain S6, where imports concentrated in Ion (n = 6), Gas (n = 4), and Organic acid (n = 4) categories, and exports concentrated in Ion (n = 4), Gas (n = 4), and Organic acid (n = 3). Although conservative thresholds limited the number of formally significant categories, the median category signals remained stable and repeatedly highlighted carbohydrate and amino-acid processes as dominant in S3-S4, while S6 emphasized inorganic (ion/gas) processes.

Together, we found Kill-the-Winner strategies (S1-S2) produced diffuse, low-magnitude shifts across many categories, reflecting broad suppression without strong functional bias. Piggyback-the-Winner (S3-S4) concentrated changes in amino acid and carbohydrate imports with organic acid/nitrogenous exports, consistent with selective stabilization of host metabolisms. In contrast, high-lytic S5 emphasized carbohydrate uptake with organic acid overflow, a signature of increased viral replication and host lysis, while high-gain S6 shifted toward ion and gas fluxes, suggesting understudied virus-mediated inorganic processes. This collectively demonstrates how viral replication extends from host control to ecosystem-scale physicochemical processes.

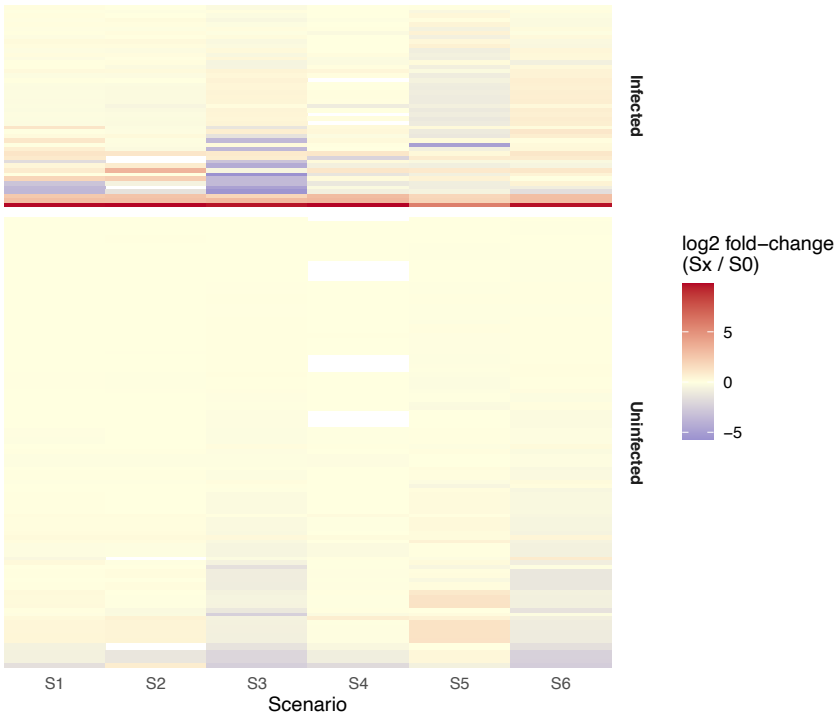

Figure 3. Scenario-dependent changes in taxon growth rates relative to baseline (S0).

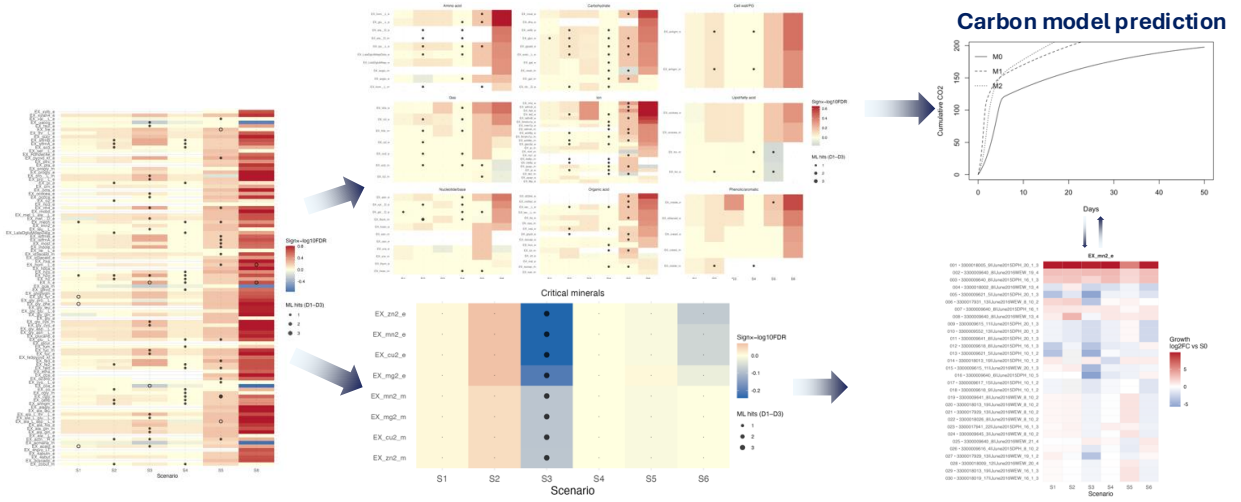

Figure 4. Bridge community metabolic model outputs with geochemical modeling.

7.0 Future implication: integrating category-resolved exchanges into carbon modeling

The developed multi-phase workflow makes it practical to feed category-resolved exchange shifts directly into carbon models. Across all scenario*direction, every reaction was retained during roll-up, with ~75% mapping to specific functional categories and the remaining ~25% assigned to Others (**Fig.4**). Our roll-up steps minimize information loss when rolling up reaction-level calls to process-level fluxes. Combined with our robust feature screening, this framework of integrating FDR-controlled statistics (limma, binomial) and ML-based multivariate screens (SHAP, LASSO, PCA) reduces false negatives, elevates weak and sparse but consistent biology, and yields interpretable flux changes at the category scale.

Growth rates directly supply μ values, with the upper quantile serving as scenario-specific μmax priors, while dividing biomass production by carbon uptake yields a proxy for biomass yield. Exchange tables, merged with annotations, define substrate and product stoichiometry, allowing construction of uptake pools (Vmax, Ks) and CUE estimates (CUE = 1 − respired C / uptake C). Although ATP maintenance fluxes are not directly captured in the community metabolic modeling outputs, effective maintenance can be approximated from CUE and yield proxies, while category-level flux shares approximate enzyme allocation. For example, Amino acid and Carbohydrate categories parameterize pool-specific uptake kinetics, Organic acids inform overflow pathways, and Ion and Gas categories capture proton and CO_2 fluxes critical for respiration terms. These outputs collectively provide the biomass yield, max growth, half-saturation, maintenance, allocation, and stoichiometric constraints needed for MEND-style models [9].

Building on this, rate laws can be structured around aggregation and pool-lumping of exchange fluxes into biologically meaningful categories, guild formation by clustering taxa into metabolic strategies (e.g., fast growers vs. fermenters/respirers), and setting priors from trait outputs. Viral impacts are then represented as top-down modifiers layered onto these rate law. For example, Kill-the-Winner reduces dominance of fast growers, lysis events redistribute biomass into DOM pools, and Piggyback-the-Winner and high-gain strategies alter trophic feedbacks by stabilizing resource fluxes. This framework thus enables virus-mediated microbial processes to be directly rolled into geochemical carbon models, linking scenario-driven microbial and viral dynamics with ecosystem-scale carbon cycling.

8.0 Use case: isolate viral impact on the import of critical mineral ions by specific microbial populations

Critical minerals are increasingly recognized as essential for national security and industrial supply chains, with their scarcity carrying both ecological and geopolitical implications (Congressional Research Service report). This makes them a compelling use case for evaluating virus-mediated impacts on microbial processes. By focusing on the Critical minerals category in our framework, we can isolate viral perturbations that directly affect the cycling of scarce and policy-relevant elements and then trace them down to the microbial populations that are responsible for such changes (**Fig. 4**).

In S3, the Critical minerals category encompassed 8 annotated imports (Zn^{2+}, Mn^{2+}, Cu^{2+}, and Mg^{2+}, each represented in both extracellular and intracellular compartments). These exchanges are biologically significant because they represent cofactors essential to enzymatic activity in photosynthesis, respiration, and stress response. On the export side, no critical mineral fluxes passed significance thresholds, highlighting the dominant role of uptake shifts under viral perturbation.

Coupled growth shifts help identify the likely microbial populations involved. In S3, strong declines were observed in *Bathyarchaeia* (3300018002_8, -2.53), *Phycisphaerae* (3300009552_13, -2.37), and *Desulfobacterota* (3300018018_15, -1.23), while increases occurred in *Acidimicrobiia* (3300009616_4, +0.78), *Acidobacteriota* (3300018046_10, +0.74), and Acidobacteriota (3300018015_9, +0.69). These shifts occur in tandem with altered uptake of Mg^{2+}, Mn^{2+}, Zn^{2+}, and Cu^{2+}, providing a tractable short-list for attributing mineral uptake roles to specific guilds under viral perturbation.

Because every exchange is category-labeled and cross-referenced with statistical and ML screens, this workflow enables isolation of the subset of mineral-relevant exchanges perturbed by viruses, quantification of their functional impact as category-level flux changes suitable for geochemical models, and trace-back to the microbial populations most responsible via concordant growth shifts and reaction ownership. Given the importance of critical minerals, this approach offers a promising path for connecting microbial ecology and viral strategies to decision-relevant modeling of resource cycles, with immediate potential to extend the framework to other scarce resources.

9.0 References

1. Swift, R.S., *Sequestration of carbon by soil.* Soil science, 2001. **166**(11): p. 858-871.
2. Jansson, J.K. and R. Wu, *Soil viral diversity, ecology and climate change.* Nature Reviews Microbiology, 2023. **21**(5): p. 296-311.
3. Jansson, J.K. and K.S. Hofmockel, *Soil microbiomes and climate change.* Nature Reviews Microbiology, 2020. **18**(1): p. 35-46.
4. Carreira, C., et al., *Integrating viruses into soil food web biogeochemistry.* Nature Microbiology, 2024: p. 1-11.
5. Liang, X., et al., *Incorporating viruses into soil ecology: A new dimension to understand biogeochemical cycling.* Critical Reviews in Environmental Science and Technology, 2024. **54**(2): p. 117-137.
6. Guidi, L., et al., *Plankton networks driving carbon export in the oligotrophic ocean.* Nature, 2016. **532**(7600): p. 465-470.
7. Wu, R., et al., *Targeted assemblies of cas1 suggest CRISPR-Cas's response to soil warming.* The ISME journal, 2020. **14**(7): p. 1651-1662.
8. Kuzyakov, Y. and K. Mason-Jones, *Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions.* Soil Biology and Biochemistry, 2018. **127**: p. 305-317.
9. Jagadamma, S., et al., *Organic carbon sorption and decomposition in selected global soils*. 2014, ORNLTESSFA (Oak Ridge National Lab's Terrestrial Ecosystem Science).

Appendix A – Data and code availability

Raw and processed data including the supplementary data are stored on PNNL research computing cluster under the project name of “virus_ldrd” and the workflow is documented at PNNL Gitlab (https://tanuki.pnnl.gov/ruonan.wu/fy25_ldrd_virus).

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov