MODELING ECOLOGICAL DYNAMICS VIA MULTI-AGENT SYSTEMS OF MICROBIAL METABOLISM

Anonymous authors
Paper under double-blind review

ABSTRACT

Metabolic interactions govern the assembly and dynamics of microbial communities. These interactions are studied computationally through genome-scale metabolic reconstructions in which an organism's biochemical reaction network is assembled from annotated genomes and curated databases. The capabilities of each organism are then predicted by optimizing flux through the resulting metabolic network using flux balance analysis (FBA) over a particular phenotype (growth, for example).

In this work, we asked whether a language model can replicate the predictions of a genome-scale metabolic model, with each organism acting as an agent that chooses its exchange with its environment, which represented by the model context.

We first show that prompted single microbial agents were able to recover key metabolic capabilities quantitatively, albeit at a lesser precision than dynamic flux balance analysis (dFBA), over multiple query rounds. Subsequently, by fine-tuning an open-source model on flux envelopes drawn from 897 reconstructions, microbial agents outperformed metabolic models in predicting metabolite concentrations. Our analysis showed that fine-tuning produces valid flux distributions that respect the underlying stoichiometry and effectively turned the language model into a reliable solver of metabolic fluxes under constraints.

Then, by scaling to bacterial communities, multi-agent systems correctly captured competition as the main driver of interactions in communities, and correctly modeled smooth network dynamics that transition through building, converging, and stabilizing of the community structure over time, which DyMMM (Dynamic Multispecies Metabolic Modeling), a framework based on dFBA, fails to capture. To do so, we built an observability and evaluation system that was derived from ecological theory and adapted to the multi-agent model setting.

Taken together, we present a new framework for modeling microbial communities relevant to bioengineering and biomedicine by using multi-agent large language model systems constructed with domain-specific harnesses. We also propose fine-tuning on metabolic knowledgebases as a strategy to build accurate modeling tools for the prediction of intervention effects in microbial communities and for the design of synthetic communities.

1 INTRODUCTION

Bacteria are ubiquitous in our environment and are central to human health and disease. They are organized into microbial communities that colonize the gut, the skin, and the soil. The composition of the gut microbiota is particularly implicated in inflammatory disease, in infection, and in the metabolism of drugs. Beyond the study of naturally occurring communities, there is increased interest in synthetically engineered consortia in bioengineering to produce chemicals and to degrade waste. Therefore, understanding how such communities arise, how they stabilize, and how they interact is a prerequisite for designing therapeutic and industrial interventions.

Interactions between species are mainly driven by metabolic exchanges. Bacteria sharing an environment are coupled through the metabolites they secrete and take up, leading to cooperation and competition, and the emergent behavior of the community as a whole becomes the aggregate of those interactions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Computationally, a community is studied through the metabolic networks of its members. For each organism, the biochemical reactions it is known to carry are surveyed from curated knowledgebases and assembled into a network, represented by the stoichiometric matrix $S \in \mathbb{R}^{m \times n}$. The rows of this matrix are the m metabolites participating in the n biochemical reactions, with S_{ij} the stoichiometric coefficient of metabolite i in reaction j. For a reaction r_1 that converts A into $B + C$, the column of S for r_1 holds $S_{A,1} = -1$, $S_{B,1} = S_{C,1} = +1$, and zero in every other row. Since reactions involves only a few of the thousands of metabolites an organism carries, S is very sparse.

Given S, the metabolic capabilities of an organism are studied by solving for the flux $v \in \mathbb{R}^n$ (mmol/hr) carried by each reaction. The rate of change of the internal metabolite concentrations is $\mathrm{d}\mathbf{c}/\mathrm{d}t = Sv$, which we assume to be at steady state, so that $\mathrm{d}\mathbf{c}/\mathrm{d}t = 0$ and therefore $Sv = 0$. Adding that each reaction is bounded by thermodynamics constraints (such as direction of flux) and by nutrient availability, flux balance analysis (FBA) selects the flux vector maximizing a biological objective c, which is usually biomass production.

$$\max_{v \in \mathbb{R}^n} \; c^\top v \qquad \text{subject to} \qquad \underbrace{Sv}_{\mathrm{d}\mathbf{c}/\mathrm{d}t} = 0, \qquad \ell \le v \le u. \tag{1}$$

By applying FBA repeatedly over a changing medium for a whole time horizon, this yields dynamic FBA (dFBA), the standard tool for predicting dynamic metabolic behavior, provided by solving the linear program in Equation 1.

In this work, we propose to study microbial communities as a system of language-model agents, with each species represented by one agent. The premise is that S is itself a reconstruction, with every reaction in it was reported in the literature and then curated from a knowledgebase. Therefore, a language model trained on that same literature can carry the information needed to state what an organism can consume, what it secretes, and whether it grows, and how can it interact with other species. In addition to testing single-species agents, and by leveraging recent advances in multi-agent systems, we developed language model agents for microbial communities by combining single-species agents, testing them against metabolic networks for the same species and against ground truth measurements from cultures of the same organisms in defined nutrient media.

2 Related work

Multi-agent systems built from language models have been applied across biology. They have been proposed as a general instrument for biomedical discovery (Gao et al., 2024), used to design molecules that were then validated at the bench (Swanson et al., 2025), and evaluated as collaborators on discovery tasks over multi-omic cancer data (Johri et al., 2026). The agents take turns reasoning in natural language about a scientific problem, passing arguments between them.

In our setting each agent instead stands in for one organism, which is compared against the standard constraint-based metabolic model (Equation 1) of that organism's metabolism. Flux balance analysis and the family of methods around it are long established (Orth et al., 2010; Heirendt et al., 2019). Even though LLMs have been used to improve modeling and reconstruction of metabolic networks (Luo et al., 2026; Chowdhury et al., 2026); what we propose here is to train LLMs for metabolic tasks, and to leverage recent advances in multi-agent systems as a framework to study the dynamics of microbial communities.

The link between metabolic networks and LLMs is that the stoichiometric matrix of a metabolic model (Equation 1) is first built from a knowledgebase of metabolic reactions. Bacterial genomes are annotated and mapped onto a curated reaction database, then gap-filled and manually verified by experts. These networks, therefore, represent the sum of metabolic knowledge about a particular organism and can guide post-training of language models. It is known, for example, that metabolic models of well-studied organisms in the literature have higher quality and predictive accuracy than those of less-studied organisms.

Because flux balance analysis is a linear program, training a language model on a metabolic network also connects to work on learning to solve constrained optimization. Previous work has shown that solving mixed-integer programs can be aided by neural networks (Nair et al., 2020). This suggests that large language models can potentially absorb metabolic models (Equation 1) through a well-calibrated post-training procedure (Section 4.1.2).

3 METHODS

3.1 GROUND TRUTH DATA

To compare model predictions with experimental data, we used data from well-defined bacterial mono and co-cultures (Clark et al., 2021) for 25 species (Appendix B), grown in DM38, a chemically defined medium, giving 212 clean wells with 3 to 22 replicates per species. Each well reports growth, measured as optical density at $600\,\text{nm}$ (OD600) and used throughout as the proxy for biomass, together with the concentrations of acetate, butyrate, lactate and succinate after $48\,\text{h}$ of culture.

3.2 BENCHMARK PROCEDURE

We first tested our setting on single-species cultures. We compared models of bacterial growth by testing their predictions against the ground truth data set. For LLM-based agents, the context represents a single well-stirred patch of $1\,\text{mL}$ to mimic a deep-well plate. The time horizon is fixed at $48\,\text{h}$ and divided into K steps/rounds. To simulate rounds of $4\,\text{h}$ each, we set $K = 12$. In the initial step, bacteria are inoculated at equal uniform density. Then in subsequent steps, the agent receives a rendered description of its chemical environment, proposes exchange fluxes and a growth rate, and a validator verifies the proposal for physical consistency, metabolite balance across the proposed fluxes.

We also compared against predictions from metabolic models ran through FBA. These models are solved for the whole time horizon at every integration substep, which is referred to as dFBA in the static optimization (SOA) implementation (Mahadevan et al., 2002).

We then tested our setting in communities of bacteria sharing one metabolite pool. For LLM-based agents, we placed every member of a community in one well-stirred flask and modeled each species as its own agent. Each agent sees the pooled metabolite concentrations, and makes decisions, that is, uptake/secretion of metabolites, based on its environment and on internal capacities and state. Communication between agents therefore occurs through metabolite exchange. For metabolic models, we used the Dynamic Multispecies Metabolic Modeling (DyMMM) framework (Zhuang et al., 2011) to model communities dynamically using metabolic models of bacteria, which represents the single model dFBA counterpart for community models.

3.3 METABOLIC MODELS

We employed dFBA to model the metabolic behavior for single species. dFBA obtains species i's exchange fluxes $\mathbf{v}_{i,t}$ by solving Equation 1 once per step with that species' S_i and c_i, and with lower bounds $\boldsymbol{\ell}(\mathbf{m}_t, X_{i,t})$ set by what is in the flask, where $\mathbf{m}_t \in \mathbb{R}^M$ holds the concentrations of the M metabolites at step t and $X_{i,t}$ is the biomass of species i. The horizon $T = 48\,\text{h}$ is divided into K steps of $\Delta t = T/K$. We used metabolic models from the current release of AGORA2 (Heinken et al., 2023), a resource of genome-scale metabolic reconstructions for human gut microorganisms. The four flux balance formulations we rely on are stated in (Appendix A).

3.4 LLM-BASED SPECIES AGENTS

We used Llama-3.3-70B, Llama-3.1-8B and Qwen2.5-72B as open-weight base models, together with Claude Haiku 4.5 as a closed-weight comparison, and ran them in two tiers. In the first tier the models are only prompted. Each agent is given the name of its organism, a description of the flask and the composition of the media, and the instruction to commit to a metabolic strategy for the coming interval. In the second tier, we fine-tuned an open-weight model, Llama-3.1-8B, on flux vectors produced by microbial metabolic models to instill metabolic knowledge to the LLM and improve its predictive power. We then compare Llama-3.1 before and after fine-tuning for a set of metabolic tasks. Let ρ_i render the flask as text and let g_θ be a language model with weights θ. An agent's flux decision is

$$\mathbf{v}_{i,t} = \Pi(\, g_\theta(\rho_i(\mathbf{m}_t, X_{i,t}))) \,, \tag{2}$$

where Π is a validator that detects feasible flux proposals. The agent has no access to S_i, so whatever it knows about species i is carried from internal knowledge in θ.

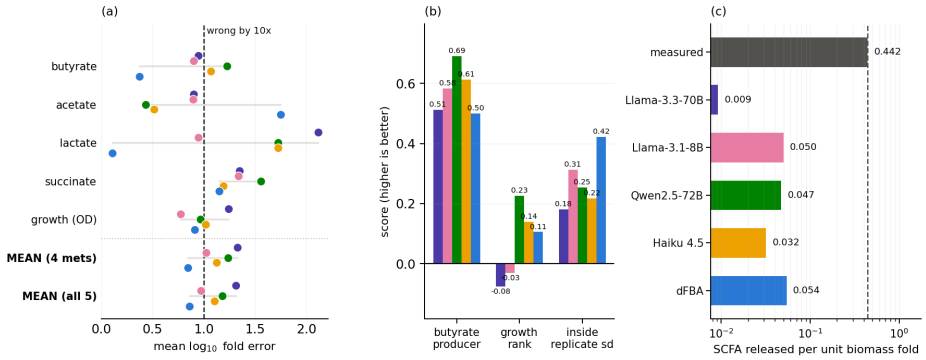


Figure 1: Performance of single-species metabolic agents and models on ground truth measurements. **(a)** measurement error for each model on ground truth data. **(b)** additional scores derived from the same measurements. **(c)** carbon allocation as short-chain fatty acid released per unit of biomass fold, against the measured value as a dashed line.

3.5 Modeling temporal dynamics

For both the LLM-based species and the classical metabolic model, the environment is updated after each time step/round,

$$\mathbf{m}_{t+1} = \mathbf{m}_t + \Delta t \sum_{i=1}^{N} X_{i,t}\, \mathbf{v}_{i,t}^{\mathrm{ex}}, \qquad X_{i,t+1} = X_{i,t}\, e^{\mu_{i,t}\Delta t}, \tag{3}$$

This setting corresponds to the SOA implementation for dynamic metabolic modeling. In the multi-species community setting and because LLM-based agents have no access to other agents' states, no agent knows other agents' biomass or intended fluxes, and the interactions are only determined over a shared pool. Competition and cross-feeding are therefore induced by Equation 3, which allows estimating interaction networks.

4 Results

4.1 Single-agent models

We first started by evaluating single-species agents by comparing their predictions against a measurement of that organism in a defined media. We constructed single-species cultures for the 25 bacteria present in a ground truth data and only exposed to the composition of DM38, the nutrient media used in the experiments, which serves as input to the model. The agent harness was built on i) defining species capabilities through the LLM's internal knowledge and ii) modeling the environment through its context. We evaluated how well a language model recovers the known metabolic capabilities of the 25 gut species. Each agent is given the composition of DM38 and asked what its organism consumes, what it secretes and how fast it grows, and the resulting endpoint concentrations are scored against the measured wells and against the predictions of the metabolic networks for the same organisms in the same medium.

4.1.1 Prompted agents

At a first stage, we evaluated prompt-based predictions. Each prompt states the rules of the world, an environment report for the current turn listing the available nutrients in mmol L^{-1} together with the imposed conditions and the host environment, and the decision the agent must return. The agent is asked to reason in prose and then to end its reply with a single JSON object giving the exchange fluxes it will run and the growth rate it claims. For each model, and for 12 time step over the 48 hours of the experiment, we scored model predictions against the measured endpoints, which include final metabolite concentrations for 4 metabolites (butyrate, acetate, lactate, and succinate) and

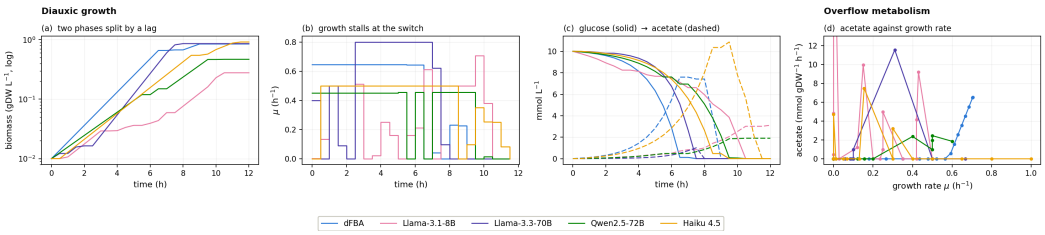


Figure 2: Recovery of diauxic growth and overflow metabolism for each model. Panels (a) to (c) represented diauxic growth, showing log biomass, growth rate, and the substrates with glucose solid and acetate dashed. Haiku 4.5 satisfies the diauxic signature, visible in (c) as the only acetate curve that rises high and then returns to zero, and dFBA satisfies it too. Panel (d) is overflow metabolism plotted against growth rate. Only dFBA reproduces the overflow elbow. Scorecards for both phenotypes are in Appendix E.

each species' final abundances (Table 1, with per-metabolite errors in Appendix C). We found that, encouragingly, LLM-based model species were able to recover many of the known metabolic capabilities, demonstrating the potential of language models in predicting microbial metabolism. But they failed to outperform the traditional metabolic network reconstructions in terms of predictive accuracy. We also find that trajectories of LLM agents prompted over turns are smooth rather than erratic, with biomass and metabolite curves that rise and saturate (Appendix D).

However, we want to highlight that as published, all gut bacteria metabolic models we used (25/25) grow on reference human diets but only 1/25 grew on DM38, even though experimental evidence in ground truth data shows growth. We have therefore gap-filled these draft reconstructions including adding metabolite transport reactions using knowledge from the literature about these species, taking DM38 growth from 1/25 to 25/25, as reported by experiments. This curation informed our next steps, because it showed that metabolic models are i) a condensed representation of the literature and ii) a model's quality depends directly on the completeness of the literature it ingests.

To further evaluate the models and test the reconstructions accuracy, we tested two behaviors that bacteria are known to display and that require precise quantitative knowledge of metabolism.

The first behavior is diauxic growth, that is, when a bacterial species is offered two carbon substrates consumes the preferred one first and switches to the less preferred one only when the first is exhausted. We have tested our metabolic models and LLM-based agents on this behavior. This behavior is defined by Escherichia coli aerobic growth on glucose as the sole carbon source, then switching to acetate when glucose is exhausted, exhibiting a two-phase growth curve with a lag between them. The signature, is two exponential segments split by a plateau, with $\mu_2 < \mu_1$. Metabolic models were able to model the diauxic growth curve ($0.856\,\mathrm{g\,L^{-1}}$), consumed the same acetate, and made 7.0% of their final biomass after glucose is entirely consumed. Whereas only one prompted model (Haiku 4.5) reproduced the phenotype and did about 40% of its final biomass after the glucose is depleted.

A second peculiar phenotype is overflow metabolism, in which E. coli excretes acetate in the presence of oxygen. Its characteristic signature represents an elbow-shaped curve of acetate excretion against growth rate, which is flat at zero up to a critical rate and then rises linearly (Basan et al., 2015). Metabolic models were able to reproduce the elbow-shaped curve, with acetate exactly zero up to μ_c and then linear with slope $\approx$53 at every cap. However, no agent produces this complex behavior.

These results highlight that the amount and quality of internalized knowledge maybe a leverage in building the next-generation of bacterial metabolism models because i) refining metabolic models with literature-based gap filling has unlocked their full functional capabilities and because ii) larger LLMs (Haiku-4.5) were better able to recapitulate complex behaviors.

4.1.2 Fine-tuning on genome-scale metabolic models

Previous experiments showed that metabolic models carry signal the prompted agents do not include. It recovered both of the particular phenotypes above, and across the measured endpoints it was the

Table 1: Model performance against ground truth data for single strains. Here bal. acc. stands for balanced accuracy, ρ for the Spearman rank correlation, and sd for standard deviation.

metric	Llama-3.3-70B	Llama-3.1-8B	Qwen 2.5-72B	Haiku 4.5	dFBA	Llama-3.1-8B fine-tuned
butyrate call (bal. acc.)	0.512	0.583	**0.691**	0.613	0.500	0.500
growth rank (ρ)	−0.076	−0.030	0.225	0.139	0.105	**0.325**
growth fold ($\log_{10}$)	1.243	0.772	0.969	1.015	0.911	**0.614**
metabolites ($\log_{10}$)	1.329	1.021	1.234	1.126	0.845	**0.748**
inside replicate sd	0.181	0.313	0.253	0.217	**0.422**	0.398
clean mass balance	167/300	90/300	**173/300**	165/300	–	**173/300**
no usable answer	5/300	16/300	2/300	30/300	–	**0/300**

more accurate of the two on most quantities. As stated previously, a metabolic model is also the literature written in a mathematical format, since every reaction in S was placed there because it was reported and then curated. Therefore, training a language model on a metabolic network can be a way of distilling that curated literature knowledge.

4.1.3 Corpus

To fine-tune LLMs for metabolic functions, we constructed a corpus covering 897 gut strains across three input diets (Western, Mediterranean, and High-Fiber), with each strain and diet pair sampled at the base medium and again under single-nutrient dropouts, which remove one nutrient component from the diet and leaves everything else fixed. We then ran flux variability analysis (FVA), which predicts intervals of secretion/uptake to estimate and record what the organism can do along with the point estimates from pFBA, the L1-regularized implementation of FBA.

To evaluate our strategy, we held out evaluation organisms, which are excluded by species, not by strain, to avoid data leakage. Beyond holding out the 25 evaluation strains present in the ground truth data, we dropped 67 strains sharing a genus–species with them (8.3% of the data). We also did not include DM38 in our training data set. So neither the organism nor the medium is one the model was fine-tuned on.

4.1.4 Training

We trained Llama-3.1-8B-Instruct with low-rank adaptation (LoRA) at r=32, α=16 and dropout 0.05, for one epoch at learning rate 2×10^{-4} on a cosine schedule with an effective batch of 64, taking the loss on the assistant turn only. Validation loss was minimal at step 300 of 843 (0.1199) and rose 5.8% by the end.

4.1.5 Result

The single-species fine-tuned bacterial model had better overall performance than either metabolic or prompted agents, with growth ranking 0.325 against 0.105, while producing well-formed chemistry at the same rate as the best prompted model (57% clean) and being the only model with no unusable answers. In other words, fine-tuning made the LLM model's output of exchange fluxes able to satisfy the constraints of the metabolic network's linear program. In fact, the realizable fraction of these LLM fluxes rose from 0.868 to 0.928 with respect to the base model Appendix F. Therefore, training an open-weights model to emit the output of a constrained optimization over the stoichiometric matrix of metabolic network, has trained the model to solve a linear program.

4.2 Multi-agent bacterial communities

Having shown that LLMs provide meaningful representations of microbial metabolism by combining biochemical stoichiometry with reasoning on literature over complex metabolic phenotypes, we now turn our attention to modeling microbial communities. Because our ground truth data contains communities of multiple bacteria ranging from n=2 to n=25 (all species), we therefore evaluate on

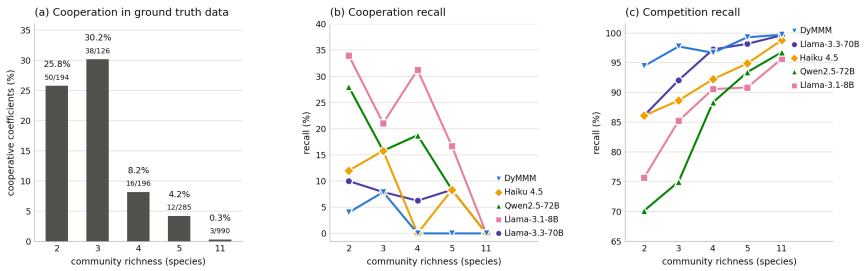


Figure 3: Competition and cooperation in multi-agent bacterial communities. **(a)**, cooperation is subject to large class imbalance as n increases. **(b)** and **(c)**, recall on each interaction type, with competition dominating interactions. Counts in (a) give the cooperative coefficients over all coefficients at each richness, and because only 3 of 990 are cooperative at richness 11, the zeros in (b) at that richness rest on 3 coefficients.

the same outputs as before, using the final metabolite concentrations, and the final relative abundances of each species. We modeled each organism by one agent, and coupled species only through the shared metabolite pool, where one language model stands for one bacterium, and that model's context window is that bacterium's environment. Everything the organism can sense is captured in the context, so the model plays the part of the cell's internal metabolism, deciding what to take up and release, while the context plays the part of the media. A community is then n models sharing one pool, and the only channel between any two of them is a metabolite that one released and another can reach. Therefore, ecology appears only through biochemical interactions.

In our benchmark, we ran the same 3 LLMs, a fine-tuned version, and a metabolic community model using the media composition as input. To detect interactions (competition or cooperation), we scored final composition and the pairwise interaction coefficient $\varepsilon_{i|j} = a_{i|j}/a_{i|\emptyset}$, the final abundance $a_{i|j}$ that species i reaches beside partner j divided by its abundance $a_{i|\emptyset}$ in monoculture. Values above 1 are cooperation and values below 1 are competition.

We found that over 12 turns of 4 hours each, multi-agent models of bacteria were able to establish complex interaction networks, with competition dominating the interactions as expected. Cooperation, however, was much rarer and more difficult to predict as the number of participants grows.

Because traces of reasoning for each model are recorded in every turn, interaction networks can be derived from them at each round, to identify who fed whom, and who was rationed on a metabolite another member took a larger share of. Using n=5 as an example, we tracked the assembly over time for both llama and Haiku. We find that LLM-based bacterial communities assemble from nothing over a run, reaching 4.5 to 11.3 cross-feeding edges per turn at richness 5 and holding that pattern over a few rounds.

However, models of metabolic communities using DyMMM, a framework comparable to dFBA for multiple species, reached a comparable density but only after 2 time steps and then reached a non-steady state equilibrium (Figure 10). Taken together, these simulations show that LLM-based bacterial communities not only produce valid endpoints with respect to experimental data but can also help reverse engineer the intermediate steps that build the interaction network. As such, they can be a valuable tool for understanding the construction and dynamics of microbial communities.

4.2.1 Ecology-driven evaluation of multi-agent bacterial interactions

In building multi-agent systems for bacterial communities, it is crucial to evaluate metrics and structures derived from domain knowledge. We have scored our models against measurements, including which species wins, what share it takes, and how large ε is. However, to certify our trajectory modeling process, and because this process is under-determined, multiple competing or implausible constructions can lead to the same final result.

We therefore ask a different question, whether a model obeys a constraint that real communities are known to satisfy, independently of whether it gets the abundances right.

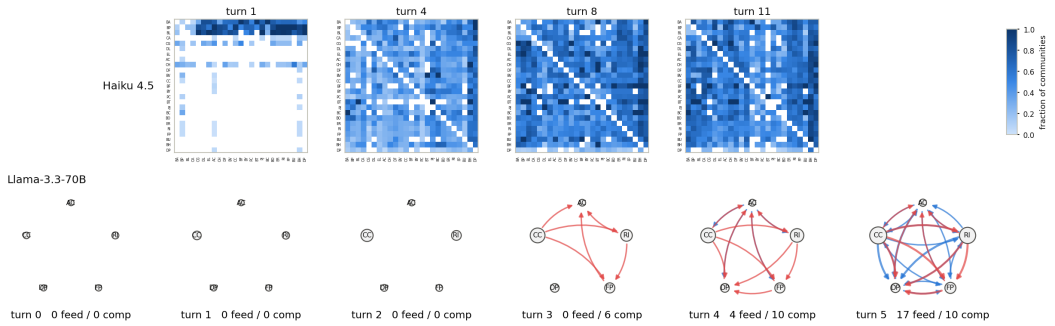


Figure 4: Assembly of cross-feeding and competition in two agent-based models. **Top**, Haiku 4.5 cross-feeding as adjacency matrices at four turns over the 90 $n = 11$ communities, rows releasing and columns consuming, shaded by the fraction of communities carrying that transfer. **Bottom**, Llama-3.3-70B on one $n = 5$ community of AC, CC, DP, FP and RI (Appendix B) over the first six turns. Blue $A \to B$ means B consumed a metabolite A released last turn, and red $A \to B$ means a metabolite was rationed for B while A held the larger share of demand for it. Under each panel, "feed" and "comp" count the feeding and competition edges. Competition appears first, at turn 3, and cross-feeding at turn 4 (Appendix G shows every turn). Node size is proportional to relative abundance.

Table 2: Performance of multi-agent bacterial communities on ground truth data for $n = 11$ to 15, on the 232 communities with replicate wells. Spearman ρ and top-1 are means over communities with bootstrap 95 % intervals, and the error column is the median over species of $|\log_{10}(\text{predicted}/\text{measured})|$.

model	n	Spearman ρ	top-1	median $\log_{10}$ error
Llama-3.1-8B fine-tuned	232	**+0.195 [+0.169, +0.222]**	9.5% [6.0, 13.4]	**0.631**
Llama-3.1-8B	232	+0.058 [+0.021, +0.095]	12.5% [8.6, 16.8]	1.508
Llama-3.3-70B	232	+0.186 [+0.149, +0.222]	**16.4% [11.6, 21.1]**	0.821
DyMMM	232	−0.041 [−0.066, −0.015]	3.4% [1.3, 6.0]	1.354

The constraint we propose is the one established for the human microbiome (Bashan et al., 2016), by studying the overlap between the share of abundance that the species both contain and the dissimilarity in the Jensen–Shannon distance of the root between the two final abundance vectors restricted to those shared species and renormalized. In real communities, dissimilarity falls as overlap rises, because organisms that share species converge on shared composition.

We restricted our testing to communities of 11 to 15 members because this is the only window in which every pair is present and for which overlap still varies, and where the ground truth data shows the constraint cleanly (-0.284). As a control, we constructed a shuffled null in which compositions are permuted between communities of equal richness. This destroys the link between species content and abundance while preserving everything else, to prevent that a negative slope can arise from sampling alone (Kalyuzhny & Shnerb, 2017).

First, we found that our ground truth data satisfies the constraint. Ground truth measured communities give a slope of -0.284 over the band, while their own shuffled null sits on it at $+0.002$. We also found that DyMMM communities ($+0.111$) do not satisfy the constraint.

Next, we found that LLM-based agent communities satisfy it for the whole 48 h run. Llama-3.3-70B gave -0.348 and Llama-3.1-8B -0.374, against nulls of $+0.155$ and $+0.119$. Because every community begins from the same initial condition, with equal inoculum, any structure observed at the endpoint must arise from the dynamics rather than from the starting composition. The composition at $t = 0$ is uniform and pairwise dissimilarity is identically zero.

In other words, the agents were able to recover is the structure of the ecology. We also show that they are able to capture accurate metabolite concentrations and final relative abundances, particularly for the fine-tuned model on metabolic network reconstructions (Table 2).

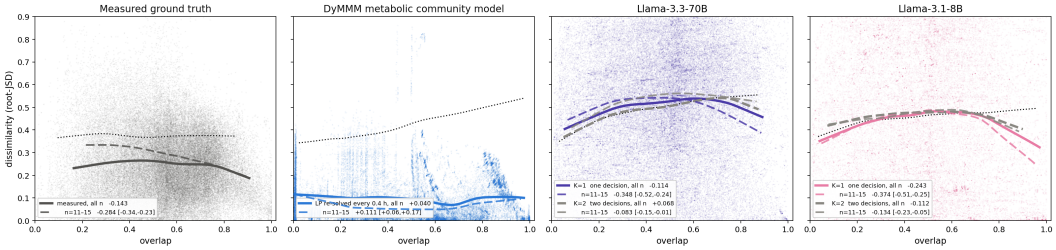


Figure 5: Dissimilarity–overlap at communities of size 11 to 15. A curve falling at high overlap is the universality reported for real communities (Bashan et al., 2016). Dotted is the shuffled null, the same data with compositions permuted between communities of equal richness. Ground truth data satisfies the constraint and DyMMM does not. Both language models satisfy it over a 48-hour run. Slopes carry bootstrap 95 % intervals over communities, with self-pairs excluded.

5 Discussion

Genome-scale metabolic networks can predict bacterial metabolism accurately, and they have been the standard method in bioengineering and in biomedicine. They extend naturally to communities because metabolic interactions shape community compositions. Therefore, models of metabolism can represent community assembly from the ground up.

Metabolic model reconstructions are also a form of compression. Every reaction in S, and every bound on it, comes from genome annotation or the literature and is curated into a database, so a reconstruction is that species' genetic and literature knowledge written as mathematics. A language model compresses an overlapping body of knowledge, acquired from the literature directly rather than through curation, so the two models can have exchangeable parts.

In the first direction, we show that the published reconstructions were incomplete with only 1 of 25 AGORA2 models grew on DM38 as published, even though all 25 organisms experimentally grow in that medium. We debugged the reconstructions until all 25 grew, by annotating missing transport reactions from the literature and by gap filling missing reactions. This gap highlights the sort of knowledge that can be missing in draft reconstructions.

In the other direction, fine-tuning a language model on a metabolic model improved its ability to recapitulate metabolic capabilities of bacterial strains. The metabolic model here aggregates literature knowledge about microbial metabolism, which improved the language model's performance on metabolite concentrations, highlighting the synergy both modeling types can achieve.

Training an open-weights model to emit the output of a constrained optimization over a stoichiometric matrix is, in effect, training it to solve a linear program. Previous work (Nair et al., 2020) was able to solve mixed integer programs, at a scale comparable to ours, using deep neural networks.

Our benchmarks only included bottom-up models. Rather than fitting a community's behavior from measurements of communities, we put the knowledge in the individual agents, coupled them only through the metabolites they exchange, and let the behavior of the whole emerge. As a result, absolute accuracy was low, with model performance around Spearman ρ near 0.3 against measured composition (for n=12 communities). A statistical model trained on these communities, which maps species' genetic features to community function shows high predictive power in cross-validated splits (Qian et al., 2026), for metabolite concentrations, and the same study reports correlations between -0.09 and 0.30 for a community constraint-based model, in line with our observations. An additional limitation is that our community models were sensitive to how many times agents were consulted within the fixed time horizon.

Taken together, we propose multi-agent systems of microbial metabolism as a new tool to model community formation and emergent behaviors, with immediate applications to health and biotechnology. Future work will build computational infrastructure for modeling microbial communities at scale, by fine-tuning species-specific LLM models, to enable discoveries in biology and medicine.

Reproducibility statement

The prompts and the code for model training and evaluation are released with the paper, along with community lists and the run records each figure is computed from. The prompted open-weight models (Llama-3.3-70B-Instruct, Llama-3.1-8B-Instruct and Qwen2.5-72B-Instruct) were queried through the Hugging Face inference API, and Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) through the Claude API. The fine-tuned model was run from the Hugging Face release of Llama-3.1-8B-Instruct with our LoRA adapter. Section 3 specifies the medium, the integration step and the substep, the uptake kinetics and the carbon bound. For fine-tuning, Section 4.1.2 gives corpus generation and the training configuration. Metabolic models can be downloaded from AGORA2 and gap-filling steps for metabolic models are provided as well.

Ethics statement

This work uses published measurements of bacterial cultures (Clark et al., 2021) and public genome-scale reconstructions. It involves no human subjects, no personal data and no new laboratory work. The application domain, predicting how gut bacterial communities assemble, is directly relevant for improving human health and does not pose any risks related to infectious disease.

AI use statement

In this work, we used generative AI tools as a coding assistant for implementing methods, and for improving text readability and addressing typos. We did not use generative AI tools to generate synthetic data or to formulate mathematical claims.

References

Markus Basan, Sheng Hui, Hiroyuki Okano, Zhongge Zhang, Yang Shen, James R. Williamson, and Terence Hwa. Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature*, 528(7580):99–104, 2015. doi: 10.1038/nature15765.

Amir Bashan, Travis E. Gibson, Jonathan Friedman, Vincent J. Carey, Scott T. Weiss, Elizabeth L. Hohmann, and Yang-Yu Liu. Universality of human microbial dynamics. *Nature*, 534(7606): 259–262, 2016. doi: 10.1038/nature18301.

Niaz Bahar Chowdhury, August George, Sumit Purohit, Angela Cintolesi, Erin Bredeweg, Jeffrey Czajka, Kelly Stratton, Yuqian Gao, Molly Stephenson, Joshua R. Elmore, Allan Scott, Damon Leach, Abby Jerger, Teresa Lemmon, Paul Piehowski, Kylee Tate, James M. Fulcher, Alex Beliaev, Kristin E. Burnum-Johnson, Paul Rigor, and Jaydeep P. Bardhan. ChatGEM: An agentic architecture enabling interactive simulation of genome-scale metabolic models. *bioRxiv*, 2026. doi: 10.64898/2026.07.20.739662.

Ryan L. Clark, Bryce M. Connors, David M. Stevenson, Susan E. Hromada, Joshua J. Hamilton, Daniel Amador-Noguez, and Ophelia S. Venturelli. Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications*, 12(1):3254, 2021. doi: 10.1038/s41467-021-22938-y.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with AI agents. *Cell*, 187(22):6125–6151, 2024. doi: 10.1016/j.cell.2024.09.022.

Almut Heinken, Johannes Hertel, Geeta Acharya, Dmitry A. Ravcheev, Malgorzata Nyga, Onyedika Emmanuel Okpala, Marcus Hogan, Stefanía Magnúsdóttir, Filippo Martinelli, Bram Nap, German Preciat, Janaka N. Edirisinghe, Christopher S. Henry, Ronan M. T. Fleming, and Ines Thiele. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nature Biotechnology*, 41:1320–1331, 2023. doi: 10.1038/s41587-022-01628-0.

Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastián N. Mendoza, Anne Richelle, Almut Heinken, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v.3.0. *Nature Protocols*, 14(3):639–702, 2019. doi: 10.1038/s41596-018-0098-2.

Shreya Johri, Erica Pimenta, Josephine Yates, Jingxin Fu, Erik L. Bao, Hyeji Jun, Brendan Reardon, Sasha Bacot, Maha Shady, Doris Fu, Wenbin Mei, Sabrina Y. Camp, Jihye Park, and Eliezer Van Allen. Evaluating agentic AI for biological discovery in autonomous and copilot settings. *bioRxiv*, 2026. doi: 10.64898/2026.06.04.729919.

Michael Kalyuzhny and Nadav M. Shnerb. Dissimilarity-overlap analysis of community dynamics: Opportunities and pitfalls. *Methods in Ecology and Evolution*, 8(12):1764–1773, 2017. doi: 10.1111/2041-210X.12809.

Jiahao Luo, Hao Wang, Devlin Moyer, Zhetao Guo, Jonathan L. Robinson, Johan Gustafsson, Mihail Anton, Yu Chen, Eduard J. Kerkhoven, Jens Nielsen, and Feiran Li. Reconstruction of human metabolic models with large language models. *Proceedings of the National Academy of Sciences*, 123(15):e2516511123, 2026. doi: 10.1073/pnas.2516511123.

Radhakrishnan Mahadevan, Jeremy S. Edwards, and Francis J. Doyle. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophysical Journal*, 83(3):1331–1340, 2002. doi: 10.1016/S0006-3495(02)73903-9.

Vinod Nair, Sergey Bartunov, Felix Gimeno, Ingrid von Glehn, Pawel Lichocki, Ivan Lobov, Brendan O'Donoghue, Nicolas Sonnerat, Christian Tjandraatmadja, Pengming Wang, Ravichandra Addanki, Tharindi Hapuarachchi, Thomas Keck, James Keeling, Pushmeet Kohli, Ira Ktena, Yujia Li, Oriol Vinyals, and Yori Zwols. Solving mixed integer programs using neural networks. *arXiv preprint arXiv:2012.13349*, 2020. doi: 10.48550/arXiv.2012.13349.

Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010. doi: 10.1038/nbt.1614.

Yili Qian, Sarvesh D. Menon, Hanchen Huang, Nick Quinn-Bohmann, Sean M. Gibbons, and Ophelia S. Venturelli. A data-driven modeling framework for mapping genotypes to synthetic microbial community functions. *Cell Systems*, 17:101652, 2026. doi: 10.1016/j.cels.2026.101652.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646:716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Kai Zhuang, Mounir Izallalen, Paula Mouser, Hanno Richter, Carla Risso, Radhakrishnan Mahadevan, and Derek R. Lovley. Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *The ISME Journal*, 5(2):305–316, 2011. doi: 10.1038/ismej.2010.117.

Appendix A The four flux-balance methods

Four flux balance analysis (FBA) related methods appear in this paper. All share the same feasible set, built from a stoichiometric matrix $S \in \mathbb{R}^{m\times n}$ over m metabolites and n reactions, and a flux vector $v \in \mathbb{R}^n$ with

$$\mathcal{F} \;=\; \big\{\, v \;:\; Sv = 0,\;\; v_{\min} \le v \le v_{\max} \,\big\}, \tag{4}$$

where $Sv = 0$ imposes steady state (no internal metabolite accumulates) and the box bounds carry reversibility, capacity and, for exchange reactions, what the medium makes available.

FBA. Maximize a biomass objective over $\mathcal{F}$:

$$\mu^\star \;=\; \max_{v\in\mathcal{F}} \; c^\top v, \tag{5}$$

with c selecting the biomass reaction. A linear program (LP). It returns a growth rate $\mu^\star$ and one flux vector, but the optimal set

$$\mathcal{F}^\star \;=\; \big\{\, v \in \mathcal{F} \;:\; c^\top v = \mu^\star \,\big\} \tag{6}$$

is generally not a singleton.

pFBA Parsimonious FBA (pFBA) breaks the tie in Equation 6 by minimizing total flux:

$$v^{\text{pFBA}} \;=\; \arg\min_{v\in\mathcal{F}^\star} \; \sum_{j=1}^{n} |v_j|, \tag{7}$$

a second LP solved over the optimal set of the first. This picks a single, reproducible member, on the argument that cells do not run pointless flux (Section 4.1.1).

FVA. For each reaction j, solve two LPs over a relaxed optimal set, where growth need only reach a fraction γ of $\mu^\star$:

$$v_j^{\min},\, v_j^{\max} \;=\; \min_{v\in\mathcal{F}_\gamma} v_j,\;\; \max_{v\in\mathcal{F}_\gamma} v_j, \qquad \mathcal{F}_\gamma = \big\{\, v \in \mathcal{F} : c^\top v \ge \gamma\, \mu^\star \,\big\}. \tag{8}$$

This returns an interval per reaction rather than a point, the range the stoichiometry permits while the organism is still alive. FVA generated the fine-tuning training corpus at $\gamma = 0.1$, because at $\gamma = 0$ the set would admit states where the organism does not grow, and a high γ would bake in the growth-maximizing bias. We used the loopless variant, having measured that it changes 0 of 1216 banded bounds while costing $1.4\times$, since loops do not cross band boundaries at the exchanges.

dFBA. Apply FBA repeatedly over time, updating the medium from the fluxes of the previous interval. Writing X for biomass and C_i for the concentration of extracellular metabolite i, each step integrates

$$X(t+\Delta t) = X(t)\, e^{\mu(t)\Delta t}, \qquad C_i(t+\Delta t) = C_i(t) + v_i(t)\, X(t)\, \Delta t, \tag{9}$$

with $v_i < 0$ for uptake, and re-solves Equation 5 at the next step against bounds recomputed from the new C_i. Mahadevan et al. (2002) defined two formulations. The dynamic optimization approach (DOA) optimizes once over the whole horizon, whereas SOA solves an FBA per interval and propagates state forward. Both methods add a rate-of-change bound

$$\big|v_j(t) - v_j(t-\Delta t)\big| \;\le\; \dot{v}_{\max}\, \Delta t \qquad \forall j, \tag{10}$$

encoding finite enzyme synthesis rate, imposed at every instant in DOA, relaxed near zero concentrations in SOA.

Equation 10 constrains how fast a flux may change, not how large it may be. The latter is the availability cap our validator applies,

$$|v_i| \;\le\; \frac{C_i(t)}{X(t)\,\Delta t} \qquad \text{for any } i \text{ being taken up,} \tag{11}$$

which prevents a culture consuming more of a metabolite than the flask holds. We added this bound in Section 4.1.1 explicitly.

Appendix B Species in the ground truth data

code	species and strain	code	species and strain
AC	Anaerostipes caccae DSM 14662	CC	Coprococcus comes ATCC 27758
BA	Bifidobacterium adolescentis ATCC 15703	CG	Clostridium asparagiforme DSM 15981
BC	Bacteroides caccae ATCC 43185	CH	Clostridium hiranonis TO-931 (DSM 13275)
BF	Bacteroides fragilis NCTC 9343	DF	Dorea formicigenerans ATCC 27755
BH	Blautia hydrogenotrophica DSM 10507	DL	Dorea longicatena DSM 13814
BL	Bifidobacterium longum NCC2705	DP	Desulfovibrio piger ATCC 29098
BO	Bacteroides ovatus ATCC 8483	EL	Eggerthella lenta DSM 2243
BP	Bifidobacterium pseudocatenulatum DSM 20438	ER	Eubacterium rectale ATCC 33656
BT	Bacteroides thetaiotaomicron VPI-5482	FP	Faecalibacterium prausnitzii A2-165
BU	Bacteroides uniformis ATCC 8492	PC	Prevotella copri CB7 DSM 18205
BV	Bacteroides vulgatus ATCC 8482	PJ	Parabacteroides johnsonii DSM 18315
BY	Bacteroides cellulosilyticus DSM 14838	RI	Roseburia intestinalis L1-82
CA	Collinsella aerofaciens ATCC 25986		

Table 3: Species of the ground truth data (Clark et al., 2021) and the abbreviations used in figures and tables. We included 25 of the 26 species. The 26th, Holdemanella biformis DSM 3989 (HB), is missing from AGORA2.

Appendix C Full single-agent results

Table 4: Full version of Table 1, with the metabolite error broken out for each of the four metabolites. Here bal. acc. stands for balanced accuracy, ρ for the Spearman rank correlation, and sd for standard deviation.

metric	Llama-3.3-70B	Llama-3.1-8B	Qwen 2.5-72B	Haiku 4.5	dFBA	Llama-3.1-8B fine-tuned
butyrate call (bal. acc.)	0.512	0.583	**0.691**	0.613	0.500	0.500
growth rank (ρ)	−0.076	−0.030	0.225	0.139	0.105	**0.325**
growth fold ($\log_{10}$)	1.243	0.772	0.969	1.015	0.911	**0.614**
metabolites ($\log_{10}$)	1.329	1.021	1.234	1.126	0.845	**0.748**
butyrate	0.949	0.899	1.226	1.071	**0.370**	0.439
acetate	0.903	0.897	**0.431**	0.515	1.750	1.033
lactate	2.114	0.948	1.724	1.725	**0.110**	0.177
succinate	1.349	1.339	1.554	1.193	**1.149**	1.341
inside replicate sd	0.181	0.313	0.253	0.217	**0.422**	0.398
clean mass balance	167/300	90/300	**173/300**	165/300	–	**173/300**
no usable answer	5/300	16/300	2/300	30/300	–	**0/300**

Appendix D Trajectories for the single-agent models

To complete previous agent-based runs (Figure 7), we provide the remaining models for completeness, because every model ran the identical flask, so the panels are directly comparable species by species.

Only the 48 h endpoint is measured, and the results here are the trajectories themselves. The flat segments visible in several panels are stalls, where growth ceases on exhaustion of a trace nutrient while a bulk carbon source remains, rather than decay.

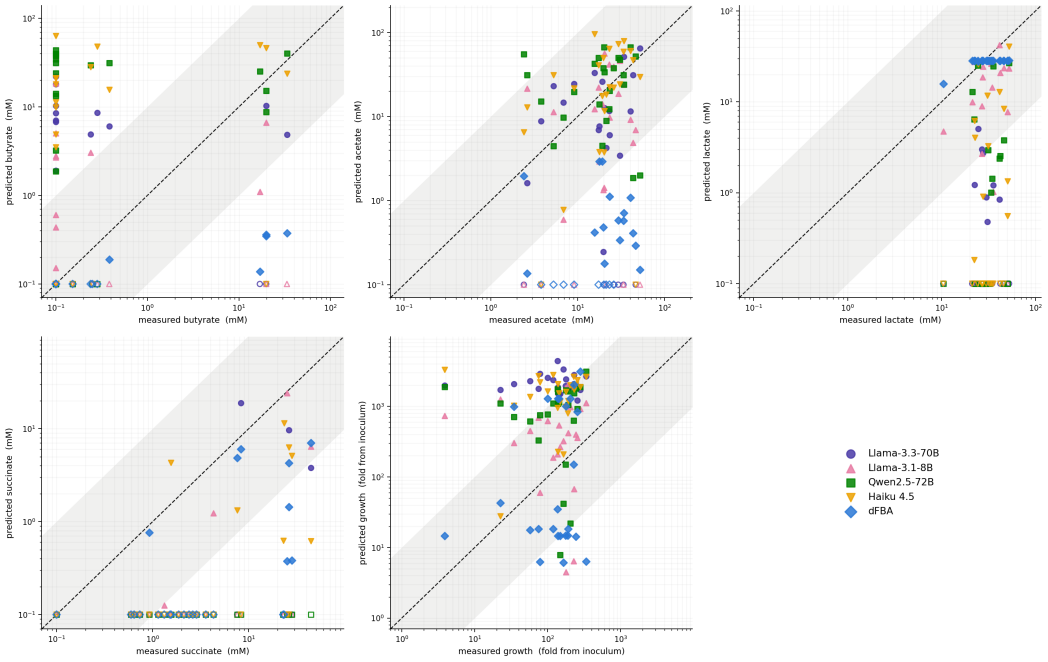


Figure 6: Predicted against measured, 25 monocultures, one point per species. The diagonal is a perfect prediction and the band is within $10\times$. Hollow markers sit on the floor at predicted zero. Carbon accounting for the same runs is in Figure 1.

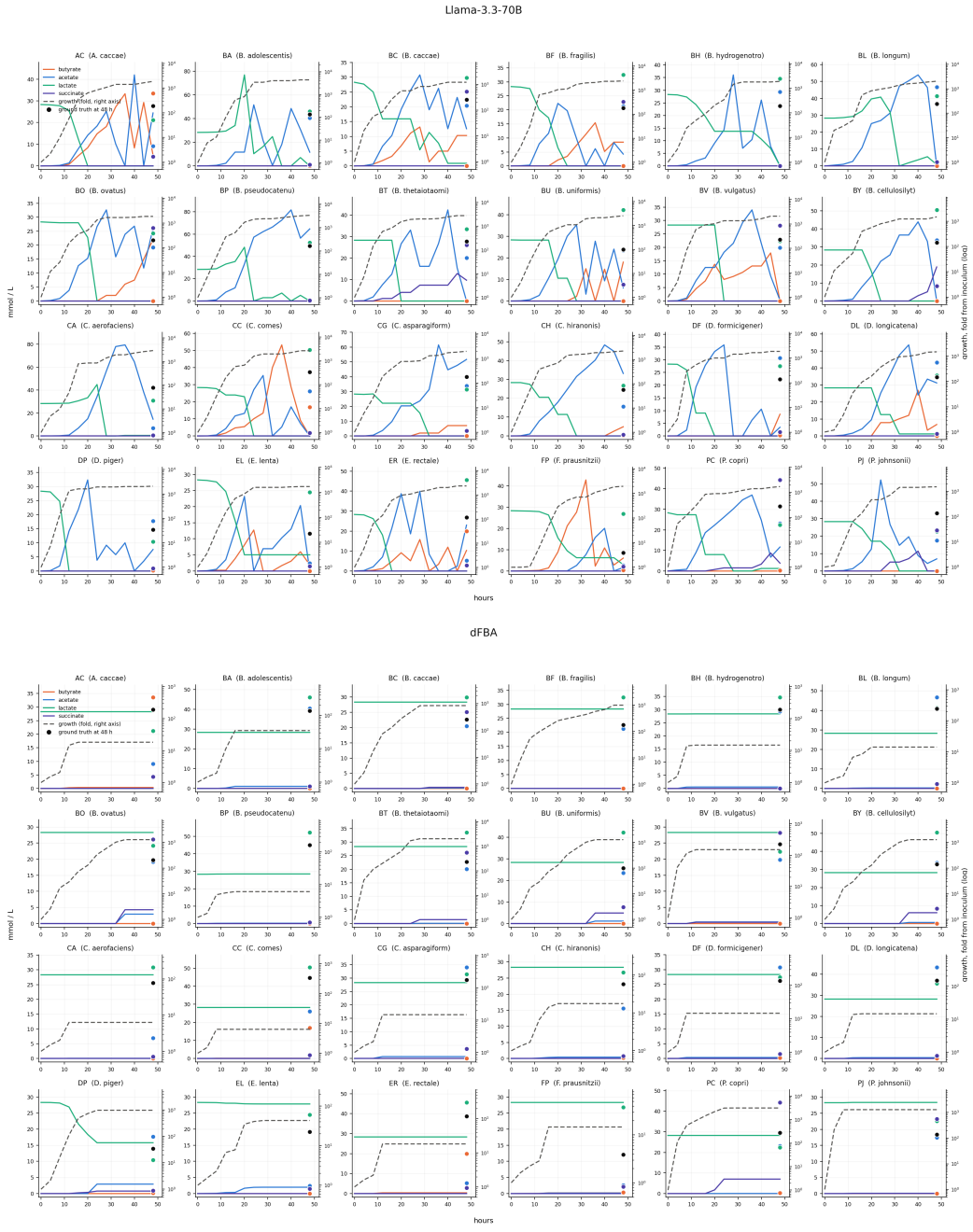


Figure 7: Metabolite and biomass trajectories over the twelve turns, Llama-3.3-70B (top) and dFBA (bottom). Only endpoints are measured.

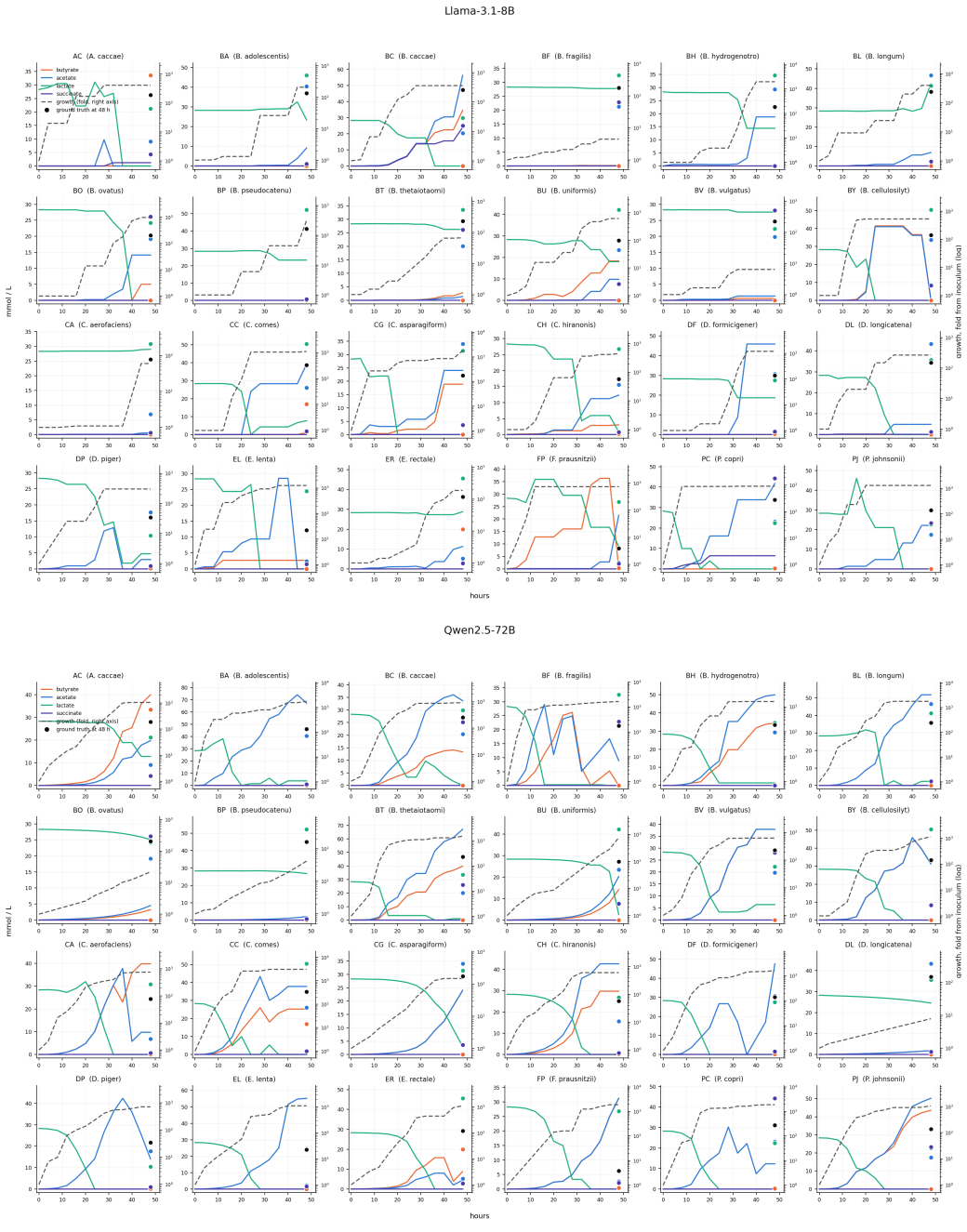


Figure 8: Trajectories, Llama-3.1-8B (top) and Qwen2.5-72B (bottom). Qwen has the best butyrate call in the study (0.691). Llama-3.1-8B is the base model our LoRA is trained on, so it is the controlled comparison for Section 4.1.2.

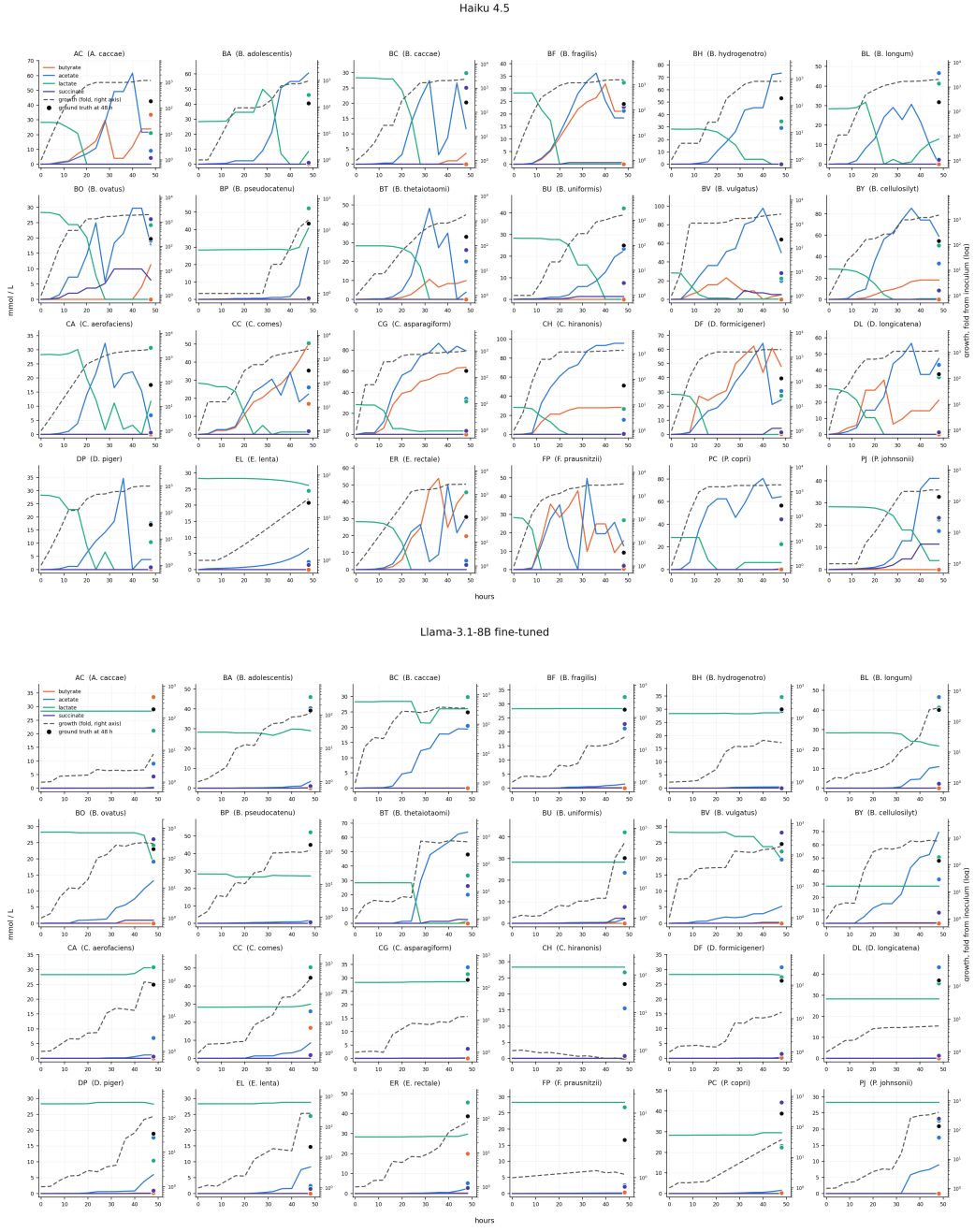


Figure 9: Trajectories, Haiku 4.5 (top) and Llama-3.1-8B fine-tuned (bottom), the latter queried in the format it was trained on. Compare the bottom panel with the solver in Figure 7: the fine-tuned agent adopts the solver's carbon strategy, leaving glucose largely untouched, where the prompted models consume it to exhaustion.

Appendix E Diauxic growth and overflow metabolism scorecards

Diauxic growth. Every model on the diauxic batch at an oxygen cap of 6 mmol g^{-1} h^{-1}, scored by the compound criterion of Section 4.1.1. The trajectories these rows summarize are panels (a) to (c) (Figure 2).

model	cap	peak μ (h^{-1})	$Y_{X/S}$ (g/mol)	glc gone (h)	plateau (h)	acetate (mM)	re-eaten	ph2 (%X)	verdict
naive dFBA	10	0.77	84.6	6.0	0.00	2.43	✓	7.0	instant switch
dFBA + dv/dt	**10**	**0.77**	**84.6**	**6.0**	**0.50**	**2.43**	✓	**7.0**	DIAUXIC
Llama-3.1-8B	10	1.39	71.9	11.0	0.00	0.00	–	0.0	instant switch
Llama-3.3-70B	10	1.00	89.2	9.5	0.00	0.00	–	19.3	instant switch
Qwen2.5-72B	10	0.22	19.5	–	0.00	0.22	–	–	no switch reached
Haiku 4.5	10	0.25	23.9	–	0.00	1.47	–	–	no switch reached

Table 5: The signature is compound and a model has to satisfy all of it, meaning a plateau at the switch, a slower second phase, and acetate both made and re-consumed. Bold rows satisfy all three. "glc gone" is the time at which glucose is exhausted. "ph2" is the share of final biomass made after glucose is gone, which separates a real second phase from a single timestep.

Overflow metabolism. Every model on the glucose uptake experiment, scored by the criterion of Section 4.1.1. The solver rows are the oxygen concentration sweep that establishes the observed elbow in the plot. The agent rows represent the four prompted models.

model	$\mu_{\max}$ (h^{-1})	μ_c (h^{-1})	slope	acetate below μ_c	acetate max	verdict
pFBA, O_2 cap 15	0.86	0.82	52.8	0.00	0.17	no elbow
pFBA, O_2 cap 10	**0.70**	**0.56**	**53.1**	**0.00**	**6.52**	ELBOW
pFBA, O_2 cap 8	**0.64**	**0.44**	**53.3**	**0.00**	**9.08**	ELBOW
pFBA, O_2 cap 6	**0.58**	**0.35**	**51.6**	**0.00**	**11.05**	ELBOW
Llama-3.1-8B	0.50	0.00	-23.0	31.50	31.50	no elbow
Llama-3.3-70B	0.67	0.10	-33.7	1.00	11.55	no elbow
Qwen2.5-72B	0.60	0.20	-2.6	0.00	2.46	no elbow
Haiku 4.5	1.00	0.13	-5.9	4.72	7.47	no elbow

Table 6: Overflow metabolism of acetate in E. coli for 4 different models as determined by the ELBOW criterion. ELBOW requires a breakpoint strictly inside the sweep, a positive slope above it, and near-zero acetate below.

Appendix F LP admissibility of the agent proposed fluxes

Whether a flux distribution proposal emitted by an agent is a point the corresponding reconstruction's linear program admits at all, is a question we asked of every stored exchange vector after single-agent runs on DM38.

model	n	k	band mean	sd	all k jointly
Llama-3.3-70B	295	5	0.878	0.139	51.2
Llama-3.1-8B	283	5	0.868	0.163	47.3
Qwen2.5-72B	298	6	0.911	0.101	51.7
Haiku 4.5	270	8	0.901	0.086	35.6
Llama-3.1-8B fine-tuned	225	26	**0.928**	0.040	9.3

Table 7: How much of an LLM agent's flux distribution proposal is an admissible solution of a metabolic network. k is the median number of exchanges a model proposes per turn and "band mean" is the fraction of them that are admissible to the integer program over S. "sd" is the standard deviation of that fraction. "all k jointly" is the percentage of proposals admitted whole.

Appendix G Interaction networks derived from the applied fluxes

Network layers are recovered from the agents' proposed flux distributions, and neither is asked of them in any prompt. This is important because it highlights the trajectories the agents take through the interaction space to achieve the observed outcomes from the ground truth data, which contains only endpoints, so no measured network exists in ground truth data.

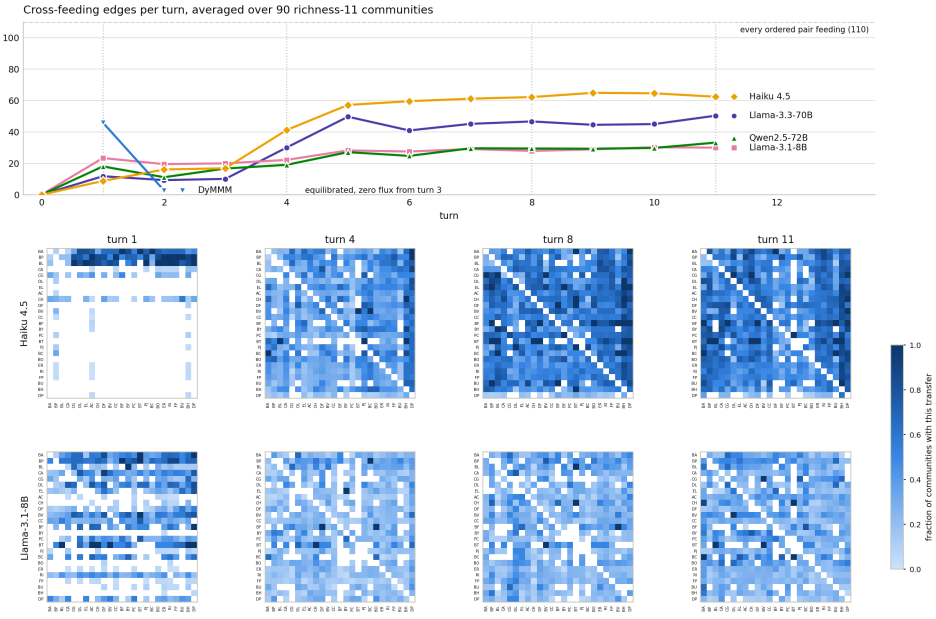


Figure 10: The cross-feeding network assembling. An edge $A \to B$ means B consumed, this turn, a metabolite A released on the previous one. **Top**, edges per turn averaged over the 90 richness-11 communities, against 110 possible ordered pairs. **Bottom**, the same as matrices at four turns, row releasing and column consuming, shaded by the fraction of communities in which that transfer occurred.

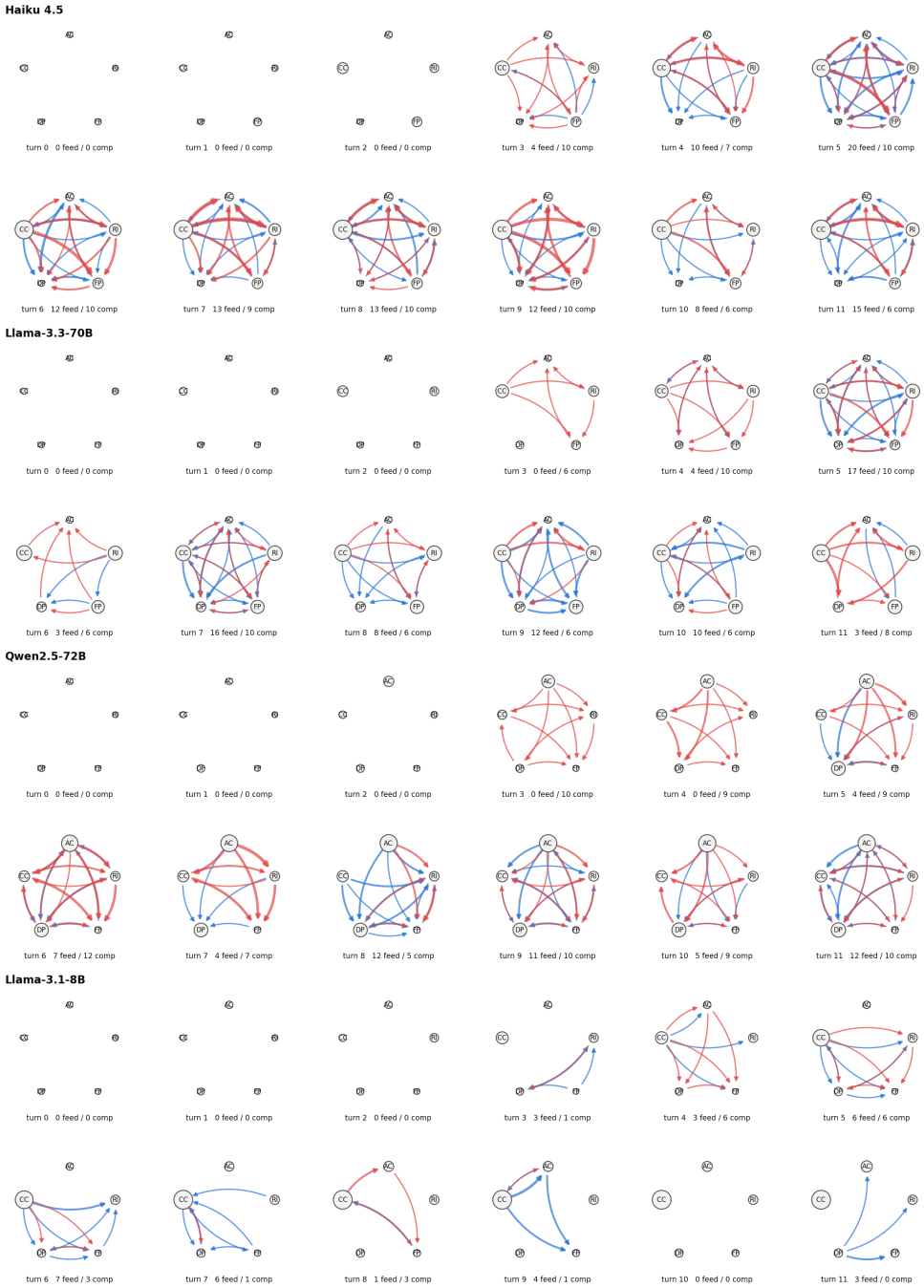


Figure 11: Cooperation and competition interaction layers on one flask, every turn, for one richness-5 community. **Blue** $A \rightarrow B$ means B consumed a metabolite A released last turn. **Red** $A \rightarrow B$ means a metabolite was rationed for B this turn and A held the larger share of demand for it. Node size is correlated to biomass. Competition establishes first and holds roughly flat while cross-feeding builds, which follows from the physics, since competition needs only shared demand on the medium's own substrates, whereas feeding requires that someone has already released something. The solver is absent because it records no rationing.