

PNNL-39899

Predicting Microbial Gene Fitness with a Biological Knowledge Graph

September 2026

Kaizad Patel
Winston Anthony
Sam Donald
Adam Deutschbauer
Sumit Purohit
Robert Egbert

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

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

Printed in the United States of America

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

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

Predicting Microbial Gene Fitness with a Biological Knowledge Graph

September 2026

Kaizad Patel
Winston Anthony
Sam Donald
Adam Deutschbauer
Sumit Purohit
Robert Egbert

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

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Predicting the condition-dependent fitness effects of gene inactivation remains a central challenge in microbial phenomics. We introduce GIMME (**G**raph **I**nterference for **M**icrobial **D**ark **M**atter **E**xploration), a knowledge graph integrating 7.94 million RB-TnSeq fitness measurements spanning 55,421 genes, 1,770 experiments, and 10 *Pseudomonas* strains with experimental metadata, metabolic annotations, and pretrained text and protein-sequence representations. We train GFM1 (GIMME Fitness Model 1), a heterogeneous graph attention network (GAT) to predict gene-inactivation fitness effects under three held-out evaluation settings: experiments, genes, and ortholog clusters. Message passing over a subset of the knowledge graph increases performance: In the held-out-gene setting, message passing reduced mean squared error (MSE) by 10.60% and increased Pearson correlation by 12.44% relative to a multi-layer perceptron (MLP) using the same multimodal node-encoder design and decoder template. Performance gains were concentrated among highly connected, conditionally essential genes, where GFM1 predicts %12.5 more true positives, but this advantage disappears under the more restrictive cluster-aware split, where it loses accuracy. Within the subset of test observations whose genes had level-1 EC annotations, the gene-split GFM1 had lower mean absolute error (MAE) than the MLP in all seven classes and showed greater predictive accuracy than a genome-scale metabolic model in a matched *Pseudomonas putida* benchmark. Together, these results indicate that message passing can leverage knowledge-graph structure to improve gene-fitness prediction when informative relationships to training genes are available. The GIMME knowledge graph and GFM1 represent a step forward in predictive inference of microbial metabolism and a new tool for *in silico* phenotype prediction.

Summary

GIMME is a large heterogeneous knowledge graph built from millions of RB-TnSeq fitness measurements across *Pseudomonas* species, combined with metabolic annotations, protein-sequence embeddings, and text-derived information. Using this integrated graph, we train GFM1, a graph neural network designed to predict condition-specific gene fitness effects. GFM1 leverages relational structure – such as orthology, pathway membership, and shared experimental context – to improve predictions relative to a multimodal MLP baseline, especially for genes with strong or conditionally essential fitness phenotypes. The model is evaluated under several held-out scenarios (experiments, genes, and gene-family clusters), showing substantial gains when relational information is available, and diminished advantages when related genes are intentionally partitioned away. This work demonstrates that combining large-scale fitness data with graph-based machine learning enables more accurate inference of gene-environment interactions, and can complement mechanistic metabolic models for prioritizing genes or conditions for experimental follow-up.

Acknowledgments

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

Acronyms and Abbreviations

GIMME: Graph Inference for Microbial Dark Matter Exploration

GFM1: GIMME Fitness Model 1

RB-TnSeq: Randomly Barcoded Transposon Sequencing

KG: Knowledge Graph

GNN: Graph Neural Network

GAT: Graph Attention Network

MLP: Multi-Layer Perceptron

PLM: Protein Language Model

GSMM: Genome-Scale Metabolic Model

FBA: Flux Balance Analysis

EC: Enzyme Commission

MAE: Mean Absolute Error

MSE: Mean Squared Error

AUROC: Area Under the Receiver Operating Characteristic

AUPRC: Area Under the Precision-Recall Curve

Contents

Abstract.....	ii
Summary	iii
Acknowledgments.....	iv
Acronyms and Abbreviations.....	v
1.0 Introduction	1
1.1 Related Work	2
2.0 Methods	3
2.1 Datasets	3
2.2 Data Processing	4
2.3 Multi-Modal Node Encoder.....	4
2.4 Heterogeneous Message Passing	5
2.5 Prediction Head	5
2.6 Training and Evaluation	6
2.7 GSMM performance comparison	6
3.0 Results	7
3.1 Model Predictive Performance	7
3.1 Graph Performance Gain in High Connectivity and Essentiality Regime	7
3.2 Regression advantage disappears under the cluster-aware split.....	9
3.3 GFM1 outperforms a GSMM when predicting fitness growth defects	9
4.0 Discussion	11
4.1 Fitness prediction under cluster-aware generalization remains challenging	11
4.2 Relational structure improves prediction of conditionally essential phenotypes	11
4.3 GFM1 complements mechanistic genome-scale metabolic models.....	11
4.4 Limitations and future work	12
5.0 Conclusion	13
6.0 References.....	14

Figures

- Figure 1. Graph schema used by GFM1. FITNESS edges connect gene and experiment subnetworks. Gene context includes KEGG orthogroups, protein domains, Swiss-Prot annotations, and metabolic knowledge, whereas experiment context represents conditions and metadata. All structural edges are treated as bidirectional during training. 3
- Figure 2. Architecture schematic, detailing the Multi-modal encoders, variable depth GNN, and prediction head. Note that within the prediction head, the concatenation (cat) fusion method is displayed, and that the larger KG is represented by a simple three node graph containing a single gene, experiment and media node. 5
- Figure 3. Top: Mean absolute error rises as gene network connectivity increases during cold-start gene inference. Dots indicate MAE point estimates. Bottom: Bar chart indicating the graph gain at each connectivity bin. 9
- Figure 4. Cluster Split Prediction performance across enzyme classification groups is heavily influenced by fraction of essential conditions. Left: Conditionally essential fitness cell fractions for all EC groups across each split protocol. Right: Graph gain for all EC groups across split protocols. Only 27-30% of test prediction rows could be assigned a canonical level-1 EC class. 10

Tables

- Table 1. Mean and standard deviation of test performance across confirmation runs for the selected GAT and matched MLP baseline. The architecture-screen seed is excluded. 8
- Table 2. Comparison of best MLP and GAT performance across data splits in the high essentiality stratum. "Ess." refers to conditionally essential fitness pairs per split. 8
- Table 3. Comparison of flux balance analysis (FBA) and GFM1 predictions against experimental RB-TnSeq measurements for cold-start genes in *Pseudomonas putida* KT2440. Base rate of conditionally essential genes is 0.052. 10

1.0 Introduction

The fitness effect of a genetic perturbation depends on the environment in which it is measured. Characterizing these genotype-environment interactions is important for understanding microbial physiology and predicting phenotypes under unmeasured conditions. Predicting these interactions could reduce the experimental burden of mapping genotype-phenotype fitness landscapes and enable *in silico* inference for unmeasured genes and conditions.^{1,2} High throughput genome-wide fitness assays provide a powerful method for rapid experimental assessment of the function and metabolic plasticity of microbes. One such technique, randomly barcoded transposon sequencing (RB-TnSeq), uses barcoded transposon insertions to generate pooled mutant libraries containing disruptions across much of the genome.³ RB-TnSeq assays have been used to identify conditionally important genes, generate hypotheses about gene function, and associate poorly characterized genes with biological pathways through correlated fitness profiles.⁴ The Fitness Browser provides a large collection of RB-TnSeq measurements from 9,532 experiments spanning nutrient and stress conditions across diverse bacterial lineages.⁵ These measurements provide a foundation for training models that predict fitness effects for gene-condition combinations not observed during training.

The Gram-negative genus *Pseudomonas* includes environmentally, industrially, and clinically important species with diverse metabolic capabilities.^{6,7} Despite extensive experimental and computational characterization, a substantial fraction of genes in *Pseudomonas putida* remain without characterized functions.⁸ The combination of large fitness datasets, established metabolic models, and persistent annotation gaps makes *P. putida* and related pseudomonads useful systems for evaluating predictive models of genotype-environment interactions.

Conventional computational annotation commonly relies on sequence similarity to characterized homologs.⁹ Profile hidden Markov models extend this approach by capturing position-specific patterns of conservation and variation within protein families.¹⁰ Although these methods are valuable for functional annotation, functional similarity alone does not determine the fitness effect of gene disruption under a particular environmental condition. Many poorly characterized proteins, sometimes described as “protein dark matter,” also have little detectable similarity to experimentally characterized proteins.¹¹ Genome-scale metabolic models (GSMMs) provide a mechanistic approach by representing gene-reaction relationships and using constraint-based optimization to simulate metabolic flux and growth.^{12,13} GSMMs can simulate the effects of single-gene disruptions, but their predictions are restricted to genes and processes represented in the metabolic network.¹⁴

Protein language models (PLMs) for protein sequences are promising candidates to address the “dark matter problem”.¹⁵ PLMs encode protein sequences as dense representations that capture statistical patterns associated with structural and functional properties. These representations can be combined with fitness measurements, experimental metadata, and metabolic annotations to predict condition-specific phenotypes.¹⁶ Knowledge graphs (KGs) provide a structured framework for integrating such heterogeneous information by representing biological entities and their relationships under an explicit schema.^{17,18} Graph neural networks (GNNs) can then aggregate information across relationships such as orthology, pathway membership, gene-reaction associations, and shared experimental context.¹⁹⁻²¹ Message passing could therefore allow a predictive model to use relational information integrated from resources such as KEGG,²² MetaCyc,²³ and Swiss-Prot.²⁴ Whether this graph structure improves fitness prediction beyond the underlying sequence, text, and metadata features, however, remains an empirical question.

Here, we introduce GIMME (Graph Inference for Microbial Dark Matter Exploration), a heterogeneous biological knowledge graph integrating RB-TnSeq fitness measurements with gene, experiment, metabolic, text-derived, and protein-sequence information. We use GIMME to train GFM1 (GIMME Fitness Model 1), a heterogeneous graph neural network for predicting condition-specific gene fitness. We compare graph-based models with a multimodal multilayer perceptron (MLP) that uses the same prediction head but does not perform message passing. Models are evaluated under three held-out settings: experiments, genes, and family-style gene clustering, to test generalization to unobserved gene-experiment combinations. We also examine how graph-model performance varies with network connectivity, conditional essentiality, and Enzyme Commission class, and compare GFM1 with a genome-scale metabolic model in a *P. putida* benchmark. Message passing improves prediction for held-out experiments and genes, but its regression advantage is not significant during the cluster-aware split. Together, these analyses demonstrate biological knowledge-graph structure provides predictive information beyond multi-modal node features and establishes a framework for graph-based microbial fitness prediction.

1.1 Related Work

Prior machine-learning approaches to microbial gene function and fitness have primarily used either protein-sequence representations or experimental phenotype data. PLMs such as ESM-3 encode amino acid sequences as dense representations, and recent work has combined PLM embeddings with GNNs to predict gene essentiality.^{25,26} Sequence-only models, however, do not explicitly represent the environmental conditions under which fitness is measured. Conversely, phenotype-based methods capture relationships within observed assay data. Borchert et al., for example, identified 84 co-fitness modules from a 179-condition *P. putida* fitness matrix using independent component analysis,²⁷ but such matrix-based approaches do not directly generalize to genes, strains, or conditions absent from the observed data.

Knowledge graphs can integrate sequence-derived, experimental, and functional information. Heterogeneous GNNs have been applied to biomedical knowledge graphs for drug-target, drug-drug interaction, and essentiality prediction tasks;²⁸⁻³¹ FuseLinker additionally combines frozen language-model text embeddings with GNN-based link prediction.³² Our framework instead targets condition-specific microbial gene fitness by integrating text-derived annotations, protein-sequence representations, experimental context, and heterogeneous graph structure.

2.0 Methods

2.1 Datasets

We obtained 7,937,507 gene-experiment fitness measurements spanning 55,421 genes, 1,770 experiments, and 10 *Pseudomonas* strains from the Fitness Browser including growth media composition and experimental metadata such as pH and temperature. Gene-level fitness scores are calculated from changes in relative mutant-barcode abundance and reported on a log2 scale.⁵ We define a conditional fitness defect using the "extreme fitness value" of ≤ -2 from Borchert et al.²⁷

The Fitness Browser is partially integrated with three external reference databases: SEED, MetaCyc,²³ KEGG²² and the UniProt EC number. We integrated the full EC hierarchy using the EC RDF,³³ and added MetaCyc reactions and compounds. The final KG contains 141,947 nodes across 19 node types, 884,487 structural edges across 27 relationship types, and 7,937,507 FITNESS supervision edges. Figure 1 shows the training subgraph; the released KG includes the complete node and relationship inventory. Other than the FITNESS edges, no other edges included in model training are derived from fitness observations. No text or annotations embedded into training data mentions essentiality or fitness outcomes.

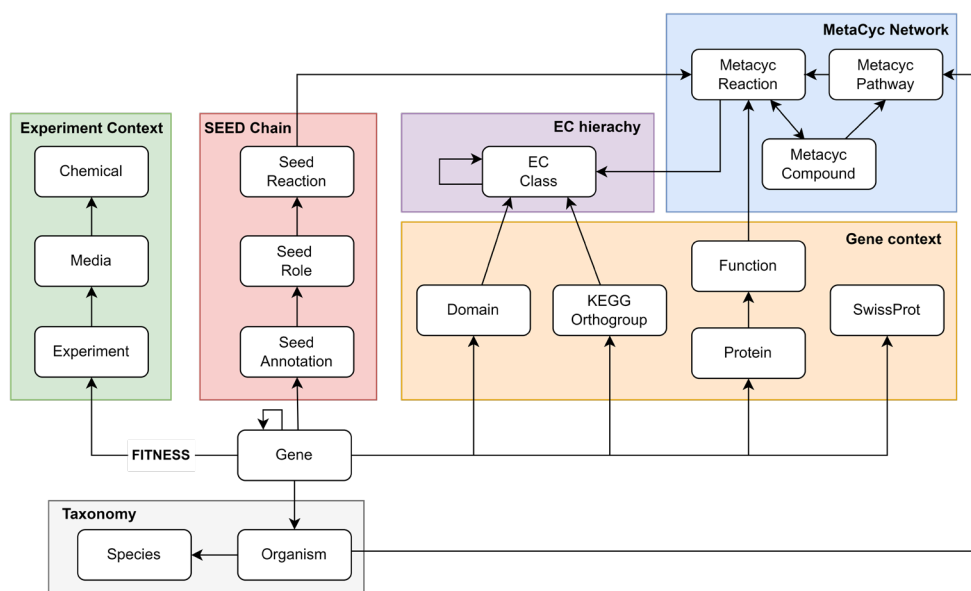


Figure 1. Graph schema used by GFM1. FITNESS edges connect gene and experiment subnetworks. Gene context includes KEGG orthogroups, protein domains, Swiss-Prot annotations, and metabolic knowledge, whereas experiment context represents conditions and metadata. All structural edges are treated as bidirectional during training.

2.2 Data Processing

To reduce the influence of extreme fitness values, we fit a quantile transformer on the training fold that maps the target distribution to an approximate standard normal using 10,000 quantile bins. Transformed targets were used for model optimization and validation-based early stopping. We inverse-transformed predictions before computing all reported metrics on the original $\log_2$ fitness scale. We evaluated three split protocols representing generalization to held-out experiments, genes, and gene clusters. Split counts below are based on assignment seed 42.

- **Experiments:** We assigned the 1,770 experiments and their associated fitness cells to one partition. The resulting cell counts were 6,371,831 training, 782,838 validation, and 782,838 test cells. Experiment nodes and their target-independent metadata remain available, but all FITNESS edges associated with validation and test experiments are withheld from training.
- **Genes:** We assigned the 55,421 genes and their associated fitness cells to one fold. The resulting cell counts were 5,544,834 training, 1,194,386 validation, and 1,198,287 test cells. Every fitness measurement inherits its gene’s fold. Held-out Gene nodes and target-independent graph context remain available, but their validation/test FITNESS labels are excluded from training.
- **Clusters:** To reduce homology across partitions, we assigned each gene to exactly one cluster using the following priority order: KEGG orthogroup, Swiss-Prot group, connected component of ORTHOLOGOUS edges, or singleton, emulating CAFA gene clustering.^{22,34} The resulting cluster assignment partitions the fitness cells into 6,310,761 training, 820,809 validation, and 805,937 test cells. We removed ORTHOLOGOUS and BEST HIT edges connecting genes between partitions either directly or through intermediary nodes. This is a cluster-aware protocol rather than the complete exclusion of all related genes.

2.3 Multi-Modal Node Encoder

As shown in Figure 2, we encode each node within the heterogeneous graph by projecting every available feature type through its own branch, then fusing the results to a common hidden vector of width H . Each branch b produces a raw vector $\mathbf{x}_v^{(b)}$ for node v , and we define A_k as the set of branches active on node type k . The four supported branches are:

- **Text** ($\mathbb{R}^{d_T}$). A frozen text representation embedded with gemini-embedding-001³⁵ into a fixed 3072-dimensional vector.
- **Numeric** ($\mathbb{R}^{2n_k}$). For a node type k with n_k scalar properties, the branch takes the vector of z-normalized values paired with n_k missingness binary flags indicating if each value is present.
- **Categorical** ($\mathbb{R}^{P_p \cdot e_p}$). Each categorical property p has a learned embedding table $\mathbf{E}_p$ of width $e_p = \max(2, \lceil \sqrt{|c_p|} \rceil)$, where $|c_p|$ is the property’s cardinality, and are concatenated to form the branch input.
- **Vector** ($\mathbb{R}^{d_v}$). Pre-trained ESM-3²⁵ embeddings using sequence-length hidden states extracted from the last transformer layer and compressed into a single, 1536-d vector via mean-pooling across the amino acid dimension to return a fixed length.

The output vectors combined into a final node representation in two stages, both using the same reusable module B consisting of a layer-normalization operator LN, a learned linear layer

with weights $\mathbf{W}$ and bias $\mathbf{b}$, a nonlinearity $\phi = \text{GELU}$, and an optional dropout layer D_p with probability p during training,

$$\mathcal{B}(\mathbf{x}) = D_p(\phi(\mathbf{W} \text{LN}(\mathbf{x}) + \mathbf{b})) \quad [1]$$

Each branch is first projected by its own instance of the module, $\mathcal{B}_b$, differing in both input and output width based on branch type. The branch outputs are then concatenated and passed through $\mathcal{B}_k$ (per node type as $\mathcal{A}_k$ varies across nodes), mapping the concatenation to a hidden vector of width H :

$$\mathbf{h}_{k,v}^{(0)} = \mathcal{B}_k \left(\left[\mathcal{B}_b(\mathbf{x}_v^{(b)}) \right]_{b \in \mathcal{A}_k} \right) \in \mathbb{R}^H \quad [2]$$

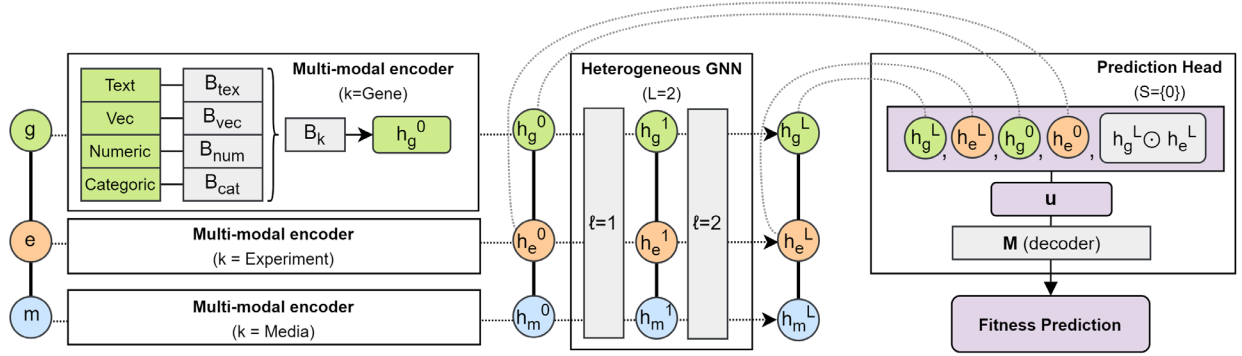


Figure 2. Architecture schematic, detailing the multi-modal encoders, variable depth GNN, and prediction head. Note that within the prediction head, the concatenation (cat) fusion method is displayed, and that the larger KG is represented by a simple three node graph containing a single gene, experiment and media node.

2.4 Heterogeneous Message Passing

Stacking the encoded representations across nodes gives $\mathbf{H}^{(0)}_k \in \mathbb{R}^{N_k \times H}$ per node type k . As shown in Figure 2, we propagate information across the KG using L heterogeneous convolution layers. The FITNESS relation is always excluded from $\mathcal{R}$, as it carries the supervision signal. Our convolutional layers use the multi-head GATv2 attention³⁶ backbone.

2.5 Prediction Head

For an input pair (g, e) , let $\mathbf{h}_g^{(\ell)}, \mathbf{h}_e^{(\ell)}$ denote the gene and experiment representations at depth ℓ , with $\ell = 0$ being the encoder output (Equation (2)) and $\ell = L$ the final message-passing output. Let $S \subseteq \{0, 1, \dots, L-1\}$ be an optional subset of retained intermediate depths, $D = \{L\} \cup S$ be the full set of retained depths, and for $t \in \{g, e\}$ the stack of depth- ℓ be $\mathbf{H}_t$, corresponding with Equation (3).

$$\mathbf{H}_t = \left[\mathbf{h}_t^{(\ell)} \right]_{\ell \in D} \in \mathbb{R}^{|D| \times H} \quad [3]$$

We consider two fusion variants:

- *Concatenation (cat)*. $\mathbf{H}_g$ and $\mathbf{H}_e$ are flattened, giving a head input whose width grows linearly with $|S|$.

$$\mathbf{u} = \left[\text{vec}(\mathbf{H}_g); \text{vec}(\mathbf{H}_e); \mathbf{h}_g^{(L)} \odot \mathbf{h}_e^{(L)} \right] \quad [4]$$

- *Cross-attention (xa)*. Each side's final representation $\mathbf{h}_t^{(L)}$ queries the *other* side's $\mathbf{H}_t$ through an independent multi-head attention block.³⁷ Cross attention is weighted independently across depths. Final representations, cross-attention contexts, and the Hadamard interaction are then concatenated to produce a fixed width vector $5H$.

Finally, the head input $\mathbf{u}$ is passed through a decoder $M: \mathbb{R}^{|u|} \rightarrow \mathbb{R}$, a two layer MLP with input LayerNorm, GELU nonlinearities, and dropout, producing a scalar prediction $\hat{y} = M(\mathbf{u})$ on the quantile-transformed target scale. We note that the MLP-only baseline is recovered as the $L = 0$ special case of the **cat** fusion. It uses the same multi-modal node-encoder design and decoder template as the graph models but does not perform message passing.

2.6 Training and Evaluation

All models use hidden width 256, dropout 0.2, and minibatches of 16,384 gene-experiment fitness pairs. We optimize the mean squared error (MSE) on the quantile-transformed targets using Adam with peak learning rate 10^{-3} , weight decay 10^{-4} , and a 5% linear warm-up. The learning rate is reduced by a factor of 0.3 after eight validation checks without improvement, to a minimum of 10^{-5} . An architecture screen selected GAT/L = 2/xa for the experiments and genes splits and GAT/L = 6/xa for the cluster-aware split. To separate model selection from confirmation, these screening runs were excluded from primary inference; fourteen additional GAT and MLP runs were then trained for each split. Model checkpoints and analysis artifacts are provided in the supplementary data release. For feature ablation assays, we retrained the selected genes-split GAT after disabling one feature branch while leaving the corresponding nodes and graph edges intact. Performance comparisons in Table 2, Figure 3, Figure 4, and Table 3 used the designated validation-selected representative model for each split.

2.7 GSMM performance comparison

The GSMM iJN1463 was downloaded from the BiGG database <http://bigg.ucsd.edu/>. For a carbon condition it closes all carbon-source exchanges and opens only the target carbon source at its uptake flux, leaving ammonium available as the nitrogen source; for a nitrogen condition it closes all nitrogen source exchanges and opens only the target nitrogen source, holding glucose as the carbon source. FBA is computed across all genes and carbon source experiments and we define essential where growth ratio to control ≤ 0.5 or gene knockout growth approaches 0.

3.0 Results

3.1 Model Predictive Performance

For interpretability, relative MSE reduction is reported as $(\text{MSE}_{\text{MLP}} - \text{MSE}_{\text{GAT}}) / \text{MSE}_{\text{MLP}}$. The addition of message passing through the GAT layer significantly improved the observed mean on all three metrics for held-out experiments and genes (Table 1). Relative to the MLP, mean MSE decreased by 5.9% for experiments and 10.6% for genes, corresponding to absolute reductions of 0.0081 ($p=0.000183$, Wilcoxon signed-rank test, Benjamini-Hochberg (BH) correction) and 0.0257 ($p=0.000183$, Wilcoxon signed-rank test, BH correction), respectively. By contrast, the difference in cluster-split MSE was not significant, and GAT had lower MSE in only 7 of 14 pairs. While increasing GAT layer depth to 6 during the more restrictive cluster-aware split enabled deeper exploration of the knowledge graph's integrated metabolic networks, there was no significant improvement over the MLP model. This category also resulted in the lowest performance score across all splits.

We conduct an exploratory analysis to determine which node features support the increase in performance observed under gene-split fitness prediction for graph-observed nodes. Each run was compared with the full GAT using the same seed. Only one node feature ablation, experiment text, resulted in a significant change in performance ($+0.0167 \Delta\text{MSE}$, $p = 0.01860$, BH adjusted).

3.1 Graph Performance Gain in High Connectivity and Essentiality Regime

Graph structure is not homogeneous at the gene level, with different levels of gene annotation, experimental validation, and integration into biological and metabolic knowledge bases. We calculated MAE and percent conditional gene essentiality for genes binned by network connectivity (Figure 3). Within low network connectivity regimes (0-14 median gene degree) the GNN closely tracks MLP performance, and the percent of essential gene pairs remains low. Above this connectivity range GNN gain over MLP and percent essentiality increases to 2.96% and 6.18%, respectively. Within the essential gene stratum (fitness defect -2 and lower), the experiment and gene GNN beats the MLP all metrics save precision, where it scores lower (Table 2). Graph inference proves most helpful for the gene split, where recall increases $+9.2$ pts and $F_1 +9.0$ pts, though precision falls 4.1 pts, resulting in 1,633 more true-essential gene-experiment observations recovered. The MLP beats the GNN architecture in every metric under the gene-cluster split.

Table 1. Mean and standard deviation of test performance across confirmation runs for the selected GAT and matched MLP baseline. The architecture-screen seed is excluded.

Split	Model	n	MSE $\downarrow$	MAE $\downarrow$	$\rho\uparrow$
Experiments	MLP	14	0.1364 ± 0.0012	0.2127 ± 0.0005	0.7457 ± 0.0021
	GAT/ $L = 2/x_a$	14	0.1284 ± 0.0028	0.2097 ± 0.0012	0.7659 ± 0.0029
Genes	MLP	14	0.2416 ± 0.0048	0.2553 ± 0.0011	0.5299 ± 0.0119
	GAT/ $L = 2/x_a$	14	0.2160 ± 0.0079	0.2482 ± 0.0018	0.5958 ± 0.0187
Clusters	MLP	14	0.2934 ± 0.0052	0.2712 ± 0.0009	0.4294 ± 0.0196
	GAT/ $L = 6/x_a$	14	0.2938 ± 0.0059	0.2705 ± 0.0010	0.4228 ± 0.0187

Table 2. Comparison of best MLP and GAT performance across data splits in the high essentiality stratum. "Ess." refers to conditionally essential fitness pairs per split.

Split	Model	Truly ess.	Pred. ess.	TP	Recall	Precision	F_1
Experiments	GAT	10,468	7,364	6,197	0.592	0.842	0.695
	MLP	10,468	6,311	5,357	0.512	0.849	0.639
Genes	GAT	17,742	7,897	6,630	0.374	0.840	0.517
	MLP	17,742	5,672	4,997	0.282	0.881	0.427
Clusters	GAT	13,748	1,759	1,567	0.114	0.891	0.202
	MLP	13,748	2,228	1,995	0.145	0.895	0.250

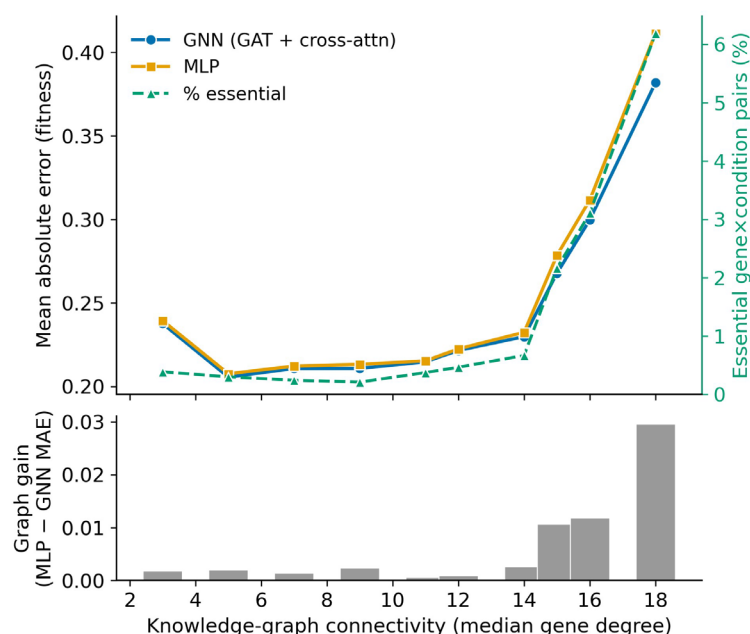


Figure 3. Top: Mean absolute error rises as gene network connectivity increases during cold-start gene inference. Dots indicate MAE point estimates. Bottom: Bar chart indicating the graph gain at each connectivity bin.

3.2 Regression advantage disappears under the cluster-aware split

Each test prediction was assigned to its gene's primary level-1 Enzyme Commission (EC) class (1-7). For genes with multiple EC annotations, the primary class was the modal first EC digit, with the smaller class used to break ties. Graph gain versus MLP increases in most categories from the experiment to gene splits as the fraction of conditionally essential fitness gene-experiment observation also rises or stays stable. The cluster split reverses this trend, decreasing graph gain across functional groups primarily driven by category 4 (lyases, $\Delta\text{MAE} -0.1436$) and increasing the essential fraction of fitness gene-experiment observation to 13.16%. This specific category is the main driver of decreased GNN performance over MLP in the cluster split.

3.3 GFM1 outperforms a GSMM when predicting fitness growth defects

GFM1 substantially outperforms iJN1463 derived fitness predictions across both classification and continuous prediction metrics (Table 3). The GNN's AUPRC is approximately 2.7-fold higher than iJN1463 and 15.12x above the 0.052 random baseline. This results in 96 fewer false positives: 19 vs. 115, a 83% reduction. GFM1 also showed a stronger correlation with continuous RB-TnSeq fitness measurements ($r=0.790$) than iJN1463 ($r=0.435$).

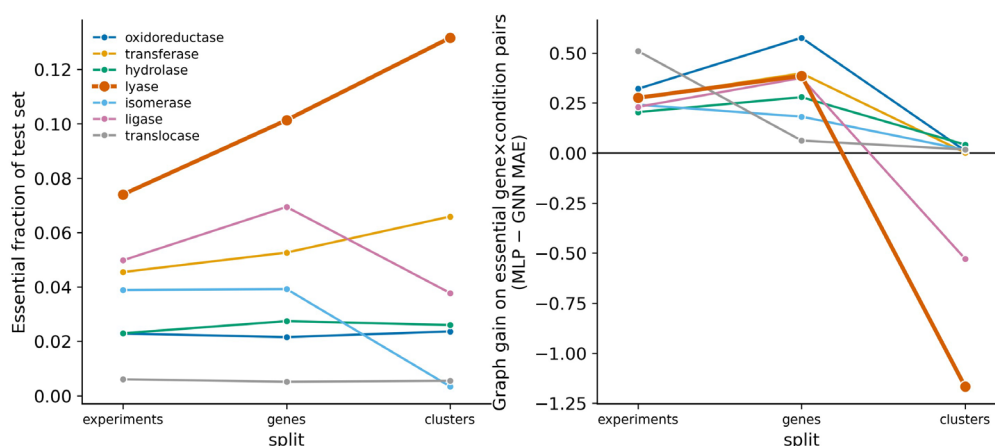


Figure 4. Cluster Split Prediction performance across enzyme classification groups is heavily influenced by fraction of essential conditions. Left: Conditionally essential fitness cell fractions for all EC groups across each split protocol. Right: Graph gain for all EC groups across split protocols. Only 27-30% of test prediction rows could be assigned a canonical level-1 EC class.

Table 3. Comparison of flux balance analysis (FBA) and GFM1 predictions against experimental RB-TnSeq measurements for cold-start genes in *Pseudomonas putida* KT2440. Base rate of conditionally essential genes is 0.052.

Model	AUROC	AUPRC	F_1	Precision	Recall	Accuracy	Pearson
FBA (iJN1463)	0.733	0.294	0.455	0.392	0.544	0.932	0.435
GFM1 + xa	0.940	0.786	0.765	0.835	0.706	0.977	0.790

4.0 Discussion

4.1 Fitness prediction under cluster-aware generalization remains challenging

Our results show that microbial gene-knockout effects can be predicted in two settings: generalization from known genes to held-out experiments and cold-start prediction for held-out genes. Performance declined under the gene-cluster split, indicating that prediction is substantially more difficult under the more restrictive cluster-aware split. In the experiment and gene splits, pretrained embeddings and biological context allow the model to transfer information across related conditions and genes. Much of this information is unavailable or restricted in the cluster split, and message passing does not significantly improve performance over the non-graph baseline.

Class imbalance may contribute to this result: most gene-experiment pairs have little or no measured fitness effect, so a model can minimize average error by predicting values near the center of the distribution. Even so, these results demonstrate that large-scale RB-TnSeq measurements can serve as training data for predicting phenotypes in unmeasured gene-condition combinations.

4.2 Relational structure improves prediction of conditionally essential phenotypes

Message passing improves prediction over matched non-graph baselines, with the largest gains among strongly deleterious fitness observations (Table 2). Within the strongly deleterious fitness stratum the GAT had higher precision, recall, or F_1 in seven of the nine split-by-metric comparisons. The improvement was most pronounced in the gene split, where graph traversal can incorporate information from orthologs and gene-family context.

Under the tested architectures and optimization settings, two-layer GATs performed best in the experiment and gene splits, so it is possible that local graph context is more useful than additional rounds of message passing. However, encoder depth alone does not establish that deeper resources, such as the EC hierarchy or MetaCyc compound network, are uninformative: their contributions may also depend on graph connectivity, feature quality, sampling, and optimization. The observation that only one individual feature ablation significantly changed performance could indicate redundancy among node features, compensation through graph topology, or insufficient sensitivity of single-feature ablations. More targeted structural, edge-type, and subgraph ablations are needed to determine which knowledge sources support generalization in the cluster setting.

4.3 GFM1 complements mechanistic genome-scale metabolic models

GFM1 showed stronger agreement with experimental RB-TnSeq measurements than flux balance analysis with iJN1463 for the evaluated *Pseudomonas putida* KT2440 benchmark. In particular, GFM1 achieved higher classification performance and stronger correlation with continuous fitness measurements while producing fewer false-positive growth-defect predictions (table 3). These results support the value of learned models for predicting fitness across genes and conditions represented incompletely by a genome-scale metabolic model.

The approaches nevertheless provide different forms of information. Genome-scale metabolic models impose stoichiometric and mass-balance constraints and can generate reaction- and flux-level hypotheses, whereas GFM1 produces empirical fitness predictions without an equivalent mechanistic explanation. Their strengths are therefore complementary rather than directly interchangeable. GFM1 could prioritize gene-condition pairs with predicted fitness defects, after which a genome-scale metabolic model could be used to investigate candidate metabolic mechanisms. Conversely, disagreement between the models and experimental measurements could help identify missing or incorrect gene-reaction associations.

4.4 Limitations and future work

Two limitations should be considered. First, the training data are taxonomically restricted to ten *Pseudomonas* strains. An organism-level split would more closely represent prediction for a previously unmeasured species or strain and would test whether gene and experiment representations transfer across phylogenetic lineages. Expanding training to the broader Fitness Browser collection could improve the learning of conserved sequence-to-function relationships, although differences in experimental design, annotation quality, and phylogenetic representation would require careful control.

Finally, alternative architectures may also improve information flow through the heterogeneous graph. For example, knowledge-source-specific encoders or mixture-of-experts modules could process individual subnetworks before combining their representations through a learned gate. Protein structures, protein-protein interactions, transcriptomics, and proteomics could provide additional and potentially complementary signals. Future work should evaluate these modalities alongside alternative heterogeneous-graph encoders and more stringent phylogenetic generalization protocols.

5.0 Conclusion

We developed GFM1, a heterogeneous graph neural network for predicting condition-specific microbial fitness following gene inactivation. Across held-out experiment, gene, and genecluster settings, GFM1 performed best when orthologous or gene-family context was available. Message passing improved performance over a matched shared-encoder MLP in the experiment and gene splits, with the largest gene-split gains among well-connected genes enriched for conditionally essential phenotypes. These gains were substantially reduced in the gene cluster split, demonstrating that generalization to genes from withheld families remains challenging.

GFM1 also showed stronger agreement with experimental RB-TnSeq measurements than flux balance analysis with iJN1463 on the evaluated *Pseudomonas putida* KT2440 benchmark. These results demonstrate that protein and text embeddings can be integrated with biological knowledge graphs to learn condition-dependent genotype-phenotype relationships. Future work should expand the taxonomic and experimental diversity of the training data, improve transfer to novel gene families and explore architecture changes, and evaluate the resulting representations in experimental-design and functional-annotation workflows. Together, these results highlight GFM1's potential to prioritize informative gene-condition combinations for experimental validation and accelerate microbial functional characterization.

6.0 References

1. Wagner, Genotype sampling for deep-learning assisted experimental mapping of a combinatorially complete fitness landscape, *Bioinformatics* **40**, p. btae317 (2024).
2. S. Valavarasu, Y. Sangu and T. Mahapatra, Prediction of antibiotic resistance from antibiotic susceptibility testing results from surveillance data using machine learning, *Scientific Reports* **15**, p. 30509 (2025).
3. K. M. Wetmore, M. N. Price, R. J. Waters, J. S. Lamson, J. He, C. A. Hoover, M. J. Blow, J. Bristow, G. Butland, A. P. Arkin *et al.*, Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons, *MBio* **6**, 10 (2015).
4. J. Borchert, A. Bleem and G. T. Beckham, Experimental and analytical approaches for improving the resolution of randomly barcoded transposon insertion sequencing (rb-tnseq) studies, *ACS Synthetic Biology* **11**, 2015 (2022), PMID: 35657709.
5. M. N. Price, K. M. Wetmore, R. J. Waters, M. Callaghan, J. Ray, H. Liu, J. V. Kuehl, R. A. Melnyk, J. S. Lamson, Y. Suh *et al.*, Mutant phenotypes for thousands of bacterial genes of unknown function, *Nature* **557**, 503 (2018).
6. M. Rajkumar, L. B. Bruno and J. R. Banu, Alleviation of environmental stress in plants: The role of beneficial pseudomonas spp., *Critical Reviews in Environmental Science and Technology* **47**, 372 (2017).
7. M. G. Wilson and S. Pandey, *Pseudomonas aeruginosa*, in *StatPearls [Internet]*, (StatPearls Publishing, 2023)
8. H. Asif, D. J. Studholme, A. Khan, M. Aurongzeb, I. A. Khan and M. K. Azim, Comparative genomics of an endophytic pseudomonas putida isolated from mango orchard, *Genetics and Molecular Biology* **39**, 465 (2016).
9. T. Joshi and D. Xu, Quantitative assessment of relationship between sequence similarity and function similarity, *BMC genomics* **8**, p. 222 (2007).
10. Y. Ma, H. Chen, J. Kang, X. Guo, C. Sun, J. Xu, J. Tao, S. Wei, Y. Dong, H. Tian *et al.*, The hidden markov model and its applications in bioinformatics analysis, *Genes & Diseases* **13** (2026).
11. G. A. Pavlopoulos, F. A. Baltoumas, S. Liu, O. Selvitopi, A. P. Camargo, S. Nayfach, A. Azad, S. Roux, L. Call, N. N. Ivanova *et al.*, Unraveling the functional dark matter through global metagenomics, *Nature* **622**, 594 (2023).
12. J. Nogales, J. Mueller, S. Gudmundsson, F. J. Canalejo, E. Duque, J. Monk, A. M. Feist, J. L. Ramos, W. Niu and B. O. Palsson, High-quality genome-scale metabolic modelling of pseudomonas putida highlights its broad metabolic capabilities, *Environmental microbiology* **22**, 255 (2020).
13. Antonakoudis, R. Barbosa, P. Kotidis and C. Kontoravdi, The era of big data: Genome-scale modelling meets machine learning, *Computational and structural biotechnology journal* **18**, 3287 (2020).

14. M. Abdel-Haleem, H. Hefzi, K. Mineta, X. Gao, T. Gojobori, B. O. Palsson, N. E. Lewis and N. Jamshidi, Functional interrogation of plasmodium genus metabolism identifies species- and stage-specific differences in nutrient essentiality and drug targeting, *PLoS computational biology* **14**, p. e1005895 (2018).
15. Z. Hartevelde, A. Van Hall-Beauvais, I. Morozova, J. Southern, C. Goverde, S. Georgeon, S. Rosset, M. Defferrard, A. Loukas, P. Vandergheynst *et al.*, Exploring “dark matter” protein folds using deep learning, *bioRxiv* , 2023 (2023).
16. J. Zhong, L. Li, R. Dannenfelser and V. Yao, Benchmarking gene embeddings from sequence, expression, network, and text models for functional prediction tasks, *bioRxiv* , 2025 (2025).
17. S. Ji, S. Pan, E. Cambria, P. Marttinen and P. S. Yu, A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE Transactions on Neural Networks and Learning Systems* **33**, 494 (2022).
18. Peng, F. Xia, M. Naseriparsa and F. Osborne, Knowledge graphs: Opportunities and challenges, *Artificial intelligence review* **56**, 13071 (2023).
19. M. Schlichtkrull, T. N. Kipf, P. Bloem, R. v. d. Berg, I. Titov and M. Welling, Modeling relational data with graph convolutional networks (2017).
20. X. Wang, H. Ji, C. Shi, B. Wang, P. Cui, P. Yu and Y. Ye, Heterogeneous graph attention network (2021).
21. Z. Ye, Y. J. Kumar, G. O. Sing, F. Song and J. Wang, A comprehensive survey of graph neural networks for knowledge graphs, *IEEE Access* **10**, 75729 (2022).
22. M. Kanehisa, The kegg database, in *‘In silico’ simulation of biological processes: Novartis Foundation Symposium* 247, 2002.
23. R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti and P. D. Karp, The metacyc database of metabolic pathways and enzymes-a 2019 update, *Nucleic acids research* **48**, D445 (2020).
24. Bairoch and R. Apweiler, The swiss-prot protein sequence database and its supplement trembl in 2000, *Nucleic acids research* **28**, 45 (2000).
25. T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert *et al.*, Simulating 500 million years of evolution with a language model, *Science* **387**, 850 (2025).
26. J. Ma, J. Song, N. D. Young, B. C. Chang, P. K. Korhonen, T. L. Campos, H. Liu and R. B. Gasser, ‘bingo’—a large language model-and graph neural network-based workflow for the prediction of essential genes from protein data, *Briefings in Bioinformatics* **25**, p. bbad472 (2024).
27. J. Borchert, A. C. Bleem, H. G. Lim, K. Rychel, K. D. Dooley, Z. A. Kellermeyer, T. L. Hodges, B. O. Palsson and G. T. Beckham, Machine learning analysis of rb-tnseq fitness

- data predicts functional gene modules in pseudomonas putida kt2440, *Msystems* **9**, e00942 (2024).
28. M. Li, X. Cai, S. Xu and H. Ji, Metapath-aggregated heterogeneous graph neural network for drug-target interaction prediction, *Briefings in Bioinformatics* **24**, p. bbac578 (January 2023).
 29. Z. Li, X. Tu, Y. Chen and W. Lin, HetDDI: A pre-trained heterogeneous graph neural network model for drug-drug interaction prediction, *Briefings in Bioinformatics* **24**, p. bbad385 (September 2023).
 30. Q. Ye, Y. Zeng, L. Jiang, Y. Kang, P. Pan, J. Chen, Y. Deng, H. Zhao, S. He, T. Hou and C.-Y. Hsieh, A knowledge-guided graph learning approach bridging phenotype- and target-based drug discovery, *Advanced Science* **12**, p. e2412402 (April 2025).
 31. H. Wen, S. Carpenter, K. McGinnis, A. Nelson, K. Smith and T. Hong, Esssubgraph improves performance and generalizability of mammalian essential gene prediction with large networks, *GigaScience* **14**, p. giaf136 (01 2025).
 32. Y. Xiao *et al.*, FuseLinker: leveraging LLM-generated text and domain knowledge for enhanced link prediction on biomedical knowledge graphs, *Journal of Biomedical Informatics* **158**, p. 104735 (2024).
 33. G. McDonald and K. F. Tipton, Enzyme nomenclature and classification: the state of the art, *The FEBS Journal* **290**, 2214 (2023).
 34. P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur *et al.*, A large-scale evaluation of computational protein function prediction, *Nature methods* **10**, 221 (2013).
 35. Google, Gemini api: Embeddings <https://ai.google.dev/gemini-api/docs/embeddings>, Accessed: 2026-01-19.
 36. S. Brody, U. Alon and E. Yahav, How attentive are graph attention networks? (2022).
 37. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. . Kaiser and I. Polosukhin, Attention is all you need, in *Advances in neural information processing systems*, 2017.

Pacific Northwest National Laboratory

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

1-888-375-PNNL (7665)

www.pnnl.gov