CELLA: PREDICTING MICROBIAL CULTIVATION REQUIREMENTS ON A MULTISCALE METABOLIC GRAPH

Anonymous authors
Paper under double-blind review

ABSTRACT

Microbial growth is determined by interacting processes across a wide range of temporal and spatial scales, from molecular structure to phylogeny. Multiple modalities, including genomic data, genome-scale metabolic models, chemical structures, and media properties, often inform expert predictions. However, previous foundation models have been restricted to data available in large quantities and readily consumable by ML algorithms (proteins, DNA, chemical structures). In this work, we present Cella, a heterograph model which enables predictions based on different data sources at scale while encoding real mechanistic relationships and handling coverage gaps gracefully. We fit the model on a newly constructed collection of 4,220 prokaryotic organisms and their associated metabolic models sampled exhaustively at the genus level. We evaluate Cella on various industrially relevant tasks, including media and component growth prediction, carbon-source utilization, environmental preference, and component substitution, and strain ranking by shared cultivation requirements. Tasks are scored under held-out organism and component splits, against popularity and lineage-only baselines. We predict growth media on the nearly half of species in our collection that do not record one, while providing a traceable path to the evidence supporting the predictions. Cella enables multimodal data integration while retaining interpretable interactions and points towards underlying biological patterns learnable with foundation-model-style approaches that generalize across the microbial landscape.

1 INTRODUCTION

Biology operates across multiple orders of magnitude in time and space, from the submolecular to the multicellular. Furthermore, biologists gather data from systems under study using a wide variety of methods; these observations only in combination give a well-rounded picture of a biological system. Previous representation learning work has generally focused on single or limited number of data modalities available in large quantities for model training: protein sequences (Lin et al., 2023), genomic sequences (Brixi et al., 2026), or small-molecule chemical structures (Ross et al., 2022). However, prediction of whole-cell phenomena such as growth of different microbial species requires integrating knowledge across multiple scales. In this work, we present Cella, a graph-based representation of microbial metabolism that spans genomic, metabolic, phylogenetic, and chemical modes. To allow generalization across the kingdom of life, we fit our model on a newly constructed collection of 4,220 metabolic models spanning all prokaryotic genera present in GTDB (Parks et al., 2026). We use the representation to perform several downstream tasks relevant for industrial applications, including medium and component growth prediction, carbon source ranking, environmental prediction, and component substitution. We design media for organisms with no recorded conditions and examine the pretrained embedding for biologically relevant patterns. We first ground the work in existing literature on multimodal models and graph-based representation learning methods applied to biology. We then describe the training set collection, representation structure, and assessment construction, before presenting our results and implications for the field.

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.

1.1 MULTIMODAL MODELS IN BIOLOGY

In addition to large single-modality foundation models (Brixi et al., 2026; Lin et al., 2023; Wang et al., 2024) trained on biological sequences, there is substantial literature which combines enzyme sequences and molecular structures for kinetic parameter and substrate prediction (Li et al., 2022; Kroll et al., 2023b;a). These models typically concatenate the protein and small molecule features into a single vector, a method that does not scale to tasks with a variable number of data modes and a high prevalence of missing data. Several methods to unify DNA, RNA, and protein language and structure models have been proposed using concatenation, attentional pooling (Mollaysa et al., 2025), shared token space (Pei et al., 2026), and reverse transcription/translation of protein and RNA sequences (Liu et al., 2025b). Another body of work uses natural language as an interface to understand different views of biomolecules (2D graphs, 3D structure, and 1D strings) (Zhuang et al., 2025; Pei et al., 2024). However, this work is restricted to the central dogma or molecular structure and does not integrate information across scales of cellular organization.

For the tasks of growth and environmental prediction we address, several models have addressed one or more of the tasks using supervised machine learning methods. MediaMatch predicts bacterial growth on culture media by training independent classifiers over 45 common media recipes on 16s rRNA sequences (Liu et al., 2025a). GenomeSPOT uses a whole-genome featurization pipeline to predict oxygen tolerance, optimum temperature, salinity, and pH (Barnum et al., 2024). Importantly, it also evaluates on out-of-clade organisms to ensure cross-species generalization across species, whereas MediaMatch uses random cross-validation over sequences, which can overinflate accuracy metrics by including sequences from identical species in both training and test sets. Canopy is a heterograph foundation model trained on 13 different node types for genetic and metabolic engineering tasks (Bowden et al., 2026), but has no knowledge of medium composition and is primarily focused on well-known model organisms used as bioengineering chassis rather than wild isolates across the kingdom of life.

1.2 GRAPH-BASED REPRESENTATION LEARNING METHODS

Graph-based representations have been widely used in biology, as many biological phenomena (e.g. protein-protein interaction networks, gene regulatory networks, and metabolism) naturally lend themselves to modelling with a graph structure. Knowledge graphs are widely used in biomedical applications to represent heterogeneous datasets (Hogan et al., 2021). Nodes represent entities and edges the relationships between them, often with types that allow message passing to be parameterized per node and relation type (Schlichtkrull et al., 2018; Hu et al., 2020). A robust ecosystem of tools to construct and evaluate knowledge graphs automatically exists (Zhang et al., 2025; Lobentanzer et al., 2023; Callahan et al., 2024). The most relevant knowledge graph for this work is the recently published KG-Microbe, but this does not include GEMs or metabolic reconstructions in their knowledge graph (Santangelo et al., 2026).

In metabolism, graph learning and GNN methods have been applied to predict gene essentiality (Hasibi et al., 2024; Sharma et al., 2024). Genome-scale metabolic models (GEMs), which represent metabolism as a graph of reactions between metabolites catalyzed by enzymes, are widely used and automated mechanistic reconstruction tools exist to construct them at scale (Faria et al., 2023; Machado et al., 2018; Zimmermann et al., 2021) as well as to compare them (Matveishina et al., 2025). For instance, SIMBA-GNN uses metabolic networks to augment an edge-aware graph transformer for microbiome community prediction (Aminian-Dehkordi et al., 2025). Another body of work treats molecular structures as graphs, with nodes representing atoms and edges bonds; these graph representations are then used for reaction retrosynthesis, which often requires integrating enzyme sequences in the model (Cong et al., 2025; Zhong et al., 2023). To our knowledge, Cella is the first graph to combine media and organismal information with reconstruction-derived uptake capability for every organism to enable individual component prediction.

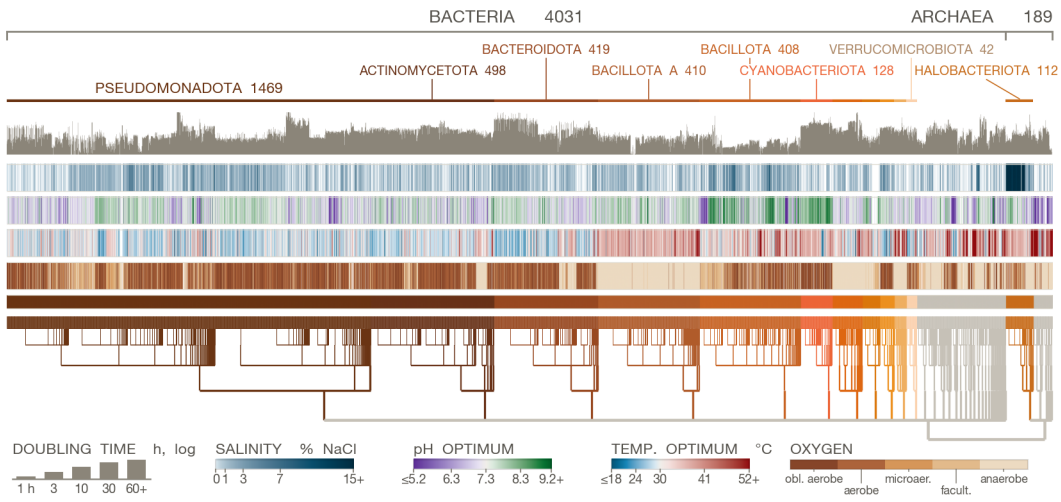

Figure 1: **Model zoo.** We construct a training set of 4,220 prokaryotic organisms and annotate it with experimental and model-based annotations. Each organism has three GEMs constructed with CarveMe, ModelSEED, and gapseq. From top to bottom: Doubling time (log), salinity, pH, temperature, oxygen category, and phylogeny.

2 METHODS

2.1 MODEL ZOO CONSTRUCTION

We wanted to distribute the training set across the prokaryotic kingdom of life, so we selected one representative genome per genus in GTDB (Parks et al., 2026), restricted to named genera with a non-metagenome (MAG) isolate that passes quality control (see Figure 1). GTDB contains 25,000+ genera, but the large majority are MAG-only lineages which build fragmentary models, so our restrictions result in 4,220 organisms (4,031 bacteria and 189 archaea). We construct three independent genome-scale metabolic models per genome from each downloaded NCBI assembly using the CarveMe, (Machado et al., 2018) ModelSEED (Faria et al., 2023), and gapseq (Zimmermann et al., 2021) pipelines. All three are gap-filled on the per-organism gap-seq predicted medium to make them comparable on a common nutritional basis. We collect associated environmental and cultivation parameters from BacDive (Schober et al., 2025) and metaTraits (taxon consensus) (Podlesny et al., 2026) and predicted remaining traits using GenomeSPOT (Barnum et al., 2024) and Phydon (Xu et al., 2025) (see Table S3). We reconcile the parameters by preferring exact-strain literature records over type-strain and species-level ones, then falling back to genome-based prediction only where no measured record exists (see Table S2). To reconcile the namespaces between the model pipelines, we mapped all compounds and reactions onto MNXref (Moretti et al., 2021) and removed compartment suffixes.

2.2 MEDIA COMPONENT RECONCILIATION

Media compositions are taken from MediaDive (N=14,908 ingredient records, N=709 media recipes, N=627 distinct components) (Koblitz et al., 2023). 86% of records have a quantity in $\mathrm{g\,L^{-1}}$. We link cultivation records to MediaDive identifiers. Components are represented by frozen MolFormer embeddings (Ross et al., 2022) after normalising hydration and salt-form suffixes and mapping conjugate-base naming (such as folic acid to folate). N=59 components are biological preparations (e.g. peptones, yeast extracts) which have no single molecular structure; we classify these as mixtures and assign them a null chemistry vector with an explicit indicator. Mixtures are represented only through their position on the graph, which means that no chemistry-based prediction is possible for them at this time. N=75 components could not be resolved to a structure and are treated likewise.

Media components are often compounds with multiple species (e.g. NaCl dissociates into sodium and chloride ions) supplied to different metabolic exchanges; we therefore dissociate each component, splitting dot-separated SMILES strings into fragments, matching cations, halides, and neutral species directly, and curating 14 polyatomic ions. This resolves 409 of 627 compounds to at least one metabolite and yields 520 `resolves_to` edges (a name-based match alone yields 6). From each reconstruction we extract gene-protein-reaction (GPR) mapping of each reaction and classify it as single-gene, isozyme (pure disjunction), complex (pure conjunction), mixed, or absent. We separate hydration salts into separate components but map them to the same metabolite exchanges.

2.3 Representation structure

Cella is a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \tau, \phi)$ with node typing $\tau : \mathcal{V} \to \mathcal{T}$ and relation typing $\phi : \mathcal{E} \to \mathcal{R}$, comprising 7,377 nodes across six types and 1.6M edges across 22 relations (see Figure 2A). Three node types carry features: organisms carry the 205 amino-acid composition statistics of GenomeSPOT (Barnum et al., 2024), components a frozen 768-dimensional MolFormer embedding (Ross et al., 2022) with an indicator for whether a structure was resolvable, and metabolites a single indicator separating those an organism can exchange from intracellular intermediates. Taxa and media are featureless.

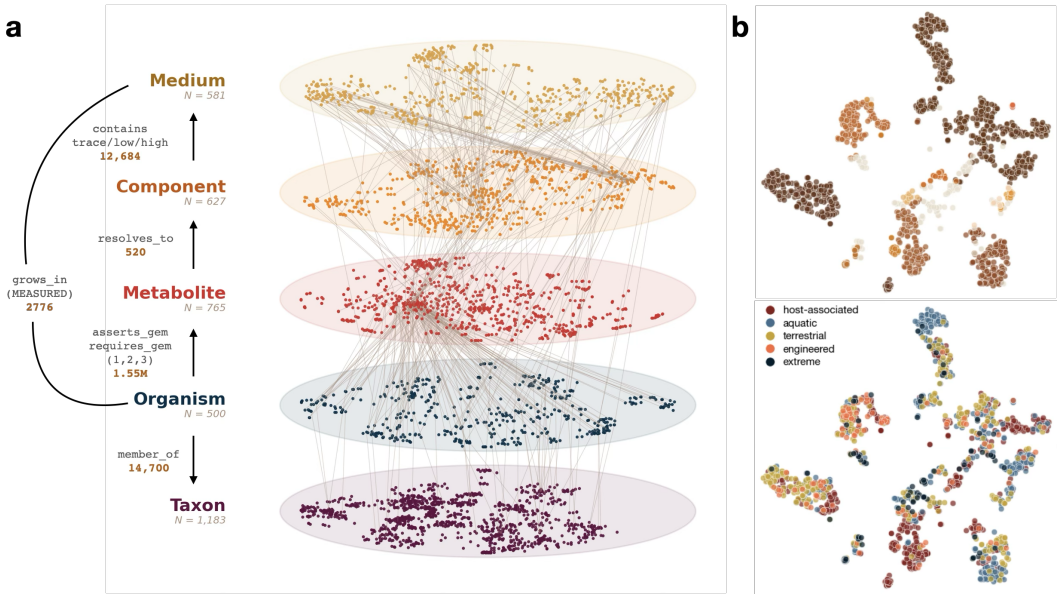

Figure 2: **The Cella graph. (A)** We show a subset (500/4,220 organisms) and a random downsample of edges (N=45 per relation). Relation names and counts are at left; node counts are below names. Position within each plane is a cosine t-SNE of each layer's features; taxa sit at the mean position of their member genomes. **(B)** T-SNE 2D representation of pretrained Cella encoding, colored by phylum (top) and habitat (bottom). Only organisms with a single habitat label are included (N=1612).

We include source provenance as part of the relation type. For example, $(\texttt{organism}, \texttt{requires}, \texttt{metabolite})$ becomes `requires_1of3`, `requires_2of3` or `requires_3of3` according to how many reconstructions assert it. All relations except for `grows_in` are derived from simulation, annotation, or ontology. Growth records used for component prediction are also used for establishing `grows_in` relations, so we rebuild the graph per-split with the relation removed for held-out organisms to avoid leakage of information through the graph structure.

Cella is trained as a relation-typed message-passing network implemented in PyTorch. Node features are linearly projected to a common width $d = 64$; node types without features are given learned embedding tables of the same width. The $L = 2$ message-passing layers compute for every relation a bias-free linear transform of the source states, averages the resulting messages at each destination

node, and adds a per-node-type root transform. The sum is layer-normalized per node type, passed through a ReLU, and added to the incoming state as a residual. We parameterize the transform per-relation to distinguish agreement-specific behavior from the GEMs layer. We collapse the three agreement relations (`requires_1of3`) for the agreement ablations. The encoder has 407k parameters, or 420k including the link-scoring head. Following Bowden et al. (2026) we add a virtual node with bidirectional edges to every other node, providing a shortcut between otherwise disconnected components; our graph is fragmented in the absence of it, since organisms sharing no medium have no short path.

Supervised models are trained full-batch for 300 epochs with AdamW (learning rate 3×10^{-3}, weight decay 10^{-4}) on a binary cross-entropy objective, with positive weighting set to the observed negative-to-positive ratio. Organisms are split 80/20, grouped so that all rows of an organism fall on one side, and every result is averaged over three seeds. A self-supervised encoder is pretrained on link prediction over the graph's own edges with a DistMult decoder. We exclude `grows_in` from pretraining.

2.4 Downstream tasks and assessment

We defined several downstream tasks to assess the broad applicability of Cella to a variety of relevant scientific questions (see Table 1). The first three tasks model growth condition prediction as link prediction at three node types: media prediction (Medium layer), component prediction (Component layer), and carbon-source utilization (Metabolite layer). These tasks ask progressively finer questions. Medium prediction answers questions like "Can this organism grow in MRS media?" over the 18 DSMZ media with at least 25 recorded organisms. Component prediction specifies which of 70 ingredients should be included in a particular organism's media to support growth, so a proposal can be made for organisms with no growth records. Carbon-source prediction covers whether an organism can use each of 72 compounds as a sole carbon source.

A second suite of nine tasks covers condition tolerance: temperature, pH, salinity optima regression, oxygen tolerance classification, and habitat multi-label classification, as well as growth envelope classification for salt, temperature, and pH. The envelope tasks predict whether an organism grows at a specific tested value, with the tested value supplied as an input to the (organism, value) pair. These tasks have measured negatives in BacDive, though they are scored against a value-only (organism-neutral) control because the tested value alone is strongly predictive.

To prevent data leakage, we randomly split training and test data in a deterministic hash-defined manner. We hold out organisms (5%, 5 folds), components (20%, 5 folds) and organism and component combinations. Components to be held-out must be MolFormer-embeddable (i.e. not undefined mixtures) and in a 1-40% prevalence band to eliminate highly prevalent species like water. For environment tasks, we performed out-of-family splits to avoid held-out genera having close family relatives in the training set. We also restricted training data to those with experimental evidence to avoid including data predicted from GenomeSPOT or Phydon.

We present a number of baselines: popularity, lineage-only, and genome-only (which on the environment tasks represents the GenomeSPOT baseline). Baseline models are random forests. Decision thresholds are the per-label F1-maximising values fitted on a calibration split held out of training. Node degree in Cella is extremely skewed due to common components like phosphate, which are attached to nearly every medium, while metabolites are only attached to a handful. To prevent matching only the in-degree prediction for hubs, we degree-match negatives for link prediction during evaluation.

3 Results

3.1 Cella can predict well across a variety of tasks

Cella clears the popularity control everywhere and the lineage-only control on both media tasks, and it gives the highest macro F_1 of any arm on media components (0.502) and cultivation media (0.351). Cella ranks level with the best concatenation arm on cultivation media (AUROC 0.922 against 0.920) but below it on media components (0.861 against 0.878) and carbon sources (0.612 against 0.649) (see Table 2). In the environment suite, the twelve-way habitat prediction task achieves 0.808 AU-

Table 1: **Downstream evaluation.** Cella is evaluated across multiple biological scales, from prediction of complete cultivation media to organism-level environmental traits and representation-based ranking tasks. Detailed dataset sizes and label distributions are reported in Appendix Table S1.

Biological scale	Evaluation	Predicts	Task type
Nutrition prediction			
Medium	Cultivation media	compatible complete media	link prediction
Component	Media components	components associated with cultivation	link prediction
Metabolite	Carbon sources	compounds usable as a sole carbon source	link prediction
Environment prediction			
Organism	Environmental optima	optimum temperature, pH, and salinity	regression
Organism	Oxygen tolerance	aerobic, anaerobic, facultative, or microaerophilic phenotype	4-class
Organism	Habitat	environmental habitat	multi-/12-class
Organism × condition	Growth envelopes	growth at a specified temperature, pH, or salinity	binary
Representation and applications			
Organism pair	Strain similarity	similarity in cultivation requirements	ranking
Organism × component pair	Component substitution	suitability of one medium component as a substitute for another	ranking
Organism	Unseen medium proposals	cultivation components for organisms without recorded media	inference

Table 2: **Nutrition tasks.** Macro F_1 and macro AUROC under organism-grouped splits, mean over five seeds. Labels are experimental only. Popularity predicts each label at its base rate, whereas lineage only is a one-hot encoding of domain, phylum, class, order and family. On carbon sources every arm here is averaged over the 52 of 72 labels that resolve to a metabolite node, which is the set the graph model scores; the restriction moves the feature arms by at most 0.006 F_1.

	Carbon sources		Media components		Cultivation media	
Arm	n=911 F_1	52 labels AUROC	n=2,776 F_1	70 labels AUROC	n=2,534 F_1	18 labels AUROC
popularity (control)	0.000	0.500	0.026	0.500	0.000	0.500
lineage only (control)	0.246	0.630	0.424	0.833	0.263	0.898
genome	0.108	0.635	0.472	0.867	0.301	0.913
genome + panels (all three)	0.106	0.649	0.471	0.878	0.262	0.920
Cella (graph)	0.229	0.612	0.502	0.861	0.351	0.922

ROC compared to a lineage-only control of 0.748; accuracy rises from 0.336 to 0.455 against a majority-class floor of 0.211. However, performance is not uniform. Our absolute F_1 on the carbon source panel is low, and the AUROC and F_1 diverge as training data is added, indicating that noise in the labels is likely limiting the performance of the models on this task (see Figure S2). Growth records are positive-unlabeled, which means that a strain not recorded on a given carbon source does not indicate a lack of growth (negative label), but rather that the strain has not been tested on those sources (unlabelled). This is a systematic bias within the training data available from public literature. We also do not beat the lineage-only control on the oxygen task based on AUROC; we however lead on accuracy.

3.2 Modalities carry different information

Before training Cella, we assessed mode-to-mode predictive accuracy by performing a held-out R^2 assessment (see Figure 4). We found that genome is the most well-predicted by the other modes, which is unsurprising as the GEM modes are downstream, but no mode pair topped 0.69 R^2. Media was especially unpredictive of the other modes, as it does not contain any organismal information, which further justifies the multiscale nature of Cella.

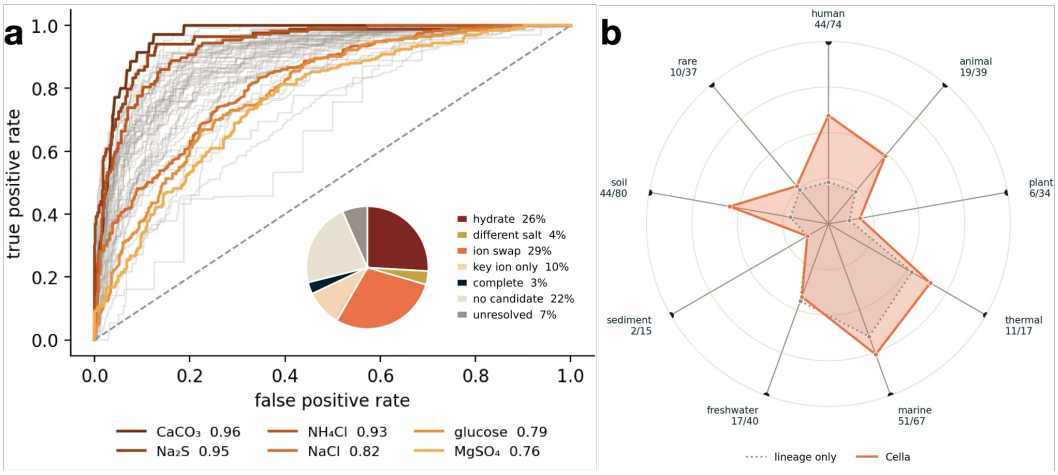

Figure 3: **Cella performs well across multiple tasks.** **(A)** ROC curves for component prediction for the 59 components with at least 25 held-out positives; highlighted are six representative curves. Inset: pie chart the substitution scorer's top-ten proposals per chemically defined component (N=45), classified by how completely each replaces the original. **(B)** Radar plot of 12-way habitat prediction. We scale each axis so the outer ring is that habitat's number of held-out genomes and the shaded region covers the number classified correctly, so the distance along each spine is recall. Orange is Cella; dashed is the lineage-only control. The rare categories with fewer than 50 genomes (food, wastewater, hypersaline, built/air) are lumped together into a single axis to increase the number of held out samples in the axis total.

We then combined various modes in panels for each of the three nutrition tasks and compare them to the genome-only baseline. While multimodal prediction beats all baselines on every task, the margins of improvement are small in absolute terms, and the genome block supplies most of what is learnable. Only one panel (ModelSEED on carbon sources) achieves genome-only accuracy without including the genome, which indicates that the information in the metabolic layers of Cella is complementary to the genome rather than sufficient. Encoding the three reconstructions as separate panels gives the highest AUROC on all three nutrition tasks, though the margins are small (see Table 2). For components with ≥ 25 test samples, we plot the receiver-operator characteristic curves to examine tradeoffs between true and false positives (see Figure 3A). On the habitat prediction task, we compare to a lineage only-baseline and find Cella outperforms it on all but freshwater labels (see Figure 3B).



Figure 4: **Modality overlap.** Mode-to-mode predictive accuracy, measured by held-out R^2 on a linear predictive model.

3.3 GEM PIPELINES AGREE ON THEIR LIMITED OVERLAPS

We compared the three GEM reconstruction pipelines and found that even post-canonicalisation, they share far less of their metabolite vocabulary than expected. Of 808 compounds modelled by at least one tool, only 130 are covered by all three (see Figure S1A). This limited overlap is the result of how the different tools build the network: ModelSEED and gapseq work bottom-up, adding reaction to a minimal network where the evidence justifies them, whereas CarveMe works top down,

removing reactions lacking evidence from a curated model. However, where the models do overlap they agree. We tested 212,860 organism-compound pairs to assess presence/absence and found that 82.8% of pairs agree unanimously (see Figure S1B). On components shared by two models, over 80% of pairs agreed. Furthermore, cross-prediction across pipelines did not top more than 0.38 R^2. CarveMe incorporates 427 components that the other pipelines do not include; as a result, the different coverage of the pipelines justifies their individual use, not just their agreement.

3.4 CELLA CAN BE USED TO SCORE SUBSTITUTIONS AND PREDICT UNSEEN MEDIA

The media-component model predicts which ingredients co-occur in a recipe and cannot propose what to do when an ingredient is unavailable. We instead scored component substitutions at the level of the chemical species the component supplies (Metabolite layer). Each component was resolved to species and weighted by specificity, to remove ubiquitous counterions (Na^+, Cl^-, SO_4^{2-}, each supplied by ≥ 30 components) from consideration. We classified each substitution by whether it is a hydrate of the same salt, a different salt, an ion swap, inclusion of the key ion (non-counterion) only, or a complete replacement (see Figure 3A, inset). The majority of substitutions are for a chemically similar component: 66% of substitutions were for a near-identical chemical, which demonstrates the validity of the substitution scorer. However, for 22% of components, mostly vitamins such as biotin, folic acid, and riboflavin, no suitable substitution was found by the scorer.

Cella can also be used to score unseen media. We scored the 1,973 organisms in the model which have no recorded growth media.

To assess predictions for genera with no recorded medium, we report accuracy conditional on patristic distance to the nearest labelled genus: held-out genera are binned by that distance, and assess how many unlabelled genera fall into each bin. Accuracy on the held-out set across the bins was largely flat. 81% of unlabelled genera fall in the two most distant bins, so this performance consistency implies the accuracy they inherit is close to the pooled held-out values (AUROC 0.876, macro F1 0.513).

We found that the proposed media show distinct phylogenetic patterns (see Figure 5); however, the structure is not only phyla-determined, as clustering the proposed media recovers phyla only loosely (adjusted mutual information 0.31). As expected, components rise and fall together with the recommendation of recipes which contain them: sodium resazurin and NaOH, markers of anaerobic formulations, are elevated together in particular clades. The model over-calls relative to the recorded media average (median 10 components proposed against a median of 8 in labelled media). The excess is concentrated in particular parts of the phylum ordering (Figure 5, bottom track). We therefore produce per-component probabilities along with the calls to allow explicit tradeoffs between coverage and precision.

3.5 CELLA'S REPRESENTATION ENCODES BIOLOGICALLY RELEVANT INFORMATION

We embedded all 4,220 genera with a pretrained encoder and reduce to two dimensions to show a space clustered first by clade, then by habitat (See Figure 2B). Within a clade, an organism's nearest neighbors are enriched for its same habitat over what would be expected based on clade alone. The embedding also improves on the features it is built from. We compute functional similarity over media components, by calculating an IDF-weighted Jaccard over recorded media-component sets. Against this metric, Spearman ρ is 0.162 for the Cella embedding and 0.138 for the raw GenomeSPOT features it is built from. The advantage survives conditioning on phylogeny (partial ρ 0.161 given patristic distance). Patristic distance alone reaches only 0.019, which indicates that the embedding is tracking function rather than relatedness.

However, since much of the predictive performance on many tasks comes from the genomic features, organisms with no genome cannot be effectively placed in the embedding space. The edges in the graph structure propagate forward from the organism layer to the other layers informed by the GEM reconstructions, so an organism with only a genome but no GEM (or only a subset of GEMs) can still be embedded gracefully. In addition, the unseen media task is deliberately designed to show how interpolation of unseen nodes can function well in cases with missing data (in this case, a single growth record). Cella therefore scales and adapts well to some types of data missingness.

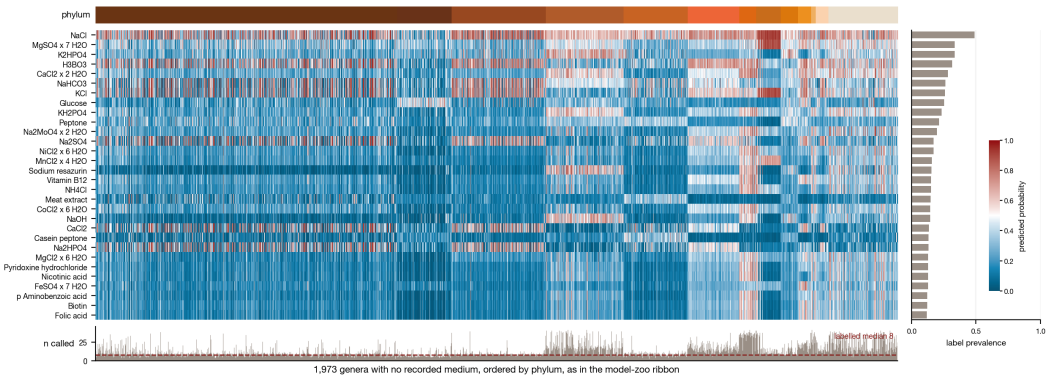

Figure 5: **Evaluating Cella's representation on unseen media.** Proposed media components for the 1,973 genera with no recorded medium. Rows: the 30 components most prevalent among labelled genera (water, yeast extract and agar omitted). Columns: genera ordered by phylum as in the model-zoo tree (top track). Cells: predicted probability (blue, confidently absent; white, 0.5; red, confidently present). Bottom: number of components called per genus at the per-component calibrated threshold, with the labelled median (8) dashed. Right: prevalence of each component among labelled genera.

4 Discussion

Cella is a heterograph model of metabolism built across five layers: taxon, organism, metabolites, components, and media. It is trained on a deliberately broad set of microbes sampled one-per-genus across 4,220 prokaryotic genera. Designed to generalize across a variety of industrial tasks, we demonstrate its effectiveness across nutrition and environment prediction. Cella can propose media components for genera with no experimental growth records and its internal embedding shows structure in both the phylogenetic and environmental dimension. Substantial work remains to extend its capabilities; we discuss caveats and future directions below.

We anticipate that future engineering of data embeddings is warranted. The genomic encodings adapted from GenomeSPOT (Barnum et al., 2024) represent only one method of encoding a whole microbial genome. As this block contains the most predictive signal for microbial traits, we anticipate that further feature engineering will improve model performance. Paths forward on this task include using frozen GenomeLMs (Bowden et al., 2026), biophysical parameters (Shen et al., 2025), or gene-content encodings such as EC, KEGG and COG annotations. The chemical components are all embedded using a frozen MolFormer, which also has several alternatives in the literature. Finally, we could encode reaction, gene, and enzyme information via the gene-protein-reaction (GPR) mapping present in GEMs.

Cella covers media concentration only coarsely. We segment media component levels into three: high, low, and trace, when in practice numerical concentrations are used. We made this design choice explicitly as more specificity was not supported by the quality of the literature databases; however, if a model is to predict a specific media recipe and not just components, it should include specific concentration levels, which will require more data on dose-response to media components.

Another limitation of our work is the provenance of the training set. We gather most of our growth records from the publicly available BacDive database, which is almost entirely positive growth records and tends to test organisms near their optimal conditions: above 55 °C, 97% of tested conditions record growth, though a much smaller fraction of species would grow across the prokaryotic tree. We had to perform significant data curation to synchronize chemical namespaces and properly split ions for component mapping; without this semi-manual curation effort, predictions are trivial (B12 proposed as a salt for cobalamin - itself with a different name) or meaningless. We anticipate a targeted experimental data generation campaign focused on clades of interest and conditions not well covered by the literature could improve performance substantially and tune results for a particular application.

5 AI Disclosure Statement

We used generative AI tools for research execution, in particular the PyTorch model implementation, train/test split design, and managing the execution of the GEM dataset creation pipeline. We verified the outputs of LLM-generated code manually by reading the code and rerunning key scripts. The documentation in the associated code submission was AI-generated and human-reviewed. We also used AI tools to create figures in Python in tandem with significant human editing. We used generative AI as a secondary reviewer for technical details in the text and to provide early feedback on research methodology and caveats during the writing process. In the writing of the paper we used AI to help rephrase and edit some sentences, as well as to double check the validity of claims with the literature and within our work. We did not use AI for related work retrieval, drafting full sections of the paper, generating synthetic datasets, or proving mathematical claims. We came up with the name Cella after a long brainstorming session with a generative AI agent.

References

Javad Aminian-Dehkordi, Mohammad Parsa, Andrew Dickson, and Mohammad RK Mofrad. Simba-gnn: mechanistic graph learning for microbiome prediction. *npj Systems Biology and Applications*, 12(1):8, 2025.

Ramy K. Aziz, Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M. Glass, Michael Kubal, et al. The RAST server: Rapid annotations using subsystems technology. *BMC Genomics*, 9:75, 2008. doi: 10.1186/1471-2164-9-75.

Tyler P Barnum, Alexander Crits-Christoph, Michael Molla, Paul Carini, Henry H Lee, and Nili Ostrov. Predicting microbial growth conditions from amino acid composition. *biorxiv*, pp. 2024–03, 2024.

Jake Bowden, Laurence Legon, and Satnam Surae. Canopy: A heterograph foundation model for metabolic engineering. *arXiv preprint arXiv:2607.06224*, 2026.

Garyk Brixi, Matthew G Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A Gonzalez, et al. Genome modelling and design across all domains of life with evo 2. *Nature*, 652(8112):1349–1361, 2026.

Tiffany J Callahan, Ignacio J Tripodi, Adrianne L Stefanski, Luca Cappelletti, Sanya B Taneja, Jordan M Wyrwa, Elena Casiraghi, Nicolas A Matentzoglu, Justin Reese, Jonathan C Silverstein, et al. An open source knowledge graph ecosystem for the life sciences. *Scientific Data*, 11(1): 363, 2024.

Shan Cong, Meng Zhang, Yu Song, Sihao Chang, Jing Tian, Hongji Zeng, and Hongchao Ji. Graph-sequence enhanced transformer for template-free prediction of natural product biosynthesis. *Patterns*, 6(8), 2025.

Xin Fang, Colton J. Lloyd, and Bernhard O. Palsson. Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nature Reviews Microbiology*, 18(12):731–743, 2020. doi: 10.1038/s41579-020-00440-4.

José P Faria, Filipe Liu, Janaka N Edirisinghe, Nidhi Gupta, Samuel MD Seaver, Andrew P Freiburger, Vibhav Setlur, Qizhi Zhang, Pamela Weisenhorn, Neal Conrad, et al. Modelseed v2: High-throughput genome-scale metabolic model reconstruction with enhanced energy biosynthesis pathway prediction. *bioRxiv*, pp. 2023–10, 2023.

Ramin Hasibi, Tom Michoel, and Diego A Oyarzún. Integration of graph neural networks and genome-scale metabolic models for predicting gene essentiality. *npj Systems Biology and Applications*, 10(1):24, 2024.

Christopher S. Henry, Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9):977–982, 2010. doi: 10.1038/nbt.1672.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

Xin Hu, Jiangli Duan, and Depeng Dang. Scalable aggregate keyword query over knowledge graph. *Future Generation Computer Systems*, 107:588–600, 2020.

Julia Koblitz, Philipp Halama, Stefan Spring, Vera Thiel, Christiane Baschien, Richard L Hahnke, Michael Pester, Jörg Overmann, and Lorenz Christian Reimer. Media dive: the expert-curated cultivation media database. *Nucleic Acids Research*, 51(D1):D1531–D1538, 2023.

Alexander Kroll, Sahasra Ranjan, Martin KM Engqvist, and Martin J Lercher. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature communications*, 14(1):2787, 2023a.

Alexander Kroll, Yvan Rousset, Xiao-Pan Hu, Nina A Liebrand, and Martin J Lercher. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature communications*, 14(1):4139, 2023b.

Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature catalysis*, 5(8):662–672, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Jianhan Liu, Guoshun Xu, Wuge Liu, Tuoyu Liu, Yanjun Li, Tao Tu, Huiying Luo, Ningfeng Wu, Bin Yao, Jian Tian, et al. Mediamatch: Prediction of bacterial growth on different culture media using the xgboost algorithm. *Microbial Biotechnology*, 18(10):e70245, 2025a.

Zicheng Liu, Siyuan Li, Zhiyuan Chen, Chang Yu, Qirong Yang, Yucheng Guo, Yujie Yang, Xiaoming Zhang, and Stan Z Li. Life-code: Central dogma modeling with multi-omics sequence unification. *arXiv preprint arXiv:2502.07299*, 2025b.

Sebastian Lobentanzer, Patrick Aloy, Jan Baumbach, Balazs Bohar, Vincent J Carey, Pornpimol Charoentong, Katharina Danhauser, Tunca Doğan, Johann Dreo, Ian Dunham, et al. Democratizing knowledge representation with biocypher. *Nature Biotechnology*, 41(8):1056–1059, 2023.

Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research*, 46(15):7542–7553, 2018.

Elena K Matveishina, Bartosz J Bartmanski, Sara Benito-Vaquerizo, and Maria Zimmermann-Kogadeeva. Gemsembler: consensus model assembly and structural comparison of genome-scale metabolic models across tools improve functional performance. *Msystems*, 10(10):e00574–25, 2025.

Sebastián N. Mendoza, Brett G. Olivier, Douwe Molenaar, and Bas Teusink. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, 20:158, 2019. doi: 10.1186/s13059-019-1769-1.

Amina Mollaysa, Artem Moskale, Pushpak Pati, Tommaso Mansi, Mangal Prakash, and Rui Liao. Biolangfusion: multimodal fusion of dna, mrna, and protein language models. *arXiv preprint arXiv:2506.08936*, 2025.

Sébastien Moretti, Van Du T Tran, Florence Mehl, Mark Ibberson, and Marco Pagni. Metanetx/mnxref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic acids research*, 49(D1):D570–D574, 2021.

Donovan H Parks, Pierre-Alain Chaumeil, Aaron J Mussig, Christian Rinke, Maria Chuvochina, and Philip Hugenholtz. Gtdb release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. *Nucleic Acids Research*, 54(D1):D743–D754, 2026.

Qizhi Pei, Zhimeng Zhou, Kaiyuan Gao, Jinhua Zhu, Yue Wang, Zun Wang, Tao Qin, Lijun Wu, and Rui Yan. Leveraging biomolecule and natural language through multi-modal learning: A survey. *arXiv preprint arXiv:2403.01528*, 2024.

Qizhi Pei, Zhimeng Zhou, Yi Duan, Yiyang Zhao, Wei Li, Han Guo, Liang He, Chengping Li, Chang-Yu Hsieh, Conghui He, et al. Biomatrix: Towards a comprehensive biological foundation model spanning the modality matrix of sequences, structures, and language. *arXiv preprint arXiv:2606.22138*, 2026.

Daniel Podlesny, Chan Yeong Kim, Shahriyar Mahdi Robbani, Christian Schudoma, Anthony Fullam, Lorenz C Reimer, Julia Koblitz, Isabel Schober, Anandhi Iyappan, Thea Van Rossum, et al. metatraits: a large-scale integration of microbial phenotypic trait information. *Nucleic Acids Research*, 54(D1):D835–D841, 2026.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

Brook E Santangelo, Harshad Hegde, J Harry Caufield, Justin Reese, Tomas Kliegr, Lawrence E Hunter, Catherine A Lozupone, Christopher J Mungall, and Marcin P Joachimiak. Kg-microbe-building modular and scalable knowledge graphs for microbiome and microbial sciences. *GigaScience*, pp. giag077, 2026.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pp. 593–607. Springer, 2018.

Isabel Schober, Julia Koblitz, Joaquim Sardà Carbasse, Christian Ebeling, Marvin Leon Schmidt, Adam Podstawka, Rohit Gupta, Vinodh Ilangovan, Javad Chamanara, Jörg Overmann, et al. Bac dive in 2025: the core database for prokaryotic strain data. *Nucleic acids research*, 53(D1):D748–D756, 2025.

Kieren Sharma, Lucia Marucci, and Zahraa S Abdallah. Fluxgat: integrating flux sampling with graph neural networks for unbiased gene essentiality classification. *arXiv preprint arXiv:2403.18666*, 2024.

Yuxin Shen, Grzegorz Kudla, and Diego A Oyarzún. Improving the generalization of protein expression models with mechanistic sequence information. *Nucleic Acids Research*, 53(3):gkaf020, 2025.

Ines Thiele and Bernhard Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010. doi: 10.1038/nprot.2009.203.

Ning Wang, Jiang Bian, Yuchen Li, Xuhong Li, Shahid Mumtaz, Linghe Kong, and Haoyi Xiong. Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence*, 6(5):548–557, 2024.

Liang Xu, Emily Zakem, and JL Weissman. Improved maximum growth rate prediction from microbial genomes by integrating phylogenetic information. *Nature Communications*, 16(1):4256, 2025.

Yuan Zhang, Xin Sui, Feng Pan, Kaixian Yu, Keqiao Li, Shubo Tian, Arslan Erdengasileng, Qing Han, Wanjing Wang, Jianan Wang, et al. A comprehensive large-scale biomedical knowledge graph for ai-powered data-driven biomedical research. *Nature Machine Intelligence*, 7(4):602–614, 2025.

Weihe Zhong, Ziduo Yang, and Calvin Yu-Chian Chen. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14(1): 3009, 2023.

Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, Jike Wang, et al. Advancing biomolecular understanding and design following human instructions. *Nature Machine Intelligence*, 7(7):1154–1167, 2025.

Johannes Zimmermann, Christoph Kaleta, and Silvio Waschina. Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome biology*, 22(1):81, 2021.

A APPENDIX

Contents of the appendix

Item	Content	Page
Figures		
Figure S1	GEM pipeline overlap: metabolites shared across the three reconstruction pipelines, and organism-level agreement on the shared vocabulary	15
Figure S2	Carbon-source labels are noise-limited rather than volume-limited: learning curves in which AUROC climbs while macro F_1 plateaus	15
Tables		
Table S1	Downstream tasks: label counts, row counts, and the labels retained at a minimum of 25 positives	16
Table S2	Provenance of cultivation parameters: evidence type, source, taxonomic match level and conflict flag, per parameter	17
Table S3	Source vocabulary: records contributed by each source, with match level	18
Table S4	Habitat class definitions and prevalence: the twelve categories, with example raw isolation-source strings	19
Table S5	Nutrition tasks, full arm sweep	20
Table S6	Environment tasks, full arm sweep	21
Table S7	Scoring tasks: strain similarity and medium substitution, each against the control it has to beat	22
Appendices		
Appendix A.1	Primer: genome-scale metabolic modelling	23

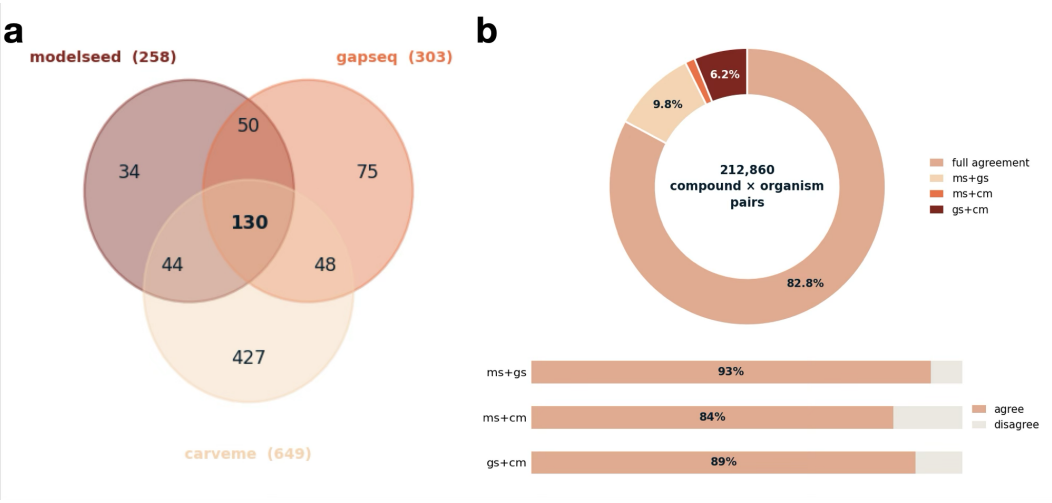

Figure S1: **GEM pipeline overlap.** **(A)** A Venn diagram of the 3 GEM construction pipelines showing overlapping metabolites. Only 130/total are shared between all 3. **(B)** Agreement between pipelines at the organism levels. Over 212,860 organism/compound pairs (for 130 shared components), 82% agree across all 3 pipelines. Pipeline pairwise bars are below.

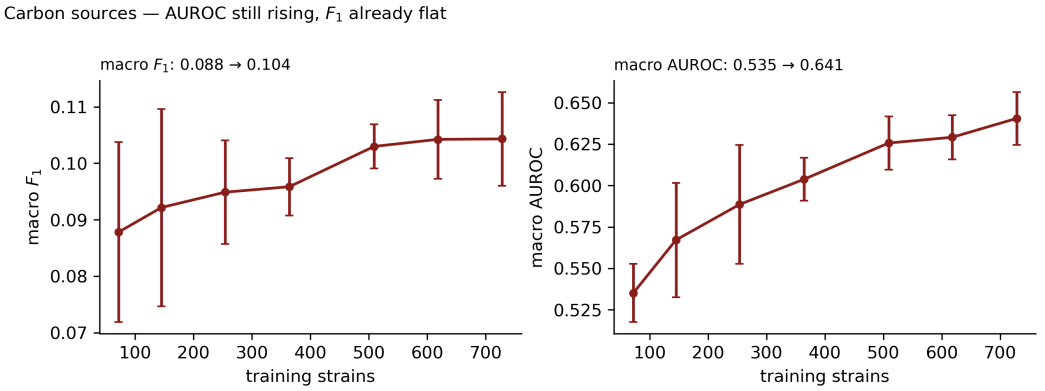

Figure S2: **Carbon-source labels are noise-limited, not volume-limited.** Learning curves for a genome-only random forest (GenomeSPOT features; 200 trees, depth 15) on the carbon source tasks, mean across three seeds. AUROC climbs from 0.535 to 0.641 while macro F_1 plateaus at 0.10. This is due to noise in the positive-unlabelled labels: unrecorded pairs are untested, not negative.

Table S1: **Downstream tasks.** "N" is the number of distinct labels observed at least once; "≥25" is those retained at a minimum of 25 positives, below which a held-out fold yields too few positives for F_1 to be defined. Rows exceed genomes where an organism has several recorded media.

Task	Predicts *(node type)*	Rows	Genomes	N (≥ 25)	Label split min / med / max
Nutrition ladder					
Carbon sources	*Metabolite*	911	911	301 (72)	26 / 75 / 433
Media components	*Component*	2,776	2,247	70 (70)	8 / 315 / 2,726
Cultivation media	*Medium*	2,534	2,534	678 (18)	26 / 55 / 326
Environment suite					
Temperature optimum	regression	2,879	2,879	1	4/30/97 °C
pH optimum	regression	1,093	1,093	1	1.68/7.25/9.88 pH
Salinity optimum	regression	807	807	1	0/1.50/20.45 % NaCl
Oxygen tolerance	4-class (aer/anaer/ facul/micro)	2,550	2,550	4	56/31/9/3
Habitat (multi-label)	multi-label	2,784	2,784	12	see Table S4
Habitat (12-class)	single-label	1,612	1,612	12 (11)	see Table S4
Envelope, temperature	binary	12,535	2,882	1	15% no-growth
Envelope, NaCl	binary	3,027	1,036	1	52% no-growth
Envelope, pH	binary	1,341	977	1	6% no-growth
Scoring					
Strain similarity	*Organism pairs*	2,746	2,746	—	400k of 3.77M pairs
Medium proposals	*Component*	1,973	—	—	previously unlabelled

Table S2: **Provenance of cultivation parameters in the Cella training collection** ($n = 4{,}220$ genomes). Each parameter carries per-record evidence type, source, taxonomic match level and a conflict flag raised when sources disagree. Counts are records with a non-missing value; percentages are of that count. *Mixed* indicates a value supported by both experimental and predicted sources, *derived* a value computed from other measured fields. Match level is reported against whatever the predicting or measuring source addressed, so a fully predicted parameter can still be strain-level: doubling time is predicted per genome and matches at strain, while atmosphere is propagated from species and genus annotations and matches at neither. Only parameters with a substantial experimental fraction are used as prediction targets; those marked † are excluded, since training a genome-based model on genome-based predictions is circular.

		Evidence type				Strain-	
Parameter	n	Exp.	Mixed	Deriv.	Pred.	level	Conflict
Temperature, optimum (°C)	4,036	2,879 (71%)	0	216	941 (23%)	95%	18%
Temperature, min (°C)	4,082	83 (2%)	1,282	0	2,717 (67%)	67%	4%
Temperature, max (°C)	4,102	93 (2%)	1,281	0	2,728 (67%)	68%	3%
pH, optimum	4,064	1,093 (27%)	0	169	2,802 (69%)	96%	4%
pH, min	4,091	76 (2%)	950	0	3,065 (75%)	76%	2%
pH, max	4,091	75 (2%)	949	0	3,067 (75%)	76%	1%
Salinity, optimum (%)	3,023	807 (27%)	0	34	2,182 (72%)	99%	6%
Salinity, min (%)	2,827	36 (1%)	618	0	2,173 (77%)	77%	1%
Salinity, max (%)	3,700	109 (3%)	600	0	2,991 (81%)	83%	2%
Oxygen requirement	4,117	2,550 (62%)	33	0	1,534 (37%)	52%	54%
Atmosphere†	3,719	0 (0%)	0	0	3,719 (100%)	0%	0%
Doubling time (h)†	4,171	0 (0%)	0	0	4,171 (100%)	100%	0%
Cultivation media	2,534	2,534 (100%)	0	0	0 (0%)	99%	0%
Carbon sources	963	911 (95%)	0	0	52 (5%)	90%	0%
Isolation source	2,980	2,980 (100%)	0	0	0 (0%)	99%	0%

Table S3: **Source vocabulary**, with total records contributed across all parameters in Table S2. Each source record was linked to a genome; a species- or genus-level match assigns a taxon trait to a specific strain and is a weaker claim than an exact-strain match.

Source	Type	Records	Match level
genomespot	genome-based prediction Barnum et al. (2024)	22,889	strain
bacdive_exact_strain	literature, curated (Schober et al., 2025)	11,958	strain
metatraits_species	aggregated trait resource (Podlesny et al., 2026)	11,609	species
phydon_predicted	genome-based prediction (Xu et al., 2025)	4,171	strain
bacdive_type_strain	literature, curated, type strain	1,271	strain
metatraits_genus	aggregated trait resource	548	genus
bacdive_species	literature, curated, species-level	54	species

Table S4: **Habitat class definitions and prevalence.** We categorize BacDive's `isolation_sources` field (4,134 distinct strings across 2,980 genomes) into one or more of 12 categories.

	Multi-label		Single-label			
Class	n	%	n	%	Str.	Example raw values
marine	701	25.2	234	14.5	780	seawater; marine sediment; From seawater
soil	652	23.4	340	21.1	736	Soil; soil; Environment, Soil; From soil
sediment	582	20.9	74	4.6	673	marine sediment; sediment; marine mud
freshwater	512	18.4	170	10.5	641	lake water; freshwater (river, lake, pond)
human	397	14.3	308	19.1	572	Human feces; human faeces; Human, Blood
plant	367	13.2	125	7.8	528	rice paddy soil; Environment, Woodland soil
animal	306	11.0	154	9.6	483	Pig feces; rumen of cow; caecal content, mouse
thermal	255	9.2	72	4.5	324	hot spring; deep-sea hydrothermal vent
hypersaline	200	7.2	23	1.4	232	soda lake; saline soil; hypersaline environment
wastewater	192	6.9	40	2.5	254	activated sludge; sewage sludge; anaerobic digester
food	83	3.0	42	2.6	133	raw milk; Beer; Butter
built_air	62	2.2	30	1.9	99	Air; Mural painting; Medieval wall painting
matched $\geq$ 1 class						2,784 of 2,980 (93.4%)
unmatched, dropped						196 (6.6%); 448 distinct strings
multi-environment, excluded from single-label						1,172

Table S5: **Nutrition tasks.** Macro F_1 and macro AUROC under organism-grouped splits, mean over folds. Labels are experimental only. Two controls bound the task and are set above the rule: popularity predicts each label at its base rate, and lineage only is a one-hot encoding of domain, phylum, class, order and family. Cella is the graph model, scored on its own protocol (three seeds, seen labels), so its row is not directly comparable with the feature arms above it. On carbon sources every arm here is averaged over the 52 of 72 labels that resolve to a metabolite node, which is the set the graph model scores; the restriction moves the feature arms by at most 0.006 F_1.

	Carbon sources		Media components		Cultivation media	
Arm	n=911 F_1	52 labels AUROC	n=2,776 F_1	70 labels AUROC	n=2,534 F_1	18 labels AUROC
popularity (control)	0.000	0.500	0.026	0.500	0.000	0.500
lineage only (control)	0.246	0.630	0.424	0.833	0.263	0.898
genome	0.108	0.635	0.472	0.867	0.301	0.913
genome + taxonomy	0.109	0.637	0.476	0.867	0.297	0.914
genome + flux	0.092	0.636	0.457	0.868	0.274	0.910
genome + panels (ModelSEED)	0.113	0.646	0.484	0.874	0.289	0.919
genome + panels (all three)	0.106	0.649	0.471	0.878	0.262	0.920
genome + panels (agreement)	0.109	0.639	0.473	0.874	0.301	0.915
panels (ModelSEED) alone	0.139	0.639	0.476	0.850	0.262	0.902
panels (agreement) alone	0.129	0.632	0.435	0.839	0.275	0.905
all modalities	0.095	0.644	0.464	0.872	0.276	0.919
Cella (graph)	0.229	0.612	0.502	0.861	0.351	0.922

Table S6: **Environment tasks.** R^2 for the three optima; macro F_1 and macro AUROC for classification and growth envelopes. Reported scores are means over three seeds, with organism-grouped splits. The lineage-only control is the floor these traits must clear. The envelope tasks additionally include a value-only control because the tested condition is an input. Cella GNN is the graph model without the reaction layer. Best arm per metric column in **bold**, including controls. Thresholds for F_1 are fitted on a calibration slice of held-out organisms for the multi-label habitat task; the single-label and binary tasks decide by $\arg\max$, equivalently a threshold of 0.5.

A. Optima

	Temperature $n = 2{,}879$	pH $n = 1{,}093$	Salinity $n = 807$
Arm	R^2	R^2	R^2
phylum only (control)	0.565	0.089	0.404
genome (GenomeSPOT)	0.787	0.436	0.588
genome + taxonomy	0.787	**0.439**	0.588
genome + panels	0.786	0.438	0.589
genome + tax + panels	0.787	0.436	**0.590**
Cella GNN	**0.819**	0.391	0.499

B. Classification

	Oxygen tolerance $n = 2{,}550$		Habitat (multi-label) $n = 2{,}784$		Habitat (12-class) $n = 1{,}612$	
Arm	F_1	AUROC	F_1	AUROC	F_1	AUROC
phylum only (control)	0.567	**0.871**	0.358	0.724	0.264	0.748
genome (GenomeSPOT)	0.631	0.861	0.416	0.785	0.354	0.828
genome + taxonomy	0.630	0.861	0.421	0.789	0.353	0.828
genome + panels	0.635	0.865	0.426	0.792	0.347	**0.835**
genome + tax + panels	**0.644**	0.867	**0.426**	**0.792**	**0.361**	0.833
Cella GNN	0.602	0.836	0.421	0.786	0.361	0.808

C. Growth envelopes

	Temperature $n = 12{,}535$		NaCl $n = 3{,}027$		pH $n = 1{,}341$	
Arm	F_1	AUROC	F_1	AUROC	F_1	AUROC
value only (control)	0.786	0.903	0.870	0.916	0.348	0.614
phylum only (control)	0.796	**0.909**	**0.884**	**0.936**	0.586	**0.801**
genome (GenomeSPOT)	0.719	0.901	0.858	0.930	0.520	0.756
genome + taxonomy	0.534	0.849	0.797	0.909	0.504	0.747
genome + panels	0.563	0.865	0.796	0.906	0.519	0.716
genome + tax + panels	0.481	0.837	0.747	0.888	0.489	0.724
Cella GNN	**0.806**	0.878	0.848	0.901	**0.602**	0.586

Table S7: **Scoring tasks.** Each is a ranking problem scored against the control it has to beat, not against chance. Strain similarity reports Spearman ρ against functional similarity, the partial ρ controlling for patristic distance, and the functional similarity of the neighbour each axis selects (last column). For medium substitution the last column is instead AUROC: a positive is a medium pair one organism is recorded as growing in (635 instances over 273 pairs), a negative pairs that organism with a medium it has no record in (1,270), and each score ranks them. Two scores fall below 0.5 because the comparison is confounded by medium size – co-occurring media are small complex broths (5.5 uptake species, 48% undefined components) while the sampled negatives are larger defined recipes (14.4, 19%), and a larger medium covers more of anything – so this evaluation cannot presently separate a good substitution score from a bad one.

Task	Predictor / arm	ρ	ρ | patristic	NN transfer / AUROC
Strain similarity (media components), n=2,746 strains, 400,000 pairs				
	Cella embedding (no phylogeny)	0.148	0.146	0.346
	Cella embedding	0.162	0.161	0.362
	graph, GEM assertions	0.105	0.108	0.336
	graph, GEM requirements	0.035	0.038	0.181
	genome features	0.138	0.136	0.364
	patristic distance	0.019	—	0.364
	Mash distance	0.083	0.087	0.341
	taxonomic rank	0.041	0.073	0.306
	random pair (floor)	—	—	0.117
	best fixed neighbour (control)	—	—	0.164
Strain similarity (carbon sources), n=778 strains, 302,253 pairs				
	Cella embedding (no phylogeny)	0.071	0.086	0.151
	Cella embedding	0.089	0.101	0.160
	graph, GEM assertions	0.097	0.105	0.154
	graph, GEM requirements	0.063	0.060	0.124
	genome features	0.042	0.052	0.155
	patristic distance	-0.025	—	0.155
	Mash distance	0.039	0.050	0.139
	taxonomic rank	0.046	0.069	0.155
	random pair (floor)	—	—	0.093
	best fixed neighbour (control)	—	—	0.195
Medium substitution, 635 (organism, pair) instances over 273 distinct medium pairs				
	Cella, medium similarity	—	—	0.617
	Cella, organism–medium	—	—	0.390
	organism-conditioned score	—	—	0.395
	composition overlap (control)	—	—	0.424
	shared-component count (control)	—	—	0.608

A.1 PRIMER: GENOME-SCALE METABOLIC MODELS

Genome-scale metabolic models (GEMs) are widely used by bioengineers to simulate the metabolic state of an organism under different medium conditions, genetic perturbations, and environmental conditions Fang et al. (2020). GEMs encode the network of metabolic reactions for a particular organism as a stoichiometric matrix $\mathbf{S}$. Using a mass balance relation, the model can be expressed as:

$$\mathbf{Sv} = \frac{d\mathbf{x}}{dt}, \tag{1}$$

where $\mathbf{v}$ is the vector of reaction fluxes and $\mathbf{x}$ is the vector of metabolite concentrations. GEMs then make a steady-state assumption that intracellular metabolites remain constant across the timescale of interest, such that

$$\mathbf{Sv} = 0. \tag{2}$$

The feasible fluxes are further constrained by bounds taken from literature:

$$v_i^{\min} \leq v_i \leq v_i^{\max}. \tag{3}$$

Creating a GEM starts with a genome sequence that is annotated to identify protein-coding genes and assign them putative functions. These functions are then matched to reactions in biochemical databases such as KEGG, MetaCyc, BiGG or ModelSEED (Thiele & Palsson, 2010). Each reaction is linked to its gene(s) through gene–protein–reaction (GPR) rules, which use "AND" for enzyme complexes and "OR" for isozymes; these rules allow the effect of gene knockouts to be simulated. Reactions are assigned to compartments such as the cytosol, periplasm or extracellular space, and compartments are connected by transport reactions. In a well-curated GEM, every reaction balances atoms and charge, so that the model cannot create energy or matter from nothing. A biomass reaction consumes the assumed building blocks of a new cell (amino acids, nucleotides, lipids, cofactors and the ATP cost of growth), and its flux represents the growth rate. Exchange reactions define which metabolites can be taken up from or secreted to the environment, and therefore specify the growth medium. Draft networks usually cannot grow because they lack some of the reactions or pathways needed to produce all biomass components. In a gap-filling step, a small number of additional reactions is added to restore growth on a given medium (Thiele & Palsson, 2010).

Manual reconstruction can take months to years (Thiele & Palsson, 2010), so several tools automate the process. ModelSEED Henry et al. (2010) annotates the genome with RAST (Aziz et al., 2008) and maps the resulting functional roles to reactions in the ModelSEED biochemistry database. It uses a template (e.g. Gram-negative or Gram-positive) that also supplies a default biomass reaction, and then gap-fills the draft to allow growth on a chosen medium. gapseq (Zimmermann et al., 2021) works bottom-up. It searches the genome against reference protein sequences to predict which pathways and transporters are present, and builds the model on ModelSEED biochemistry. During gap filling, it prefers reactions with at least weak sequence support and also adds reactions needed for the organism's known carbon sources and fermentation products. CarveMe (Machado et al., 2018) works top-down. It starts from a curated, balanced universal model derived from BiGG and scores each reaction by how well the organism's proteins match known enzymes. An optimisation step then carves out a subnetwork that retains well-supported reactions while still producing biomass. All three tools produce draft models whose quality depends on annotation and database coverage, and the same genome can yield noticeably different models depending on the tool, template and gap-filling medium (Mendoza et al., 2019).