PLANTMETBENCH: A MULTIMODAL KNOWLEDGE GRAPH BENCHMARK FOR BIOSYNTHESIS PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Plant metabolism produces a vast diversity of molecules, yet identifying the enzymes that catalyse biosynthetic reactions remains a major open problem. Existing benchmarks typically reduce this problem to isolated enzyme–reaction pairs, overlooking pathway and metabolite context that determines biological plausibility, leaving no standardised way to evaluate approaches that exploit this richer multimodal and relational structure. We present **PlantMetBench**, a multimodal graph benchmark built on PlantMetWiki, a curated knowledge graph of plant metabolism spanning 424 organisms. The resulting benchmark graph comprises 13,682 proteins, 2,741 genes, 4,577 metabolites, and 11,980 experimentally curated enzyme–reaction pairs. PlantMetBench frames enzyme–reaction prediction as an organism-disjoint link-prediction task and provides graph-native dataset splits, precomputed multimodal node features, and evaluation protocols. Using PlantMetBench, we evaluate heterogeneous graph neural networks, standard architectures (HeteroSAGE, GAT, HGT, and R-GCN), alongside residual and jumping-knowledge variants, against non-graph models. Most heterogeneous graph architectures consistently outperform non-graph baselines on our rank-based evaluation, suggesting the value of relational context for enzyme–reaction prediction. Yet, performance remains far from saturated, indicating substantial room for improved representations and modelling strategies. Also, we show that benchmark design choices (shortcuts, splits, and metrics) are key to aligning evaluation with biologically meaningful tasks. PlantMetBench provides a standardised, reproducible framework to evaluate multimodal biosynthetic data, establishing a foundation for future work on graph-based architectures for enzyme–reaction prediction. Code and data are publicly available.

1 INTRODUCTION

Plant specialised metabolism produces an extraordinary diversity of molecules that contribute to plant adaptation and defence while providing a major source of natural products with applications as pharmaceuticals, food ingredients, crop protection compounds, and other bio-based products (Owen et al., 2026). Yet most biosynthetic pathways responsible for plant specialised metabolite biosynthesis remain unresolved. Reconstructing these pathways could enable the biotechnological production of known high-value compounds and reveal yet-unknown parts of chemical space. However, this remains challenging since plant metabolism involves highly duplicated and rapidly evolving gene families, enzymes often act on multiple substrates, and pathway components are frequently dispersed across the genome (Ji et al., 2024). The scale of the resulting chemical and biological search space, together with the costly and incomplete experimental characterisation, makes computational and machine learning (ML) approaches essential for prioritising candidate enzymes, reactions, and biosynthetic routes for experimental investigation (Capela et al., 2025).

Predicting biosynthesis, particularly in plants, is intrinsically a *multimodal* and relational inference problem. Genomic and protein information describe potential enzymatic functions, while metabolite structures constrain possible biochemical transformations. These are shared across taxa through convergent or divergent evolution (Scossa et al., 2026). This structure also poses a challenge for ML evaluation, since biosynthetic knowledge is spread across heterogeneous, incomplete resources, and existing studies adopt different tasks, datasets, negative-sampling procedures, and train/test splits when addressing individual modalities. Similar concerns have motivated more rigorous benchmarks

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.

in graph learning, biomedical link prediction, and molecular ML (Breit et al., 2020; Wu et al., 2018; Pahikkala et al., 2015; Bechler-Speicher et al., 2025). Yet, existing work largely prioritises narrow tasks over real-world graph applications (Bechler-Speicher et al., 2025) and addresses single-modality node- or graph-level tasks, lacking an integrated multimodal graph approach (Cornelis et al., 2026).

Here, we introduce **PlantMetBench**, a multimodal knowledge graph benchmark for plant biosynthesis prediction. It builds on PlantMetWiki (Del Pup et al., 2026), a FAIR knowledge graph spanning hundreds of plant species and integrating genes, proteins, enzymes, metabolites, reactions, pathways, chemical structures, and taxonomy. The graph reflects the biological structure of metabolic networks manually curated in PlantCyc (Hawkins et al., 2025) from experimental evidence in the literature. Genes encode enzymes, enzymes catalyse reactions, reactions transform metabolite substrates into products, and pathways link reactions through shared metabolites into curated metabolic processes. These metabolic processes occur in biological taxa (such as plant species or genera), in which the individual components (i.e., enzymes and genes) are characterised. This makes plant biosynthesis a natural testbed for relational and multimodal ML. Building on this representation, PlantMetBench defines prediction tasks, negative-sampling procedures, evaluation metrics, and reproducible data splits. PlantMetBench code and dataset are publicly available.[1]

PlantMetBench is a novel framework to evaluate representative knowledge graphs, multimodal architectures, and graph neural networks (GNNs), including relational GNNs. It extends existing research by enabling systematic evaluation of whether current models can learn effectively from the benchmark and how evaluation choices influence conclusions. In detail, in the following, we present the full list of contributions of the paper,

- *Full benchmark ecosystem construction*: We propose PlantMetBench, a framework that (i) constructs a graph-based dataset (derived from PlantMetWiki), which is ready to be used for ML training, (ii) provides the whole training workflow as a toolkit (e.g., data splitting policies, evaluation metrics, etc., designed for this application-specific graph-based dataset).
- *Evaluation on standard GNNs:* We evaluate PlantMetBench using standard GNN models and two simple architectural alternatives. The results demonstrate the value of the knowledge encoded in the dataset, showing that simple adjustments can improve standard models. However, performance has not yet saturated, highlighting PlantMetBench as a challenging benchmark for future research.
- *Baseline comparison with non-graph models:* We compare the performance of GNN models with non-graph ones. Our findings highlight the importance of the relational context graphs provide. Graph-based models not only predict the target enzyme–reaction links, but also associate them with other modalities, which may enable pathway prioritisation.
- *Wider Evaluation potential:* Beyond the default setting, PlantMetBench includes further evaluation dimensions beyond the enzyme–reaction task, such as alternative splits (unseen pathways and unseen reactions), and a pathway reconstruction metric, opening new directions for future research.

2 Related Work

Graph & molecular ML benchmarking. Standardised benchmarks enable reproducible ML comparisons through common datasets, tasks, and evaluation procedures for graph learning (Morris et al., 2020; Hu et al., 2020a; Stoll et al., 2025) and molecular ML (Wu et al., 2018). Recent work shows that benchmark design can strongly influence results, especially under poor graph–task alignment or dataset leakage (Bechler-Speicher et al., 2025). OGBench (Cornelis et al., 2026) shows this in high-dimensional omics data. Breit et al. (2020) added splits, leakage control, and cross-model evaluation for link prediction, yet such protocols mainly address topological leakage, whereas in biosynthesis, information can also leak through entity attributes, as structurally similar metabolites or related enzyme families may appear across train and test sets without any direct graph connection.

Enzyme and reaction ML. A range of ML methods target individual components of biosynthetic prediction, such as enzyme function, substrate specificity, reaction prediction, and pathway reconstruction, from predicting Enzyme Commission (EC) classes via 3D GNNs (van der Weg et al., 2025) to identifying plant metabolite precursors through automated ML (Capela et al., 2025). Various benchmarks show that task and split design determine whether generalisation is genuinely tested (Hua

[1] https://anonymous.4open.science/r/plantmetbench-0391/

et al., 2024; Wu et al., 2018; Bushuiev et al., 2024). These approaches target complementary components with differing tasks and representations, but miss the heterogeneous relationships connecting the underlying biological and chemical entities.

Dual-Encoder retrieval for enzyme-reaction matching. Prior work models enzyme-reaction relations directly, embedding enzymes and reactions with separate encoders and aligning them in a single shared space using contrastive learning. For instance, CLIPZyme (Mikhael et al., 2024) uses a CLIP-style approach for enzyme screening, while ReactZyme (Hua et al., 2024) standardised it into a large-scale retrieval benchmark with specific data splits (time, enzyme similarity, and reaction similarity). Further, TIGER (Zhang et al., 2026) improves the enzyme representation with generated textual descriptions. Finally, Horizyn-1 (Rocks et al., 2026) scales the architecture to millions of reactions, splitting on reaction fingerprints and validating the model experimentally. These models are restricted to a single relation between enzymes and reactions, and hence they do not explore the context in which catalysis occurs, including pathway membership, gene annotation, and organism of origin. PlantMetBench makes this context explicit as typed relations in a heterogeneous graph.

3 PLANTMETBENCH: AN EVALUATION FRAMEWORK

PlantMetBench is a reproducible benchmark for enzyme-reaction prediction in unseen species. We describe its construction, node features, prediction task, data splits, and evaluation metrics. Figure 1 shows the graph layout, and Appendix A shows the system overview.

Graph construction. The benchmark uses the PlantMetWiki (Del Pup et al., 2026) knowledge graph, processed as a PyTorch Geometric `HeteroData` object. The source graph has seven node types: PROTEIN (13,682), INTERACTION (84,470; of which 19,927 are CONVERSION-subtype, the enzyme-catalysed reactions used throughout this work), METABOLITE (17,476), GENEPRODUCT (6,480), ORGANISM (424), and PATHWAY (2,478). We reduce this to the training graph by removing entities outside the prediction task or that introduce label shortcuts. These are the non-Conversion Interactions (not prediction targets), the pathway-scoped alias duplicates of METABOLITE and GENEPRODUCT nodes (redundant copies of existing entities that lack input features and add no distinct biochemical information). ORGANISM nodes are removed as well, though organism membership is still used to define the taxa holdout split described below.

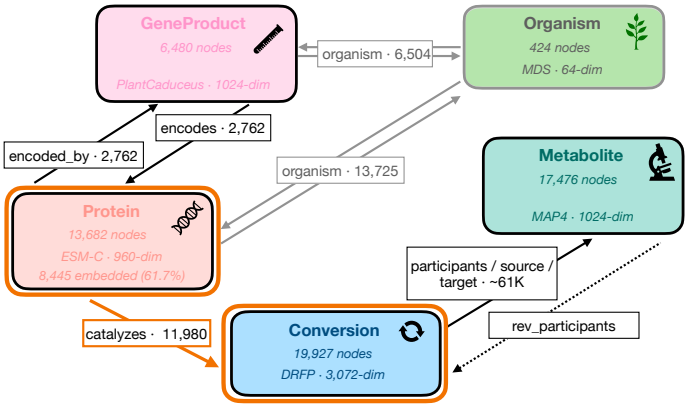

Figure 1: Graph visualisation. Orange nodes/edges denote the target, dotted line is inserted reverse edge, grey nodes are excluded.

We further remove structural shortcuts that would otherwise let the model recover catalysis labels without learning the underlying enzyme-reaction association. First, *pathway co-membership*: all (X, `is_part_of`, PATHWAY) edges and the PATHWAY nodes are dropped. A pathway is a curated set of reactions from one metabolic process, so a protein and a reaction sharing a pathway are linked regardless of whether the protein catalyses it, giving a topological shortcut to the label. Second, *reverse catalysis*: the edge (INTERACTION, `catalyzed_by`, PROTEIN) is excluded to avoid a 2-hop path that trivially reconstructs the supervision edge. Third, *one-hop supervision*: we hold out 20% of training positives as supervision-only pairs, so the model cannot recover them directly from a node's immediate neighbourhood (Appendix D.4). The full shortcut ablation is in Appendix F.

Finally, we provide a bidirectional mode through which information from secondary modalities can flow to the primary ones, e.g., adding edge (METABOLITE, `rev_participants`, CONVERSION). The final graph contains four node types: PROTEIN (13,682), CONVERSION (19,927), METABOLITE (4,577), and GENEPRODUCT (2,741). The prediction edge (PROTEIN, `catalyzes`,

CONVERSION) is derived from the Resource Description Framework (RDF) triples, giving $11{,}980$ curated, experimentally supported enzyme–reaction pairs, which we treat as ground truth. Node/edge statistics, the graph-cleaning pipeline, and further construction details are provided in Appendix B.

Node Features. Each node type carries precomputed domain-specific features. Coverage is limited by how many nodes can be resolved to a sequence or chemical structure in public databases, and unresolved nodes are assigned zero vectors. PROTEIN nodes use embeddings from ESM Cambrian (ESM-C) (Candido et al., 2026), with $8{,}445$ of $13{,}682$ proteins having a non-zero embedding (61.7% coverage). GENEPRODUCT nodes use embeddings from PlantCaduceus (Zhai et al., 2025) (49.4% coverage), a DNA language model pretrained on 16 angiosperm genomes, which adds nucleotide-level information from each gene's own sequence that protein ESM-C does not capture. METABOLITE nodes use MinHashed atom-pair (MAP4) circular fingerprints computed from SMILES (Capecchi et al., 2020) (91.7% coverage). CONVERSION (biochemical transformations of metabolite substrates into products) nodes use fingerprints approximating the differential reaction fingerprint (DRFP) (Probst et al., 2022) via MAP4, structured as $[\text{MAP4}_{\text{sub}}||\text{MAP4}_{\text{prod}}||\text{MAP4}_{\text{sub}} - \text{MAP4}_{\text{prod}}]$ (68.8% coverage). Appendix C provides full details and coverage statistics. ORGANISM nodes receive 64-dim classical Multi-dimensional Scaling (MDS, or PCoA) coordinates from pairwise NCBI-lineage taxonomic-hop distances, capturing phylogenetic relatedness between species.

Task. Reconstructing biosynthesis pathways requires matching enzymes to reactions. We formulate this as a ranking task: given a Conversion (an enzyme-catalysed reaction), the model ranks all $8{,}445$ embedded proteins by their likelihood of catalysing it. Catalysis labels are held out by organism, so the model must predict associations for species whose catalysis is absent from training.

Splits. Our default evaluation uses a *taxa holdout*, which assigns entire species to validation (11 organisms) or test (69 organisms) if at least 10% of its proteins have a non-zero sequence embedding. This prevents zero-embedding proteins from dominating evaluation. In this split, train and test share no proteins (Tanimoto similarity = 0.00),[2] but reactions are less strictly disjoint, since the same curated reaction can be catalysed by enzymes from multiple species (reaction Tanimoto = 0.10 train–test), so it can appear as a target in both partitions (Appendix D.7). The graph is also nearly partitioned by organism, with protein and gene edges 99.9% and 95.2% within-organism. In our main configuration, we remove organism nodes, so a held-out protein cannot aggregate information from training-organism proteins through message passing. Most paths connecting held-out species to training data run through shared metabolites and reactions, which reflect conserved biochemistry. The taxa holdout therefore measures cross-species transfer through shared biochemistry, though part of this signal uses the held-out organism's own within-species graph context. We quantify transductive signal in the results and Appendix D.8.

Three other splits probe different levels of generalisation difficulty. The *random-edge* split (a sanity-check upper bound) partitions edges uniformly $(80/10/10)$, leaving all proteins and reactions visible across splits. The *pathway holdout* withholds entire pathways for validation and test (Conversion Tanimoto to train = 0.00), so reactions are unseen but proteins can still appear in training via other pathways (see Appendix D.7). The *A. thaliana reaction holdout* restricts the graph to *Arabidopsis thaliana*, the training species with near-complete (97.6%) ESM-C coverage, and holds out 20% of its reactions, stratified by top-level EC class (ensuring no EC class appears only in test), ranking against $5{,}289$ proteins. As every candidate protein is present in training, this split tests generalisation to unseen reactions without the added difficulty of transferring to new species (Appendix D.6).

Metrics. P-H@k (Protein Hits at k): for each evaluation reaction, all $8{,}445$ embedded proteins are ranked by predicted score. P-H@k is the fraction of reactions for which the true catalyst appears in the top-k positions. The random baseline is $k/8445 \approx 0.59\%$ for k=50. This metric requires no sampled negatives and directly measures the practically relevant ranking quality, following the same full-pool, rank-based retrieval protocol used for spectrum-to-structure identification in MassSpecGym (Bushuiev et al., 2024). **AnyP-H@k** corrects for isoenzyme multiplicity by counting a hit if *any* known catalyst of a reaction is ranked in the top-k. **CP-AUC** and **CP-AP** (Corrupt-Protein AUC/Average Precision): for each positive protein-conversion (P, C) edge, one negative (P_{rand}, C) is sampled at evaluation time. AUC and AP are computed over this balanced set. Beyond per-reaction ranking, we assess **pathway reconstruction**, as the fraction of a pathway's reactions for which

[2]Here Tanimoto denotes the Jaccard overlap between the sets of node identities (protein or reaction IDs) present in two splits, not chemical or structural similarity.

Table 1: Organism disjoint link-prediction performance on PlantMetBench (taxa holdout, organism nodes removed, mean ± std over 5 seeds. Random baseline and BLASTp are deterministic. P-H@K: fraction of reactions where the true catalyst ranks in the top K of 8,445 proteins. Best-performing depth shown per architecture. Full depth sweep in Appendix H.9.

Category	Model	Split	P-H@10 (%)	P-H@50 (%)	CP-AUC	CP-AP
Random baseline	–	–	0.12	0.59	0.500	0.500
Structure-only	HeteroSAGE+Res&Jump (L2)	Val	0.3 ± 0.2	2.4 ± 0.3	0.70 ± 0.02	0.66 ± 0.02
		Test	0.4 ± 0.2	2.1 ± 0.3	0.70 ± 0.03	0.67 ± 0.02
Convolutional	R-GCN (L1)	Val	11.5 ± 1.6	37.7 ± 2.1	0.931 ± 0.005	0.923 ± 0.006
		Test	15.5 ± 2.4	39.6 ± 4.5	0.932 ± 0.002	0.925 ± 0.002
	HeteroSAGE (L1)	Val	11.0 ± 2.4	40.5 ± 1.3	$\mathbf{0.952 \pm 0.004}$	0.946 ± 0.006
		Test	17.5 ± 2.2	42.0 ± 5.0	$\mathbf{0.948 \pm 0.001}$	0.943 ± 0.002
	HeteroSAGE+Res (L1)	Val	9.4 ± 3.0	34.6 ± 3.3	0.930 ± 0.003	0.918 ± 0.005
		Test	10.7 ± 2.7	34.0 ± 5.1	0.929 ± 0.002	0.918 ± 0.003
	HeteroSAGE+Res&Jump (L2)	Val	$\mathbf{21.6 \pm 3.7}$	$\mathbf{49.9 \pm 3.4}$	0.943 ± 0.009	0.944 ± 0.006
		Test	$\mathbf{20.7 \pm 2.4}$	$\mathbf{48.0 \pm 2.9}$	0.937 ± 0.006	0.940 ± 0.005
Attention	HGT (L1)	Val	12.3 ± 1.7	36.3 ± 3.2	0.905 ± 0.009	0.896 ± 0.011
		Test	12.2 ± 2.4	35.5 ± 5.5	0.904 ± 0.006	0.894 ± 0.006
	GAT (L2)	Val	3.5 ± 1.2	21.9 ± 3.2	0.902 ± 0.009	0.875 ± 0.015
		Test	1.6 ± 1.0	11.6 ± 3.6	0.901 ± 0.008	0.871 ± 0.013
	MIX (L1)	Val	9.9 ± 2.4	31.4 ± 3.7	0.940 ± 0.003	0.928 ± 0.004
		Test	10.4 ± 3.7	32.5 ± 6.8	0.942 ± 0.005	0.932 ± 0.006
Non-graph	BLASTp	Test	10.17	32.44	–	–
	Dual encoder (L1)	Val	4.6 ± 1.3	27.3 ± 3.4	0.915 ± 0.003	$\mathbf{0.957 \pm 0.002}$
		Test	5.3 ± 0.7	29.6 ± 1.8	0.932 ± 0.010	$\mathbf{0.966 \pm 0.005}$

the true catalyst is ranked in the top-k, testing whether isolated hits compose into fully-recovered biosynthetic pathways. Full metric definitions are in Appendix G.

Negatives (training). Per positive (P, C), we sample k=5 conversion negatives (P, C_{rand}) and k=5 protein negatives (P_{rand}, C). The latter teach the model which protein catalyses a reaction, the task P-H@k evaluates. Currency metabolites (i.e., cofactors) involved in more than 100 reactions are optionally filtered from the message-passing graph, as their shared embeddings can shortcut reaction-level discrimination. An ablation over the number of negatives is provided (Appendix H.7).

4 Numerical Evaluation

Here, we evaluate PlantMetBench using various GNNs and compare their performance with non-graph baselines. We further conduct ablation studies on multiple aspects of the benchmark.

4.1 Setup

GNN models. Given the multimodal graph structure of the data, we evaluate using two types of architectures: (i) four standard heterogeneous GNNs: *HeteroSAGE* (Hamilton et al., 2017), *GAT* (Veličković et al., 2018), *HGT* (Hu et al., 2020b), and *R-GCN* (Schlichtkrull et al., 2018), and (ii) three extended variants: HeteroSAGE with residual connections, HeteroSAGE with residual and jumping-knowledge connections, and MIX (a combination of HeteroSAGE and GAT). The full model description is provided in Appendix H. Since the longest path between entities in our graph is 3 hops (Figure 1), we ran the ablation up to 3 layers ($L1-L3$) and report the performance of the best model. Full hyperparameters can be found in Appendix H.8.

Structure-only baseline. To isolate the contribution of graph topology, we train the best-performing GNN with random Gaussian input features, matching node dimensionality. Graph structure, training protocol, and evaluation pool are identical to the main configuration (Appendix H.6).

Non-graph baselines. Besides the graph-based models, we evaluate two non-graph baselines that are decoupled from the graph structure to assess the role of graph topology in the task.

- *BLAST:* We define a sequence retrieval baseline by scoring each protein-reaction pair by the max BLASTp bitscore against any training catalyst of that reaction (Hua et al., 2024) (Appendix H.5).
- *Dual-encoder retrieval* (Appendix H.3): The dual-encoder retrieval architecture is widely used for enzyme-reaction association (Mikhael et al., 2024; Rocks et al., 2026; Zhang et al., 2026).

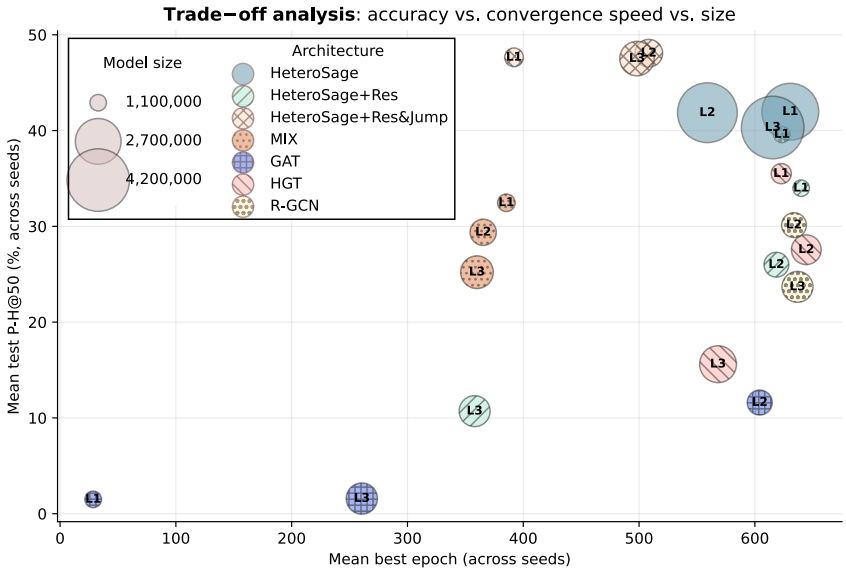

Figure 2: Link-prediction performance on PlantMetBench (taxa holdout, organism nodes removed, mean $\pm$ std across seeds) for all evaluated GNN architectures. Each point shows one model/layer-depth configuration, plotting test P-H@50 against mean epoch of best validation performance (convergence speed). Marker area is proportional to trainable-parameter count (model size).

Reactions and enzymes are encoded independently and projected into a shared space of embeddings. Cosine similarity between a reaction and an enzyme is interpreted as the likelihood of the catalysis occurring. Each representation was fed to a separate MLP, and the resulting embeddings were L2-normalised. The model is trained with a full-batch maximum-likelihood NCE (MLNCE) objective. Each epoch contrasts every training reaction against the entire enzyme pool via a global partition function, admitting multiple catalysing enzymes per reaction without a one-positive-per-query assumption. At inference, the model ranks by cosine similarity to the query. We report results in two configurations: (*i*) on the full 8,445-protein pool, enabling direct comparison with GNNs, and (*ii*) on the 2,232 deduplicated pool, which collapses proteins sharing an identical ESM-C embedding to a single representative, following the model evaluation protocol in Rocks et al. (2026). Both results are compared with a BLAST baseline on the respective set (see Tables 1 and 2).

Training process. All models are trained from scratch (no pretrained models are used). GNN models and dual encoder results are averaged over 5 different seeds (mean $\pm$ std reported). For the GNN models we use the default set-up of the dataset. Appendix H provides all experimental details, per-epoch timing, and convergence behaviour.

4.2 Results on Default Setting

Performance of PlantMetBench with GNN-based models. Table 1 (top) shows GNN model performance. For each model, the table shows the best-performing architecture; the complete results are in Table 23 in the Appendix H.9. Further, Figure 2 presents the test P-H@50 of all models,[3] along with convergence speed and model size. Starting with the two original convolution-based architectures, we notice that HeteroSAGE outperforms R-GCN in all metrics. However, notably, HeteroSAGE is a more 'costly' architecture, as it requires relatively more parameters for training compared to R-GCN, according to Figure 2. Therefore, we integrate residual connections (He et al., 2016; Li et al., 2019) into the HeteroSAGE models, which we denote as HeteroSAGE+Res. As Figure 2 indicates, the residual layers reduce the number of parameters, but the performance gets noticeably worse (from 42% to 34% test P-H@50).

[3]Since per-epoch delays are similar across models (Table 30), we use the number of epochs as the speed measure.

This drop in performance points to information loss during message passing, amplified as depth increases (i.e., at L3 for HeteroSAGE+Res). Therefore, to preserve the reduction in parameters without sacrificing performance, we combine HeteroSAGE+Res with *Jumping Knowledge Layers* (Xu et al., 2018), giving the HeteroSAGE+Res&Jump architecture. This variant improves performance across all ranked metrics (test of best model: 20.7% P-H@10, 48% P-H@50, and $\approx 14\%$ P-H@5, see Table 23), while CP-AUC and CP-AP remain comparable to HeteroSAGE. In Table 1, we present the results for L2, as it performed best, but as Table 23 and Figure 2 show, L1 and L3 perform similarly. This variant provides a good trade-off between accuracy and convergence speed (≈ 400 epochs for L1 and ≈ 500 epochs for L2/L3), while substantially reducing the number of trainable parameters. Further, Figure 3 presents an ablation of the best-performing model as the hidden dimension, denoted by h, is increased. The results show that increasing h improves accuracy, reaching a mean test 51% P-H@50 and 25% P-H@10. However, this also increases the number of trainable parameters to over ≈ 52M, making the model less efficient.

Among the attention-based models, HGT performs similarly to R-GCN, while GAT reaches only 11.6% test P-H@50, placing it last in the leaderboard. Yet GAT's CP-AUC (0.901) remains within 0.04 of the best model's (0.937), despite a $4\times$ lower P-H@50 (11.6% vs. 48.0%). CP-AUC still orders architectures correctly but compresses the magnitude of their differences, *motivating our use of P-H@k as the primary metric* (see also Figure 13 in Appendix H.12). However, as shown in Figure 2, GAT requires only $\approx$1M trainable parameters. To improve its performance, we propose a MIX architecture that combines GAT and HeteroSAGE convolutions in each layer. This improves GAT's test P-H@50 by 20 pp, while keeping the model relatively small.

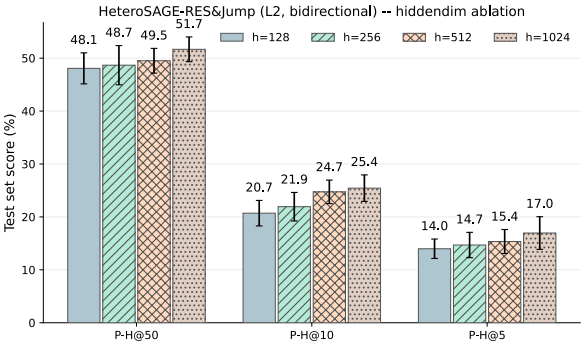

Figure 3: Mean (across seeds) test P-H@50 value (std shown above bar) of HeteroSAGE+Res&Jump (L2) while increasing the number of hidden dimensions.

These results highlight the potential of PlantMetBench, which contains strong signals that conventional GNN models can capture. We further showed that simple architectural adjustments can improve performance, with HeteroSAGE+Res&Jump achieving strong results while still leaving room for improvement. To assess this, we trained the best model on increasing fractions of the training data. As shown in Figure 12 in Appendix H.10, performance has not yet saturated, suggesting that PlantMetBench can support further research into GNN architectures.

Isoenzyme ambiguity (anyP-H@k). The isoenzyme-aware anyP-H@k alternative metric counts a hit if *any* known catalyst of the reaction is retrieved, which is instead penalised under the standard P-H@k. Test P-H@1 rises from $1.7\pm0.6\%$ to $25.0\pm2.2\%$ under anyP-H@1, but the two metrics nearly converge by $k{=}50$ ($48.0\pm2.9\%$ vs. $48.1\pm2.5\%$), *meaning that in that window, retrieving any valid catalyst generally means retrieving also the exact one.* This confirms standard P-H@k as the stricter primary metric chosen. The full breakdown across k is in Appendix G.2.

Structure-only control. The structure-only control (Table 1, Appendix H.6) reaches $2.1\pm0.3\%$ test P-H@50, $3.6\times$ the 0.59% random baseline, far below the $48.0\pm2.9\%$ reached with real features. Graph connectivity alone is therefore mildly informative, with proteins linked through shared reaction neighbourhoods being more predictable than chance. *Hence, most of the model's performance does not depend only on topology but also on the biochemical content of the input features.*

Full-pool comparison. The bottom part of Table 1 presents the performance on the full 8,445-protein pool for the two non-graph baselines.The dual encoder (29.6% test P-H@50) and BLASTp (32.4%) perform similarly. Compared with the GNN models, the dual encoder performs below all models except GAT, and BLASTp below all models except GAT, while HeteroSAGE+Res&Jump reaches 48.0%, a 16–18 pp improvement. *Relational context therefore provides signal beyond sequence and reaction representations alone, though the benefit depends on the architecture used to exploit it.*

Deduplicated-pool comparison. Under the deduplicated 2,232-protein pool (Table 2), where sequence-identical proteins no longer compete for the same rank, the dual encoder improves to 64.5% and clearly overtakes BLASTp (52.7%), *showing that reaction-conditioned joint encoding adds a gain beyond sequence retrieval alone once that structural confound is removed.*

Although non-graph baselines achieve better results (in this case), their linear nature limits their ability to provide meaningful contextual predictions. When the GNN model performs link prediction for protein-conversion, it also associates the predicted protein with its surrounding context, such as the path to the Metabolite and/or Gene-Product. This introduces an evaluation *trade-off*, allowing users to prioritise contextualisation over predictive accuracy. GNNs are not re-evaluated on the deduplicated pool since they already use graph neighbourhood context (including organism identity) to disambiguate sequence-identical proteins across species without needing deduplication. This evaluation is not applicable to our framework, as it would require flattening the current graph into a fundamentally different topology, which may result in a graph without the same biological meaning.

Table 2: Non-graph baselines calculated on 2,232 unique protein embeddings.

Model	**P-H@50 (%)** – Test
Random baseline	2.24
BLASTp baseline	52.68
Dual encoder (L1)	64.5 ± 0.6

4.3 Results on More Challenging Settings

Beyond the default configuration, PlantMetBench supports harder settings that probe different capabilities of the task and surface confounds in how P-H@k should be interpreted.

Sources of non-zero-shot signal. Because PlantMetWiki annotates many reactions across multiple species, 62.9% of test edges (64.7% of val) target a reaction that already has a catalysing edge from a different training-organism protein (Appendix D.8). Restricting to unseen reactions absent from training (Table 16) drops performance conservatively to P-H@50 = 12.6±1.7%, still ≈21× the random baseline. Reactions seen with a different catalyst during training score much higher (66.0±5.6%). This 53.4 pp gap indicates that a substantial share of taxa-holdout performance reflects reaction-level information already in the graph, making the unseen-reaction subset the more conservative generalisation estimate of cross-species transfer. A held-out organism's own neighbourhood (its other proteins and gene products) never connected to training organisms, is another independent source of non-zero-shot signal: isolating it drops test P-H@50 from 48.0% to 26.8% (Appendix D.8). Neither effect is a train/test leak, since the isolated and unseen-reaction edges never touch training organisms directly, *but together they show a substantial share of taxa-holdout performance to be transductive.*

Table 3: Test P-H@50 on the taxa holdout vs. the *A. thaliana* reaction holdout, same architectures. Best model refers to HeteroSAGE+Res&Jump; the random baseline on the *A. thaliana* reaction holdout is 0.94%.

Model	**Taxa holdout**	***A. thaliana***
Best model	48.0±2.9%	58.9±1.7%
Dual encoder	29.6±1.8%	64.6±0.3%

Taxa holdout vs. *A. thaliana* reaction holdout. On the taxa holdout, HeteroSAGE+Res&Jump (48.0% test P-H@50) substantially outperforms the dual encoder (29.6%). Under the *A. thaliana* reaction holdout (Appendix D.6) (no test reaction observed during training under any catalyst), *this reverses: the dual encoder reaches* 64.6% *test P-H@50 against* 58.9% *for HeteroSAGE+Res&Jump* (Table 3). We hypothesize this is because the two splits stress opposite capabilities. More specifically, under the taxa holdout, catalyst proteins are entirely unseen during training, so the dual encoder, lacking graph structure, cannot place them in its learned space, while the GNN propagates information across species through graph topology. Under the *A. thaliana* split, only reactions are held out, reducing the task to closed-pool chemistry-to-catalyst matching, favouring a pure similarity encoder.

Pathway-level reconstruction. P-H@k evaluates each reaction independently, while pathway reconstruction asks whether correct predictions accumulate into complete biosynthetic pathways. A pathway is a chain of enzyme-catalysed reactions, so we score it by the fraction of reactions whose true catalyst is ranked in the top 50 and report the share of pathways recovered fully, not at all, and the mean recovery (full tables in Appendix I.2 and I.1). Under the *A. thaliana* reaction holdout,

HeteroSAGE+Res&Jump fully reconstructs 13.7% of evaluable pathways and misses 18.8% entirely. Under the taxa holdout, full reconstruction drops to 9.4% and misses rise to 22.8%, consistent with cross-species transfer being harder. *This points to full pathway reconstruction as a harder, still largely open challenge than reaction retrieval,* though mean per-pathway reconstruction rates are comparable across the two splits (44.9% vs. 41.6%) and their medians are identical (40.0% both).

Split sensitivity. For the same architecture, val/test P-H@50 is $71.0\pm2.5\%$/$69.0\pm1.4\%$ under a random-edge split, $73.6\pm0.9\%$/$58.0\pm5.2\%$ under pathway holdout, and $49.9\pm3.4\%$/$48.0\pm2.9\%$ under the taxa holdout (full results in Appendix D.7, Table 14), *a swing of over 20 p.p on test.*

Shortcut ablation. PlantMetBench identifies four structural shortcuts in the raw PlantMetWiki graph (full shortcut ablation in Appendix F). The most damaging, the reverse catalysis edge, drops test P-H@50 from 48.0% to 36.3% (Table 4), a 2-hop shortcut that trains the model toward a spurious signal that does not transfer to test. Disabling the disjoint-train split raises test P-H@50 by 10.1 pp Pathway co-membership and EC one-hot features fall within one standard deviation of baseline and are removed for mechanistic reasons (Table 4). *These results show the importance of careful evaluation design for reliable conclusions about model performance on biosynthesis prediction.*

Table 4: Effect of each shortcut on the organism-removed taxa-holdout baseline (mean ± std).

Configuration	Val (%)	Test (%)
Baseline	49.9 ± 3.4	48.0 ± 2.9
+ Pathway co-membership	50.2 ± 2.7	47.6 ± 4.8
+ Reverse catalysis edge	37.6 ± 4.2	36.3 ± 3.0
+ No disjoint split	59.3 ± 2.9	58.1 ± 2.0
+ EC-number features	49.0 ± 3.9	47.7 ± 3.2

5 Discussion & Conclusion

Our evaluation shows that architecture choice determines whether GNNs add value over non-graph baselines: GAT (11.6%) trails both BLASTp (32.4%) and the dual encoder (29.6%), but residual and jumping-knowledge connections (HeteroSAGE+Res&Jump) reach 48.0% test P-H@50 and, unlike most other architectures, stay flat and do not degrade with depth. We also present further challenges that point towards future approaches. First, the *A. thaliana* reaction holdout shows the dual encoder overtakes Res&Jump (64.6% vs. 58.9%). Second, pathway reconstruction shows per-reaction hits rarely compose into fully-recovered pathways (9.4–13.7%), pointing towards a *harder and more biologically meaningful bar* than P-H@k alone. Further, we show that careful benchmark design (disjoint splits, structural shortcuts, and metrics) keeps evaluation aligned with candidate enzyme prioritisation in realistic scenarios. These results demonstrate that PlantMetBench contains knowledge signal based on ground-truth experimental data and sets the bar for a future benchmark leaderboard.

Limitations. PlantMetBench inherits the annotation biases of the curated PlantMetWiki, with the model species *A. thaliana* alone accounting for 37% of catalysis edges ($4{,}459/11{,}980$) and representing an upper coverage baseline (97.6% ESM-C coverage versus a 61.7% benchmark-wide average). Feature coverage is also incomplete across node types, as the data is based on *real-world* curated knowledge. Additionally, the current graph representation omits other modalities such as gene expression, enzyme kinetics, spatial compartmentalisation, and post-translational regulation. These curated annotations are experimentally supported positives, so an unlabelled pair is not a confirmed negative. Independent cross-validation against the reaction database Rhea (Bansal et al., 2022) further suggests the benchmark undercounts true positives: proteins ranked highly by the model but unlabelled by PlantMetWiki are supported by Rhea at 4.1–$5.9\times$ the random rate under the taxa holdout and the *A. thaliana* reaction holdout, respectively (Appendix I.3, I.4). Future work might expand our benchmark beyond plant metabolism (e.g., microbial or mammalian), thanks to interoperable pathway resources such as WikiPathways (Agrawal et al., 2024).

Conclusions. PlantMetBench is designed as a reproducible, comparable, and extensible benchmark ecosystem, including datasets, splits, embedding pipelines, and evaluation code, for developing and testing new GNN architectures that can best take advantage of the multimodal graph structure of metabolism. The results suggest that *A. thaliana* within-species retrieval and cross-species transfer are different enough problems to require dedicated tasks. The graph's remaining modalities and edge types point toward potential further task designs beyond enzyme–reaction ranking. We leave these directions to future versions of the benchmark.

AI USE STATEMENT

For writing, AI tools (Claude, Anthropic) were used to edit and polish author-written text, and to assist drafting portions of the Methodology, Results, and Conclusions sections and the Appendix from author-specified content and results, which authors then reviewed and revised. For research execution, AI tools assisted in writing, documenting, and debugging data-processing and evaluation code, and in running the training and evaluation experiments underlying several reported ablations (including the pathway reconstruction, Rhea-based validation, and held-out-organism neighbourhood ablation analyses), under author direction and with authors specifying the experimental design. AI tools assisted in writing Python scripts to generate plots, based on authors' experimental results, which are included in the released code.

We reviewed all AI-assisted work. All AI-generated code was inspected and, where it produced reported numerical results, cross-checked against independent computations or existing baselines before inclusion. All AI-drafted text was reviewed, fact-checked against the underlying results, and revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we release code, data, and precomputed embeddings as versioned and citable artefacts. The full PlantMetBench pipeline, including graph construction, data loaders, data cleaning, node embedding generation, training, and evaluation, is available at `https://anonymous.4open.science/r/plantmetbench-0391/`, together with the scripts used to produce every reported result, figure, and table. An environment specification is provided for data preparation and embedding generation, since it depends on external pretrained models (ESM-C, PlantCaduceus) with separate requirements from the training pipeline. The precomputed node embeddings are additionally archived as versioned Zenodo records with permanent DOIs, so the exact snapshot used in this paper stays fixed and citable (links will be made public upon acceptance). Dataset construction and cleaning are detailed in Appendix B, node embedding computation in Appendix C, train/validation/test split construction in Appendix D, model architectures in Appendix H, and full hyperparameter settings in Appendix H.8.

REFERENCES

Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L Coort, Marvin Martens, Denise N Slenter, Friederike Ehrhart, Daniela Digles, Andra Waagmeester, Isabel Wassink, et al. WikiPathways 2024: next generation pathway database. *Nucleic Acids Research*, 52(D1):D679–D689, 2024.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140.

Parit Bansal, Anne Morgat, Kristian B Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, Elisabeth Gasteiger, Arnaud Kerhornou, Teresa Batista Neto, et al. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research*, 50(D1):D693–D700, 2022.

Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M. Bronstein, Mathias Niepert, Bryan Perozzi, Mikhail Galkin, and Christopher Morris. Position: Graph learning will lose relevance due to poor benchmarks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81067–81089. PMLR, 2025.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic Acids Research*, 28(1): 235–242, 2000.

Anna Breit, Simon Ott, Asan Agibetov, and Matthias Samwald. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, 36(13):4097–4098, Jul 2020. doi: 10.1093/bioinformatics/btaa274.

Roman Bushuiev, Anton Bushuiev, Niek F de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. MassSpecGym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10: 421, 2009.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology, 2026. URL `https://www.biorxiv.org/content/10.64898/2026.06.03.729735`. Preprint.

Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):43, 2020.

João Capela, João Cheixo, Dick de Ridder, Oscar Dias, and Miguel Rocha. Predicting precursors of plant specialized metabolites using DeepMol automated machine learning. *Journal of Integrative Bioinformatics*, 22(2):20240050, 2025. doi: 10.1515/jib-2024-0050.

Louisa Cornelis, Johan Mathe, Louis Van Langendonck, Guillermo Bernárdez, and Nina Miolane. OgBench: A framework for evaluating graph neural networks on omics data, 2026. URL `https://arxiv.org/abs/2605.15511`.

Elena Del Pup, Max Muller, Marvin Martens, Egon L Willighagen, Marnix H Medema, Denise Slenter, and Justin JJ van der Hooft. PlantMetWiki: a FAIR knowledge graph for plant metabolic pathway cross-species representation and integration. *bioRxiv*, 2026. doi: 10.64898/2026.07.22.733699.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR 2019 Workshop on Representation Learning on Graphs and Manifolds*, 2019.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 1024–1034, 2017.

Charles Hawkins, Bo Xue, Farida Yasmin, Gabrielle Wyatt, Philipp Zerbe, and Seung Y Rhee. Plant metabolic network 16: expansion of underrepresented plant groups and experimentally supported enzyme data. *Nucleic Acids Research*, 53(D1):D1606–D1613, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 22118–22133. Curran Associates, Inc., 2020a.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of The Web Conference 2020 (WWW)*, pp. 2704–2710, 2020b. doi: 10.1145/3366423.3380027.

Chenqing Hua, Bozitao Zhong, Sitao Luan, Liang Hong, Guy Wolf, Doina Precup, and Shuangjia Zheng. ReactZyme: A benchmark for enzyme-reaction prediction. *Advances in Neural Information Processing Systems*, 37:26415–26442, 2024.

Wenjuan Ji, Anne Osbourn, and Zhenhua Liu. Understanding metabolic diversification in plants: branchpoints in the evolution of specialized metabolism. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1914):20230359, 2024.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research*, 53(D1): D672–D677, 2025.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023.

Martina Kutmon, Martijn P van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, Alexander R Pico, and Chris T Evelo. PathVisio 3: an extendable pathway analysis toolbox. *PLoS Computational Biology*, 11(2):e1004085, 2015.

Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. DeepGCNs: Can GCNs go as deep as CNNs? In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9267–9276, 2019.

Peter G. Mikhael, Itamar Chinn, and Regina Barzilay. CLIPZyme: Reaction-conditioned virtual screening of enzymes. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35647–35663. PMLR, 2024.

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.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+)*, 2020.

Charlotte Owen, Burkard Steuernagel, Neil Hall, Brooks Paige, and Anne Osbourn. Mind the gap: the challenges and opportunities for genomics-driven harnessing of plant metabolic diversity for therapeutic applications. *Current Opinion in Biotechnology*, 99:103510, 2026.

Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, and Tero Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in Bioinformatics*, 16(2):325–337, 2015. doi: 10.1093/bib/bbu010.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery*, 1(2):91–97, 2022.

Jason W. Rocks, Dat P. Truong, Dmitrij Rappoport, Samuel Maddrell-Mander, Daniel A. Martin-Alarcon, Toni M. Lee, Steven Crossan, and Joshua E. Goldford. Dual-encoder contrastive learning accelerates enzyme discovery. *Proceedings of the National Academy of Sciences*, 123(12): e2520070123, 2026. doi: 10.1073/pnas.2520070123.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference (ESWC)*, pp. 593–607, 2018. doi: 10.1007/978-3-319-93417-4_38.

Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, et al. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020.

Federico Scossa, Mustafa Bulut, Thomas Naake, John C D'Auria, and Alisdair R Fernie. Convergence and parallelism in the evolution of plant metabolism. *Journal of Integrative Plant Biology*, 68(4): 1013–1031, 2026.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 3145–3153, 2017. URL https://arxiv.org/abs/1704.02685.

Timo Stoll, Chendi Qian, Ben Finkelshtein, Ali Parviz, Darius Weber, Fabrizio Frasca, Hadar Shavit, Antoine Siraudin, Arman Mielke, Marie Anastacio, Erik Müller, Maya Bechler-Speicher, Michael Bronstein, Mikhail Galkin, Holger Hoos, Mathias Niepert, Bryan Perozzi, Jan Tönshoff, and Christopher Morris. GraphBench: Next-generation graph learning benchmarking, 2025. URL https://arxiv.org/abs/2512.04475.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Karel van der Weg, Erinc Merdivan, Marie Piraud, and Holger Gohlke. TopEC: prediction of enzyme commission classes by 3D graph neural networks and localized 3D protein descriptor. *Nature Communications*, 16(1):2737, 2025. doi: 10.1038/s41467-025-57324-5.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *International conference on machine learning*, pp. 5453–5462. PMLR, 2018.

Jingjing Zhai, Aaron Gokaslan, Yair Schiff, Ana Berthel, Zong-Yan Liu, Wei-Yun Lai, Zachary R Miller, Armin Scheben, Michelle C Stitzer, M Cinta Romay, et al. Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained dna language model. *Proceedings of the National Academy of Sciences*, 122(24):e2421738122, 2025.

Yuhang Zhang, Keyan Ding, Peilin Chen, Han Liu, Can Lin, Ruixi Chen, Shiqi Wang, and Qi Song. TIGER: Text-informed generalized enzyme-reaction retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35530–35548, San Diego, California, USA, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1643.

TECHNICAL APPENDIX

A PlantMetBench Overview **16**

B Dataset Construction **16**

B.1 Source Knowledge Graph 16

B.2 Graph Construction Pipeline 17

C Node Representations **19**

C.1 Protein Embeddings (ESM-C) 19

C.2 Gene Embeddings (PlantCaduceus) 19

C.3 Metabolite Embeddings (MAP4) 21

C.4 Reaction Fingerprints (MAP4 DRFP-approximation) 23

C.5 EC Hierarchy Embeddings (Ablation Only) 24

C.6 Organism Taxonomy Embeddings (Ablation Only) 25

D Splitting Policies: Design and Analysis **26**

D.1 Taxa Holdout Design 26

D.2 Organism Assignment Criteria 26

D.3 Split Statistics 26

D.4 Disjoint Training Supervision 27

D.5 Embedding Coverage Across the Taxa Holdout 27

D.6 *A. thaliana* Reaction Holdout (EC-stratified) 28

D.7 Split Comparison 29

D.8 Held-Out Organism Neighbourhood Isolation 31

E Cofactor / Currency Metabolite Removal **32**

E.1 Currency Metabolite List 32

F Shortcut Analysis **32**

F.1 Pathway Co-membership 32

F.2 Reverse Catalysis Edge 33

F.3 1-Hop Training Supervision 33

F.4 Enzyme Class (EC) Feature 33

F.5 Shortcut results 33

F.6 Organism Modality Ablation 34

G Evaluation Metrics **35**

G.1 Protein Hits at K (P-H@K) 35

G.2 Any-Catalyst P-H@K (anyP-H@K) — Isoenzyme-Aware Variant 36

G.3 Corrupt-Protein AUC and Average Precision (CP-AUC, CP-AP) 36

G.4 Why Not Standard Link-Prediction AUC? 37

H Extended Numerical evaluation **37**

H.1 GNN baselines 37

H.2 HeteroSAGE with Residual & Jumping Knowledge 37

H.3 Dual-Encoder Retrieval 38

H.4 MIX model 38

H.5 BLASTp Sequence-Retrieval Baseline 38

H.6 Structure-Only Baseline 39

H.7 Training Objective and Negative Sampling 39

H.8 Hyperparameters 39

H.9 Results 40

H.10 Model saturation discussion 42

H.11 Results on the Deduplicated 2,232-Protein Pool 42

H.12 Diagnostics 43

H.13 *A. thaliana* Reaction Holdout Results 45

I Pathway Reconstruction **46**

I.1 Pathway Reconstruction Analysis (Taxa Holdout) 46

I.2 Pathway Reconstruction Analysis (*A. thaliana* Reaction Holdout) 48

I.3 Rhea-Based Independent Validation (Taxa Holdout) 49

I.4 Rhea-Based Independent Validation (*A. thaliana* Reaction Holdout) 50

J Computational Resources **51**

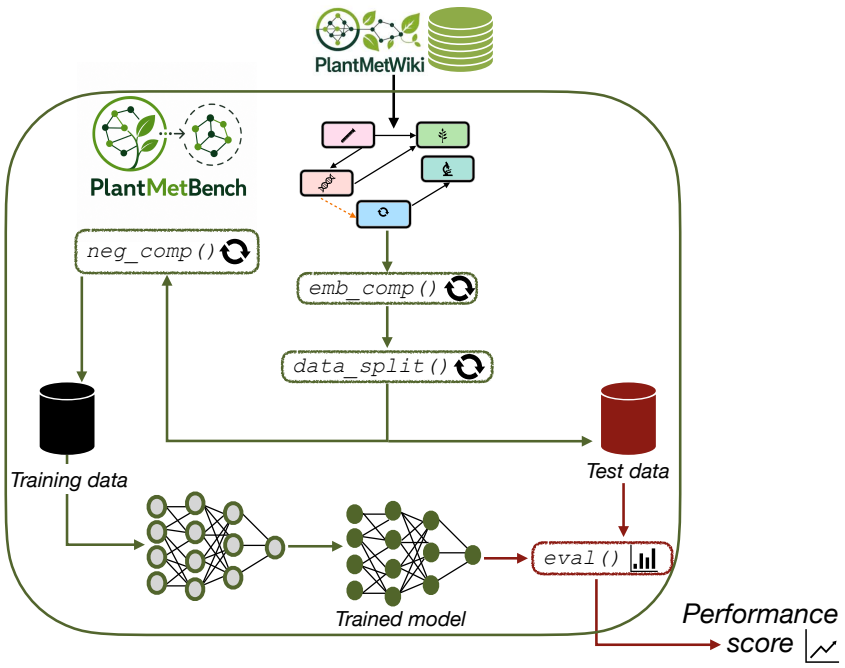

Figure 4: PlantMetBench evaluation workflow.

A PLANTMETBENCH OVERVIEW

PlantMetBench is both a graph dataset and an evaluation workflow and benchmarking ecosystem. Figure 4 presents the workflow of the PlantMetBench framework, as described in Section 3. First, PlantMetBench fetches the data from a database repository. Specifically, for the case studied in this work, data are retrieved from PlantMetWiki (Del Pup et al., 2026). For this step, the proposed framework provides all the functionalities for transforming the raw data of a pathway database like PlantMetWiki into a graph dataset suitable for training.[4] In Appendix B, we describe how the graph is constructed, and in Appendix C, how the embeddings for each node are computed. Next, one of the available data splitting policies is applied. In Appendix D, we present the splitting policies that we investigated. We identified that the taxa holdout split is the most challenging one and hence is the one used in this study.

Once the training data are ready, the training process starts. The PlantMetBench benchmarking ecosystem is not restricted to a specific machine learning (ML) case, and hence one can incorporate any ML model. PlantMetBench has a modular design. When the model is trained, it is evaluated using the available metrics (see Appendix G). This combination of metrics provides a new standardised evaluation ecosystem for biosynthesis prediction, in plants and beyond.

B DATASET CONSTRUCTION

This section contains how the benchmark graph is built from PlantMetWiki, including a description of the source knowledge graph, the import and cleaning pipeline, and the resulting node and edge statistics.

B.1 SOURCE KNOWLEDGE GRAPH

In this work, we use PlantMetWiki (Del Pup et al., 2026), a curated, RDF-based knowledge graph of plant primary and specialised metabolism built from PlantCyc (Hawkins et al., 2025), the pathway

[4]PlantMetBench preprocessing is not restricted to graph-based datasets. For example, PlantMetBench implements the dataset construction for a dual-encoder model that does not require graph representations.

database of the Plant Metabolic Network, in which curators manually annotate pathways, reactions and enzymes based on experimental (wet-lab) evidence documented in the scientific literature. It encodes biochemical reactions, the metabolites they transform, the enzymes that catalyse them, and the genes that encode those enzymes, all linked to source pathways and to the plant organisms in which each association has been characterised. Knowledge is represented as Resource Description Framework (RDF) triples and exported as files in Turtle (`.ttl`) format. An overview of the graph is provided in Tables 5 and 6.

Table 5: Node types in the PlantMetBench graph. **Raw** is as imported into `heterodata.pt`, **Default** is the training-time graph after `remove_is_part_of` (drops the Pathway node store) and graph cleaning (drops alias Metabolite/GeneProduct nodes and blank-subtype Interaction nodes – every remaining Interaction node is Conversion-subtype).

Node type	Role	# (raw)	# (default)	Feat. dim	Source
Protein	Enzyme candidates (query)	13,682	13,682	960	ESM-C (8,445 covered)
Interaction	Enzymatic conversions (target)	84,470	19,927	3,072	MAP4 DRFP-approx
Metabolite	Substrate / product molecules	17,476	4,577	1,024	MAP4
GeneProduct	Encoding genes	6,480	2,741	1,024	PlantCaduceus
Pathway	Metabolic pathways (*removed*)	2,478	–	–	–
Organism	Plant species (bidirectional by default)	424	424	64	Random / taxonomy MDS

B.2 Graph Construction Pipeline

The default training graph retains 13 edge types and 135,268 edges. We build the graph in three stages: *import* (raw RDF into a typed PyTorch Geometric object), *cleaning* (removing structural artefacts of the source format), and *feature loading* (attaching domain embeddings). Tables 5 and 6 report both the **raw** counts (as imported) and the **default** counts (what the model trains on, after cleaning and the shortcut removals of Appendix F).

Import. We parse 3.8 million RDF triples from the PlantMetWiki Turtle export using `rdflib`. Nodes are typed according to their RDF class (`biopax:Protein`, `biopax:BiochemicalReaction`, `biopax:SmallMolecule`, etc.) and edges are typed by predicate. Organism assignments are resolved via a supplementary NCBI taxonomy RDF file (`taxonomy.ttl`), which maps each biological entity to its NCBI Taxon Internal Resource Identifiers (IRI) (Schoch et al., 2020). The typed tables (`nodes.tsv`, `edges.tsv`) are then loaded into a `HeteroData` object and serialised as `heterodata.pt` using PyTorch Geometric (Fey & Lenssen, 2019).

Interaction vs. Conversion. INTERACTION is the generic PyTorch Geometric node type for *every* `biopax:Interaction`-derived entity. While CONVERSION is PlantMetWiki's label for an actual enzymatic reaction and the **only** subtype used for supervision (via a `conv_idxs` index list). Of the 84,470 raw Interaction nodes, only 19,927 carry `interaction_subtype = Conversion`. The remaining 64,543 are blank-subtype: 19,927 of these are pathway-scoped *alias* copies of the Conversion nodes themselves (one alias per Conversion, linked by an `is_about` edge). The remaining ≈44,616 are a mix of genuinely other interaction types (transport, binding, complex formation) and sub-reaction hierarchy nodes, of which ≈24,990 are entirely disconnected in the raw export. After the graph-cleaning step, every remaining Interaction node *is* Conversion-subtype, so we use "Interaction" and "Conversion" interchangeably from that point on.

Schema alignment. The raw RDF predicates (`participants`, `source`, `target`, ...) are re-typed to a cleaner relational schema. The central supervision edge is renamed `(Protein, catalyzes, Conversion)`, the curated catalysis annotations. Its reverse, `(Conversion, catalyzed_by, Protein)`, is retained in the raw schema but removed by default during training (see Appendix F), to avoid 2-hop shortcuts. Message passing on `catalyzes` is further restricted, in

Table 6: Edge types in the PlantMetBench graph. **Raw** is as imported, **Default** is the training-time message-passing graph. Note, "removed" refers to the case where the edge type is absent from the default graph. The mechanism differs by category (see row groups). The `catalyzes` edge is further restricted, in every split, to a fixed disjoint 7,655-edge subset of the 11,980 raw training-organism-available positives – see Appendix F.

Edge type	Description	# (raw)	# (default)
Supervision and catalysis			
(Protein, **catalyzes**, Interaction)	**Supervision target**	11,980	7,655[†]
(Interaction, catalyzed_by, Protein)	Reverse supervision	11,980	removed
Organism assignment (bidirectional by default, but removed in default evaluation configurations)			
(Protein, organism, Organism)	Species origin of protein	13,725	13,725
(GeneProduct, organism, Organism)	Species origin of gene	6,504	2,747
(Organism, rev_organism, Protein)	Reverse, added for message passing	–	13,725
(Organism, rev_organism, GeneProduct)	Reverse, added for message passing	–	2,747
(Metabolite, organism, Organism)	Species origin of metabolite	17,489	removed
Reaction–metabolite connectivity			
(Interaction, participants, Metabolite)	Substrate or product of reaction	34,793	26,062
(Metabolite, rev_participants, Interaction)	Reverse, added for message passing	–	26,062
(Metabolite, is_part_of, Interaction)	Legacy substrate/product edge	34,793	removed
(Interaction, source, Metabolite)	Substrate of reaction	7,554	7,537
(Interaction, target, Metabolite)	Product of reaction	18,622	18,622
Gene–protein linkage			
(GeneProduct, encodes, Protein)	Gene encodes protein	2,762	2,762
(Protein, encoded_by, GeneProduct)	Reverse of above	2,762	2,762
Alias / pathway-instance linkage (see Graph cleaning)			
(Protein, is_about, Protein)	Protein → its pathway-scoped alias (not yet cleaned)	9,932	9,932
(GeneProduct, is_about, GeneProduct)	GeneProduct → its pathway-scoped alias	4,580	930
(Metabolite, is_about, Metabolite)	Metabolite → its pathway-scoped alias	12,812	removed
Reaction hierarchy (collapses under graph cleaning)			
(Interaction, is_part_of, Interaction)	Sub-reaction nesting	9,813	removed
(Interaction, participants, Interaction)	Reaction–reaction participation	9,813	removed
(Interaction, is_about, Interaction)	Reaction → pathway-instance alias	29,740	removed
*Pathway membership (*all removed by `remove_is_part_of`*)*			
(Interaction, is_part_of, Pathway)	Reaction pathway membership	29,740	removed
(Metabolite, is_part_of, Pathway)	Metabolite pathway membership	22,109	removed
(Protein, is_part_of, Pathway)	Protein pathway membership	10,660	removed
(GeneProduct, is_part_of, Pathway)	Gene pathway membership	8,215	removed
(Pathway, is_about, Pathway)	Cross-species pathway equivalence	2,478	removed

[†] Identical across train/val/test message-passing graphs by construction (Appendix F); only the supervision (`edge_label_index`) differs per split.

every split, to a fixed 80% disjoint subset of the training positives (7,655 of 11,980 edges) to prevent a 1-hop shortcut. See Appendix F for the full protocol.

Graph cleaning. PlantMetWiki's per-pathway Graphical Pathway Markup Language (GPML) (Kutmon et al., 2015) source format instantiates a separate copy of each Metabolite and GeneProduct in every pathway diagram it appears in, on top of the canonical (`identifiers.org`-namespaced) node. For instance, CO_2 exists once canonically and again as a distinct node in each of the ≈386 pathways it participates in. These *alias* nodes carry no independent chemistry or sequence information (zero embedding at construction time) and are linked to their canonical counterpart by an `is_about` edge. A post-processing step, `build_clean_ctx` (applied at load time, *not* baked into `heterodata.pt`, and on by default via `--data-type clean`), drops these alias Metabolite and GeneProduct nodes together with the blank-subtype Interaction nodes described above, reducing Metabolite from 17,476 to 4,577 nodes and GeneProduct from 6,480 to 2,741 (Table 5). Protein aliases follow the same naming pattern but are *not* currently dropped by this step. This is a known asymmetry we flag, since `(Protein, is_about, Protein)` (9,932 edges, Table 6) is otherwise unaffected.

Feature loading. Metabolite (MAP4, 1,024-dim) and GeneProduct (PlantCaduceus, 1,024-dim) features are baked into `heterodata.pt` at graph-creation time. Protein (ESM-C, 960-dim) and Conversion fingerprint (DRFP approximation, 3,072-dim) features are loaded at runtime from separate

`embeddings_*.pt` files, independent of graph topology. Organism nodes receive 64-dimensional taxonomy-aware MDS coordinates, as well as a random 64-dimensional alternative (see Appendix C).

Bidirectional message passing (`--bidirectional`, on by default). By default, ORGANISM and METABOLITE nodes are given a reverse edge alongside their existing forward edge. `(Organism, rev_organism, Protein)` and `(Organism, rev_organism, GeneProduct)` let an organism's proteins and genes share context through their common Organism node. This is safe with respect to the taxa holdout, since organisms are disjoint train/val/test (Appendix D.1), so a reverse edge can only pool information within one organism and can never move information across that boundary. `(Metabolite, rev_participants, Conversion)`, alongside the existing `(Conversion, participants, Metabolite)` edge, gives Conversion nodes a live aggregation over their substrate/product metabolites' MAP4 embeddings, in addition to the static, precomputed DRFP fingerprint feature. Both can be disabled with `--no-bidirectional`. ORGANISM nodes can alternatively be removed from the graph entirely via `--remove-organism-nodes` (a pure-biochemistry ablation with no species topology at all), in which case the reverse edges above are skipped automatically. `(Metabolite, organism, Organism)` is removed outright by default regardless (`--remove-metabolite-organism-edges`), as 99.6% of these edges (17,420 of 17,489) assign the metabolite to *Viridiplantae* (the plant kingdom), carrying no species-discriminative signal.

C NODE REPRESENTATIONS

Each node type is initialised with a precomputed, domain-specific feature vector, as summarised under *Feature loading* in Appendix B. We detail per-modality embedding construction and coverage below.

C.1 PROTEIN EMBEDDINGS (ESM-C)

Model. We use ESM-C Candido et al. (2026) (Evolutionary Scale Modelling – Cambrian), a protein language model pretrained on large-scale protein sequence databases. We embed each protein as the mean-pool of the final-layer token representations across all residue positions, yielding a **960-dimensional** vector per protein.

Sequence retrieval. Protein sequences are fetched from UniProt (The UniProt Consortium, 2023) (reviewed Swiss-Prot entries preferred; TrEMBL as fallback) and RCSB PDB (Berman et al., 2000) using the UniProt accession codes extracted from PlantMetWiki IRIs. Of the 13,682 Protein nodes, **8,445** (61.7%) have a resolved amino-acid sequence and a non-zero ESM-C embedding.

Zero-embedding proteins. The remaining **5,237** Protein nodes (38.3%) have no resolvable sequence and receive a zero vector. These proteins participate in the message-passing graph (the GNN can propagate information *to* them via their graph neighbours), but they are **excluded from the evaluation ranking pool**: a zero-embedding protein cannot be meaningfully ranked against real embeddings, and including it would penalise models for a data-coverage issue. Zero-embedding proteins may still appear in training supervision pairs, where they contribute gradient signal via graph-neighbourhood aggregation.

Figure 5 shows PCA, t-SNE, and UMAP projections of the 8,445 embedded proteins coloured by organism. The UMAP panel reveals species-level clustering: proteins from the same organism form cohesive local neighbourhoods, reflecting that ESM-C encodes phylogenetic similarity from sequence alone without any explicit organism input. This co-species structure means that held-out species occupy distinct but nearby regions of the embedding manifold, so proximity in ESM-C space alone cannot trivially solve species assignment, ruling out that specific shortcut.

C.2 GENE EMBEDDINGS (PLANTCADUCEUS)

Model. We use PlantCaduceus Zhai et al. (2025) (checkpoint `PlantCaduceus_l32`), a DNA language model pretrained with a masked-language-modelling objective on 16 angiosperm genomes. It is built on the bidirectional, strand-aware Caduceus/Mamba (state-space model) architecture and

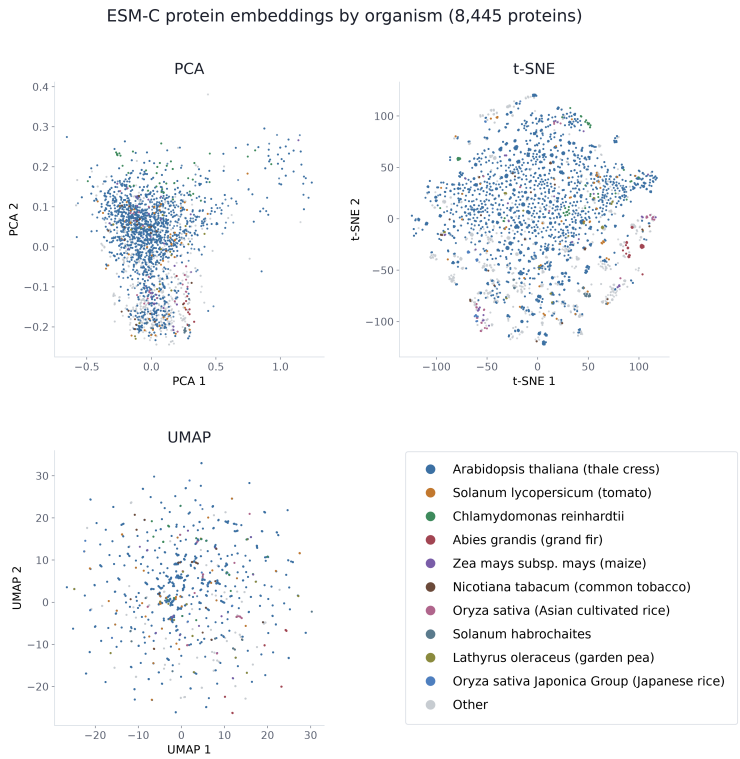

Figure 5: PCA, t-SNE, and UMAP projections of ESM-C protein embeddings (8,445 embedded proteins) coloured by organism. The UMAP projection shows species-level clustering, *A. thaliana* (blue) forms a cohesive neighbourhood, indicating that ESM-C sequence representations implicitly capture phylogenetic proximity. The t-SNE shows substantial inter-species mixing at finer scales, ruling out simple sequence-proximity lookup as a shortcut for cross-species prediction.

accepts input windows of up to 512 bp. Each gene sequence is split into non-overlapping 512-bp windows, with final windows shorter than 256 bp being discarded (unless it is the only window). Each window is embedded independently: the final-layer hidden states of the forward and reverse-complement strands are averaged and mean-pooled over positions, and the window vectors of a gene are then averaged, yielding a **1,024-dimensional** vector per gene. Genes have a median of 5 windows (maximum 260), and no information is shared between windows. Because the model reads nucleotide sequence, the embedding adds nucleotide-level information that the amino-acid-based ESM-C protein embeddings do not capture, such as codon usage and untranslated regions (UTRs).

Sequence retrieval. Each GeneProduct is resolved to a nucleotide sequence from NCBI through one of three routes, chosen by its identifier (Table 7). (i) *Locus tag.* For genes identified by a locus tag (e.g., an *A. thaliana* AGI code), we query NCBI Gene for the gene's annotated genomic interval and retrieve exactly that interval from NCBI Nucleotide, on the annotated strand. (ii) *GenBank accession.* For genes identified by a GenBank nucleotide accession, we retrieve the full nucleotide record. (iii) *UniProt.* For genes known only through a UniProt accession, we follow its EMBL/GenBank cross-reference and retrieve that record. No sequence upstream or downstream of the gene is added, so each embedding describes the gene's own sequence (including introns and UTRs where they fall within the retrieved interval or record) and not its promoter or neighbouring genes. Routes (ii) and (iii) retrieve whole database entries, so a few long entries (ten longer than 5 kb, two longer than 50 kb) may contain sequence beyond the gene itself.

Coverage. Of the 6,480 GeneProduct nodes in the raw graph, **1,359** (21.0%) have a resolved sequence and a non-zero embedding. 1,354 of them are among the 2,741 GeneProduct nodes of the training graph (49.4%); the others are pathway-scoped alias nodes, which the graph-cleaning step

Table 7: Resolution routes of the 1,359 GeneProduct sequences. Lengths in bp.

Route	Genes	%	Median length	Max length
(i) Locus tag → NCBI Gene interval	1,056	77.7	2,737	18,192
(ii) GenBank accession record	300	22.1	1,622	132,990
(iii) UniProt → EMBL/GenBank record	3	0.2	1,736	82,503
All resolved sequences	1,359	100	2,441	132,990

removes (Appendix B). Nodes without a resolved sequence receive a zero vector and participate in message passing only via graph neighbours.

Figure 6 shows projections of the 1,359 embedded GeneProduct nodes coloured by organism. The pattern is intermediate between proteins and metabolites: *A. thaliana* genes (blue) form a locally cohesive cluster, consistent with species-level differences in gene sequence composition. Coverage is lower than for proteins (49.4% of training-graph GeneProduct nodes vs. 61.7% of proteins), and the clustering is less pronounced. GeneProduct nodes take part in message passing but are not ranked, since the prediction task ranks Protein nodes.

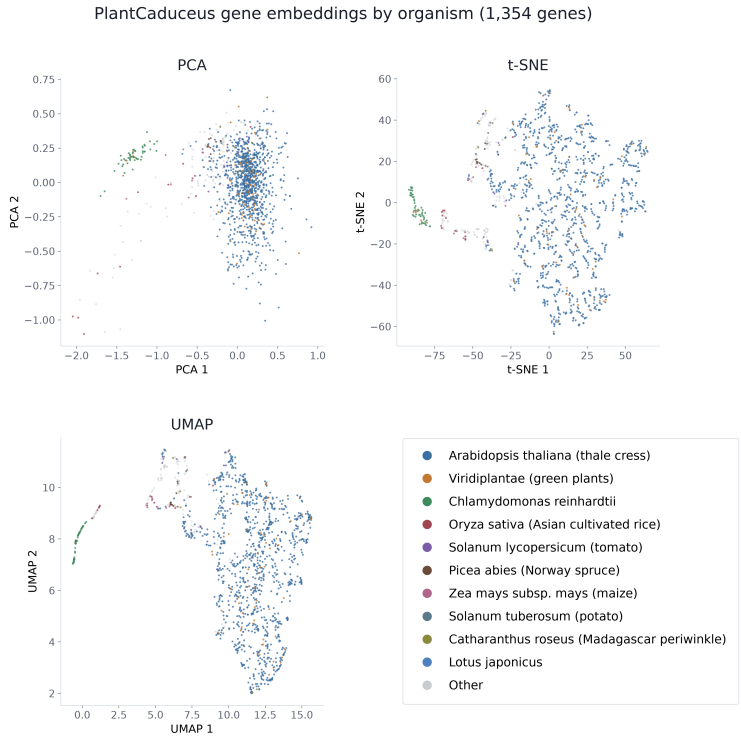

Figure 6: PCA, t-SNE, and UMAP projections of PlantCaduceus gene embeddings (1,359 embedded GeneProduct nodes) coloured by organism. *A. thaliana* (blue) forms a locally cohesive cluster in UMAP. Coverage is lower than for proteins (49.4% of training-graph GeneProduct nodes vs. 61.7% of proteins), and the clustering is less pronounced.

C.3 Metabolite Embeddings (MAP4)

Model. We use MAP4 Capecchi et al. (2020) (MinHashed Atom-Pair Fingerprint, version 2), a circular fingerprint that encodes atom-pair environments up to a configurable radius using a MinHash scheme, computed from SMILES. MAP4 is specifically designed for small molecules including natural products and is more sensitive to substructure similarity in plant metabolites than classical Morgan/ECFP fingerprints. Each resolved metabolite is encoded as a **1,024-dimensional** binary vector. MAP4 metabolite features serve a dual role: they are assigned directly as `Metabolite.x`

node features in the message-passing graph, and are separately aggregated offline to produce the reaction fingerprints in §C.4.

SMILES resolution. SMILES strings are resolved from PubChem (Kim et al., 2023) (by InChIKey), MetaNetX (Moretti et al., 2021) (by MetaNetX ID), and KEGG (Kanehisa et al., 2025) (by KEGG compound ID) in that priority order. The 17,476 Metabolite nodes comprise two populations: 4,577 canonical nodes with global chemical identifiers, and 12,899 pathway-scoped *alias* nodes (per-pathway copies of a canonical compound) that carry no independent chemical signal and are excluded from SMILES resolution entirely. Of the 4,577 canonical Metabolite nodes, **4,199** (91.7%) have a resolved SMILES and a non-zero MAP4 embedding. A further **295** SMILES are resolved but contain R-group wildcard atoms (`[*]`); these are valid MAP4 inputs but cannot be converted to InChIKey and are flagged in the embedding inputs archive. The remaining 378 canonical Metabolite nodes (8.3%) have no resolvable SMILES and are assigned a zero vector.

Figure 7 projects the 4,529 embedded metabolites coloured by organism. Unlike proteins, metabolites show almost no species-level clustering: the same small molecules (substrates, cofactors, intermediates) appear across many plant species, so MAP4 fingerprints primarily encode chemical structure, not taxonomic origin. This is the expected behaviour and confirms that metabolite representations are organism-agnostic biochemical descriptors, not species-specific features.

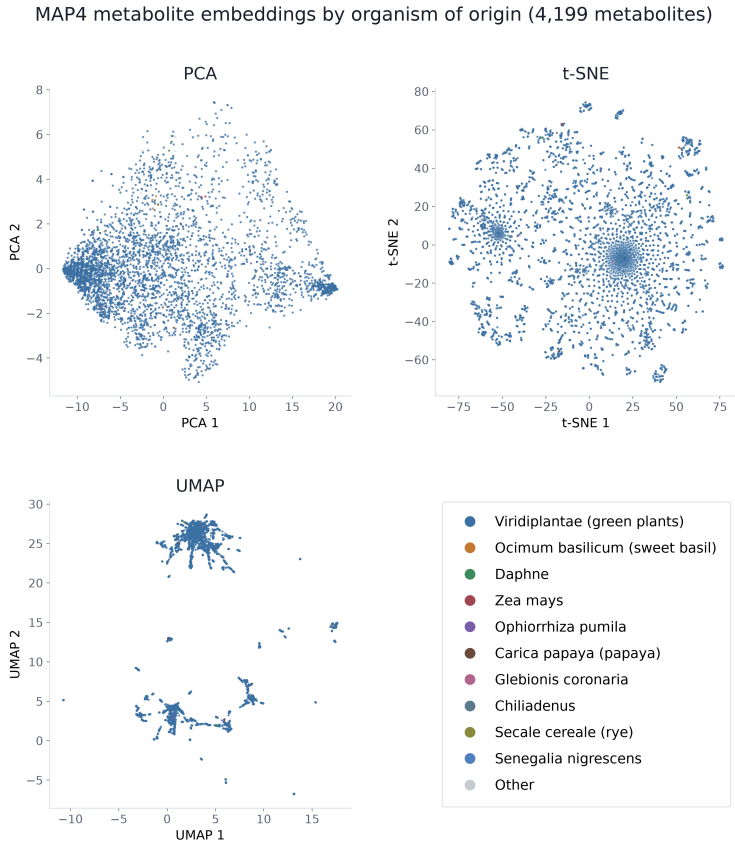

Figure 7: PCA, t-SNE, and UMAP projections of MAP4 metabolite embeddings (4,529 embedded nodes) coloured by organism of origin. The near-absence of species-level clustering shows that MAP4 fingerprints encode chemical structure, not taxonomic identity: the same metabolites appear across many plant species and are represented by identical or near-identical vectors regardless of organism context.

C.4 REACTION FINGERPRINTS (MAP4 DRFP-APPROXIMATION)

Motivation. The Differential Reaction FingerPrint (DRFP) (Probst et al., 2022) encodes a biochemical reaction by computing the symmetric difference of Morgan fingerprint atom environments between substrates and products, extracted from a single, pre-assembled reaction SMILES string (`reactants>>products`). Our graph does not provide reaction SMILES, but instead Conversion nodes are relational objects, connected to individually (partially) resolved substrate and product Metabolite nodes via graph edges. Computing literal DRFP would require reconstructing a valid combined SMILES from the participant compounds first, requiring every participant of a reaction to parse correctly. In our implementation, our per-metabolite MAP4 resolution is incomplete (Appendix C), with 8.3% of canonical metabolites unresolved, and a further subset resolved but containing R-group wildcards that MAP4 can embed but that cannot be reassembled into a clean reaction SMILES. A single failed participant would silently break conventional DRFP for the whole reaction. We therefore approximate DRFP's underlying idea that a reaction is characterised by what changes between substrates and products directly in the MAP4 embedding space. We implement mean-pooling each side's already-computed metabolite MAP4 fingerprints, reusing the same fingerprint space as the Metabolite node features (Appendix C). This avoids introducing a second hashing algorithm and toolchain, and better handles partial coverage.

Encoding. Biochemical reactions in the graph are represented by their substrates and products as Metabolite nodes. We approximate the Differential Reaction FingerPrint (DRFP) Probst et al. (2022) using the precomputed MAP4 embeddings: for a Conversion node r with substrate set S_r and product set P_r, the reaction fingerprint is:

$$\mathbf{f}_r = [\overline{\mathbf{m}}_{S_r} \parallel \overline{\mathbf{m}}_{P_r} \parallel \overline{\mathbf{m}}_{S_r} - \overline{\mathbf{m}}_{P_r}] \in \mathbb{R}^{3\times 1024}, \tag{1}$$

where $\overline{\mathbf{m}}_X = \frac{1}{|X|}\sum_{x\in X}\mathbf{m}_x$ is the mean MAP4 embedding of metabolite set X, and $\parallel$ denotes concatenation. This yields a **3,072-dimensional** reaction fingerprint. Of the 19,927 Conversion nodes, 13,717 have at least one resolved metabolite embedding and thus a non-zero fingerprint; the remainder receive a zero vector.

Fallback logic. When directional substrate/product edges are unavailable for a reaction, the following fallback is applied in order:

1. Both source and target MAP4 available: full 3,072-dim fingerprint (most reactions).
2. Only source or only target MAP4 available: partial fingerprint with zeros for the missing direction.
3. Only undirected `participants` edges have MAP4: participant mean used as substrate proxy; product and diff are zero.
4. No MAP4 coverage: all-zeros 3,072-dim vector.

Coverage. Of the 19,927 Conversion nodes, **13,717** (68.8%) have at least one MAP4-embedded metabolite and thus a non-zero fingerprint. The remaining 6,210 involve metabolites that could not be resolved to a SMILES string (PlantCyc-local compound IDs without public database cross-references).

Figure 8 visualises the MAP4 DRFP-approximation fingerprints for all Conversion nodes with at least one embedded metabolite. The t-SNE and UMAP projections show partial EC-class clustering (oxidoreductases and transferases form recognisable neighbourhoods) while substantial inter-class mixing remains. The reaction fingerprints therefore encode meaningful biochemical structure without collapsing to a simple EC-class lookup, validating their use as informative but non-leaking reaction features (cf. the EC shortcut, Appendix C.5).

Relation to metabolite graph structure. The DRFP approximation coexists in the model with the explicit metabolite graph structure: Metabolite nodes carry their own MAP4 features and are connected to Conversion nodes via `source`, `target`, and `participants` edges. The GNN therefore has two complementary sources of reaction chemistry. The DRFP approximation provides a fixed precomputed signal on the Conversion node itself, available from the first message-passing layer and invariant to graph depth or metabolite connectivity. The metabolite graph neighbourhood provides

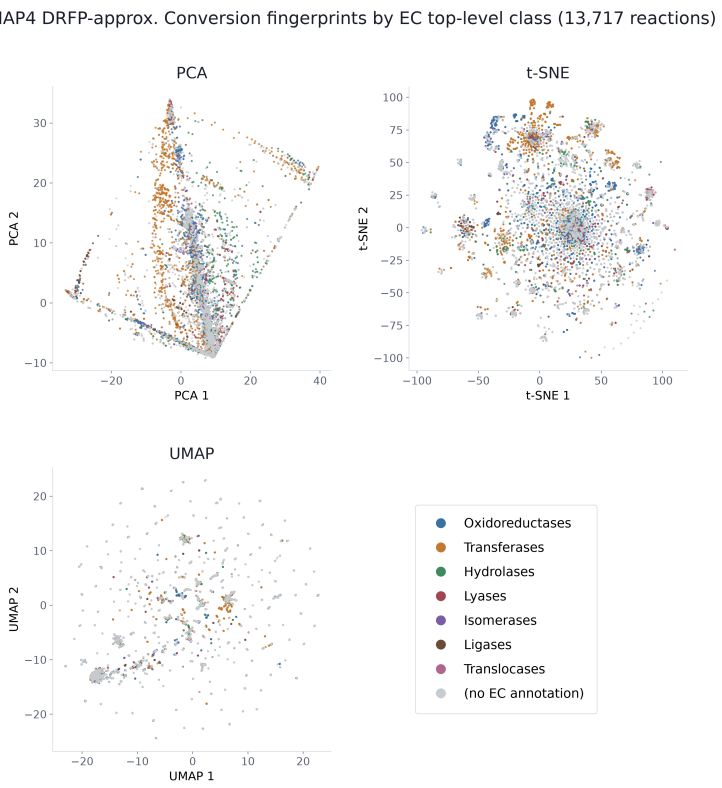

Figure 8: PCA, t-SNE, and UMAP projections of MAP4 DRFP-approximation fingerprints for the 13,717 Conversion nodes with at least one embedded metabolite, coloured by EC top-level class. Partial EC-class clustering is visible in the t-SNE and UMAP panels (oxidoreductases EC 1 and transferases EC 2 form local neighbourhoods) while inter-class mixing confirms the fingerprints are not reducible to an EC one-hot, supporting their use as non-leaking biochemical features.

a dynamic, learnable aggregation of the same chemical information at the individual substrate and product level, weighted by the GNN's attention and propagated across multiple hops. In the default configuration both channels are active simultaneously; the `--remove-all-metabolites` ablation flag removes the Metabolite nodes and their edges, leaving only the static DRFP approximation as the source of reaction-chemistry information.

C.5 EC Hierarchy Embeddings (Ablation Only)

Encoding. The Enzyme Commission (EC) number of a Conversion node encodes the reaction class at four levels (e.g., EC 2.1.1.280 = Transferase.Methyltransferase.C-methyltransferase.Caffeoylquinate-3-O-methyltransferase). We build a three-level one-hot encoding using the hierarchy of unique EC classes present in the dataset, as shown in Table 8. Note that each level encodes the observed classes in the dataset plus one "unknown" bin for Conversion nodes without an EC annotation at that level.

Table 8: EC hierarchy one-hot encoding structure (237-dim total).

Level	#Classes	#Dims (+ unknown)	EC classes
L1 (class)	7	8	Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases, Translocases
L2 (sub-class)	60	61	e.g. 2.1 (one-carbon transferases)
L3 (sub-sub-class)	167	168	e.g. 2.1.1 (methyltransferases)
Total	—	**237**	—

Coverage. EC annotations are parsed from `reactions.ttl` via `rdflib`. Approximately 55–60% of Conversion nodes have at least one EC annotation available from the curated source data. The remaining receive the "unknown" one-hot at all three levels. Table 9 shows the distribution of annotated Conversion nodes by EC top-level class.

Table 9: Distribution of Conversion nodes (the prediction target, $n = 19{,}927$) by EC top-level class. Unannotated Conversions receive the "unknown" one-hot vector at every hierarchy level.

EC	Name	# Conversions	%
EC 1	Oxidoreductases	5,418	27.2
EC 2	Transferases	4,893	24.6
EC 3	Hydrolases	1,303	6.5
EC 4	Lyases	1,064	5.3
EC 5	Isomerases	193	1.0
EC 6	Ligases	657	3.3
EC 7	Translocases	57	0.3
	Annotated	13,585	68.2
	Unknown / unannotated	6,342	31.8
	Total	**19,927**	100

Warning. EC features are **disabled by default** and provided only as an ablation flag (`--ec-features`). When enabled, the EC one-hot is appended to the Conversion node features (3,072 → 3,309 dims). Because EC class directly identifies the enzyme family responsible for a reaction, the model can memorise the mapping from EC class to training-organism proteins, achieving high validation performance without any cross-species generalisation. This shortcut causes an ≈80% relative drop from validation to test P-H@50 under the taxa holdout (Appendix F).

C.6 Organism Taxonomy Embeddings (Ablation Only)

Multi-hot lineage encoding. For each of the 424 Organism nodes, we retrieve the NCBI taxonomic lineage via the NCBI E-utilities API (`efetch`) in batches of 50 taxa. Named ranks only are retained: *superkingdom, domain, kingdom, subkingdom, phylum, subphylum, class, subclass, infraclass, order, suborder, family, subfamily, tribe, subtribe, genus, subgenus, species group, species*. Unnamed "clade" nodes in the NCBI lineage are discarded. This yields a binary multi-hot vector per organism of dimension equal to the number of unique named-rank ancestor taxon IDs across all 424 organisms. In the current dataset this vocabulary contains **702** unique ancestors, producing a 424×702 binary matrix. Two organisms sharing an order share all bits at and above the order level, naturally encoding phylogenetic relatedness.

MDS coordinates. We compute a pairwise taxonomic distance matrix $D \in \mathbb{R}^{424\times424}$ using named-rank hop distance:

$$d(A, B) = \mathrm{depth}(A) + \mathrm{depth}(B) - 2 \cdot \mathrm{depth}(\mathrm{LCA}(A, B)), \tag{2}$$

where $\mathrm{depth}(X)$ is the number of named-rank ancestors of organism X (not counting X itself), and $\mathrm{LCA}(A, B)$ is the last shared ancestor in the named-rank lineage sequences. Classical Multi-dimensional Scaling (MDS) is applied to D^2 to produce a **64-dimensional** Euclidean embedding that approximately preserves pairwise taxonomic distances. Formally: let $\mathbf{B} = -\frac{1}{2}\mathbf{H}D^2\mathbf{H}$ where $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix; MDS coordinates are the top-64 eigenvectors of $\mathbf{B}$ scaled by the square root of their eigenvalues.

Example distances. *Arabidopsis thaliana* (NCBITaxon_3702) and *Glycine max* (soybean, NCBITaxon_3847) are **10 named-rank hops** apart in the NCBI taxonomy. All 424 organisms are members of the *Viridiplantae* (green plants).

D SPLITTING POLICIES: DESIGN AND ANALYSIS

This section describes the four splitting policies supported in PlantMetBench: the taxa holdout (unseen species), the pathway holdout (unseen pathways), the *A. thaliana* reaction holdout (unseen reactions in *A. thaliana*), and the random-edge split. We first describe the default taxa holdout: its design, how organisms are assigned to splits, split statistics, how training supervision is handled, and embedding coverage across the split. We then describe the *A. thaliana* reaction holdout and compare all four splits. Finally, we analyse what the taxa holdout measures, using held-out-organism neighbourhood isolation and reaction overlap.

D.1 TAXA HOLDOUT DESIGN

The default data split in our design is a *taxa holdout* split: entire plant species are withheld from training and assigned exclusively to validation or test. This design aims to test cross-species generalisation: the model must predict enzyme–reaction associations for organisms it has never encountered, relying on conserved biochemistry encoded in protein and reaction representations, instead of memorising species-specific patterns. Part of this signal is nonetheless transductive rather than purely cross-species transfer, as we quantify directly in Appendix D.8. A random-edge split, by contrast, allows the model to look up a test reaction in the training graph via shared pathway membership or metabolite co-occurrence, making the task far easier than the deployment scenario.

Figure 9 shows the taxa holdout does not produce a trivial feature-space separation. Val and test proteins are interspersed throughout the training embedding cloud in ESM-C space. The model cannot distinguish held-out proteins by proximity alone, but it must transfer functional associations from graph context.

D.2 ORGANISM ASSIGNMENT CRITERIA

A *coverage constraint* is applied: an organism is eligible for validation or test only if at least 10% of its proteins appear in the ESM-C embedding pool. This prevents evaluation from being dominated by organisms for which the model has no feature information.

After filtering, organisms are partitioned into:

- **Train:** 322 organisms contributing the majority of positive edges.
- **Validation:** 11 held-out organisms.
- **Test:** 69 held-out organisms.

Figure 10 shows the distribution of per-organism coverage across all 424 organisms. The majority have fewer than 5% of proteins embedded, while a smaller group of well-characterised model organisms exceeds 10%, forming the right tail from which validation and test organisms are drawn.

Table 10: Taxa-holdout split statistics.

Split	# Organisms	# Positive edges
Train	322	9,569
Validation	11	1,211
Test	69	1,200
Organisms with qualifying edges	402	11,980
Total organisms in graph	424	—

D.3 SPLIT STATISTICS

Table 10 summarises the resulting organism and edge counts per split. The 22 organisms not assigned to any split ($424 - 402 = 22$) have no positive catalysis edges meeting the embedding-coverage threshold and are therefore absent from all supervision and evaluation sets, though their nodes remain in the message-passing graph and contribute neighbourhood context. A further three organisms have

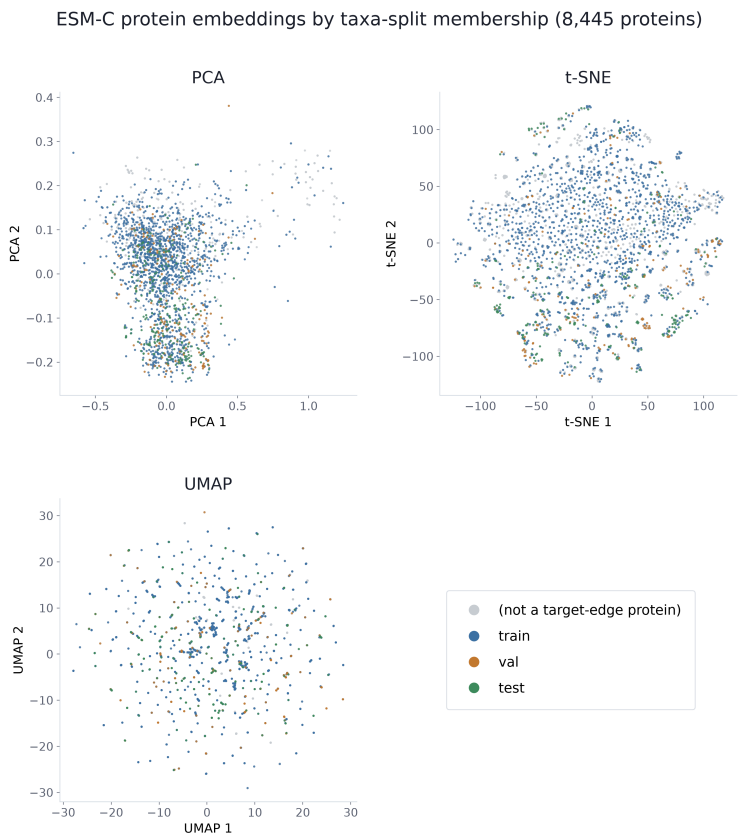

Figure 9: PCA, t-SNE, and UMAP projections of ESM-C protein embeddings coloured by taxa-split membership. Val and test proteins are geometrically interspersed with training proteins, confirming that the taxa holdout is not separable in embedding space, ruling out simple embedding-proximity lookup as a shortcut.

proteins present in the training edges despite not being part of the formal train-organism assignment, including one labelled only at the *Viridiplantae* kingdom level.

D.4 Disjoint Training Supervision

Within the training split, 20% of positive edges are designated *supervision-only* and used as labelled training pairs but are absent from the message-passing graph. The remaining 80% enter the MP graph and allow multi-hop message passing but are never used as supervision labels. This prevents the 1-hop training shortcut (Appendix F).

D.5 Embedding Coverage Across the Taxa Holdout

The taxa-holdout benchmark evaluates generalisation to *unseen* organisms, but the results in Table 1 conflate two sources of difficulty: (i) lack of ESM-C embeddings for the test proteins, and (ii) lack of training signal for unseen species. To disentangle these, we design a controlled experiment on *Arabidopsis thaliana* (NCBITaxon:3702), the one training organism whose proteins are almost completely covered by ESM-C (97.6%), and we construct a *within-species* reaction holdout that tests reaction-level generalisation while keeping the protein pool and organism context fixed.

Figure 11 illustrates the ESM-C coverage asymmetry across the 404 plant species. Training organisms are heavily skewed, with 189 of 324 species (58.3%) having *zero* ESM-C coverage and therefore contribute no protein-embedding signal during message passing. Validation and test organisms are substantially better covered (medians 68% and 86%, respectively), because the species for which UniProt sequences and ESM-C embeddings are available tend to be well-studied model organisms

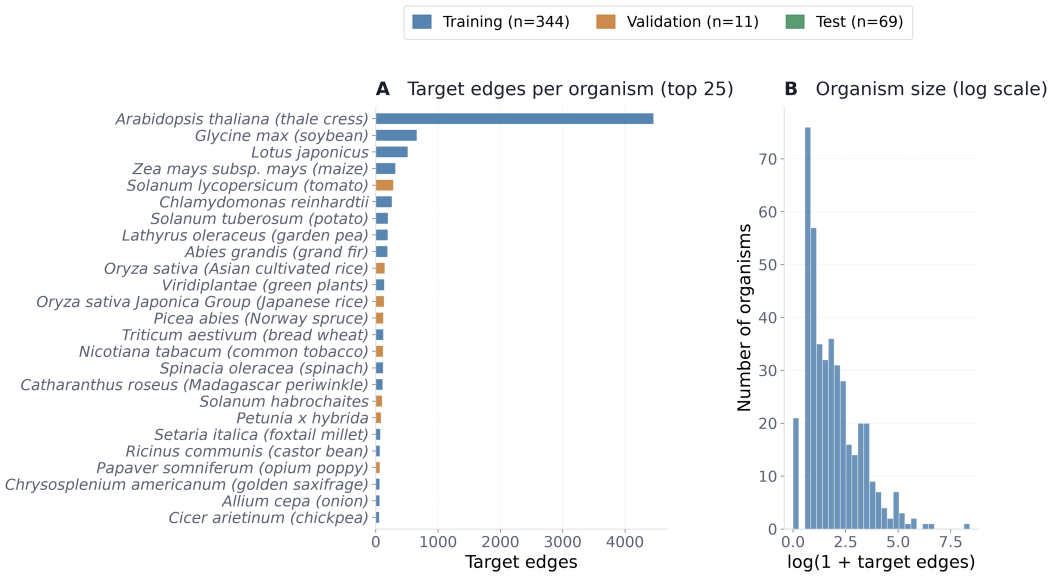

Figure 10: Distribution of (PROTEIN, `catalyzes`, CONVERSION) target edges across the 424 organisms in the taxa holdout, coloured by split membership (training, validation, test). **(A)** The 25 organisms contributing the most target edges, by scientific name. *Arabidopsis thaliana* alone contributes 4,459 edges (37% of the total), an order of magnitude more than the next-largest contributor, *Glycine max* (soybean, 661 edges). Validation organisms are disproportionately represented in this top-25 list (8 of 11 validation organisms, including tomato and both rice entries), while none of the 69 test organisms appear, indicating that test organisms, though eligible under the ESM-C coverage gate, are comparatively edge-sparse even relative to validation organisms. **(B)** Log-scale distribution of target-edge count across all 424 organisms, with 21 organisms contributing no target edges at all, and the remaining 403 spanning more than four orders of magnitude.

that were withheld for evaluation. Table 11 summarises these per-split figures alongside *A. thaliana*'s own 97.6% coverage, the motivation for choosing it as the within-species baseline below.

Table 11: Per-split ESM-C coverage summary across the 404 species. "Zero-coverage species" have no protein embedding in the pool.

Split	Species	Zero-cov.	Median cov.	Total proteins	ESM-C	Missing
Training	324	189 (58%)	0%	11,100	6,621	4,479
Validation	11	0	68%	1,260	826	434
Test	69	0	86%	1,337	1,002	335
A. thaliana	—	—	97.6%	5,421	5,289	132

D.6 *A. thaliana* REACTION HOLDOUT (EC-STRATIFIED)

Under the taxa holdout, all 4,459 *A. thaliana* catalysis edges fall in the training set. To construct an evaluation set with near-complete embedding coverage, we create a *within-species* split in which reactions are unseen at test time (unseen reactions in *A. thaliana*):

1. We collect all *A. thaliana* catalysis edges whose target reaction is a Conversion-subtype Interaction and whose protein is in the ESM-C pool (5,289 proteins at 97.6% coverage).
2. We group the resulting 2,271 unique reactions by EC class and apply a stratified 70/10/20 train/validation/test split, ensuring that no EC class appears exclusively in the test set.
3. The protein pool is restricted to the 5,289 *A. thaliana* ESM-C proteins, giving a random P-H@50 baseline of $50/5{,}289 = 0.94\%$.

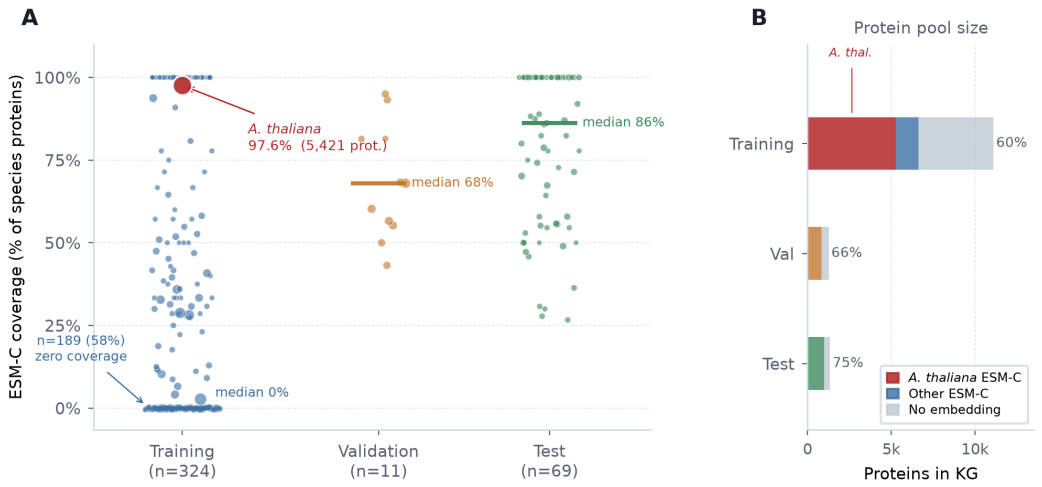

Figure 11: **ESM-C embedding coverage across the 404 species in PlantMetBench. (A)** Per-organism coverage; each circle represents one species, sized proportionally to $\sqrt{\text{protein count}}$. Training species are bimodal: 189 (58%) have zero coverage while others span the full range. Validation and test species are well-covered (medians 68% and 86%). *A. thaliana* (red) contributes 5,421 proteins at 97.6% coverage. Coloured bars indicate the per-split median. **(B)** Proportion of proteins per split with ESM-C embeddings, decomposed into *A. thaliana* ESM-C (red), other ESM-C (coloured), and proteins without an embedding (grey).

The resulting split sizes are shown in Table 12.

Table 12: Split sizes in the *A. thaliana* within-species reaction holdout. All reactions are Conversion-subtype; all proteins are ESM-C covered.

Split	Unique reactions	Catalysis edges
Training	1,589	3,079
Validation	208	402
Test	474	864
Total	2,271	4,345

D.7 Split Comparison

PlantMetBench provides four splits that probe different levels of generalisation difficulty. Table 13 summarises their statistics and Table 14 compares model performance for the random-edge split, the pathway holdout and the taxa holdout, which share the 8,445-protein ranking pool. The *A. thaliana* reaction holdout is not included in Table 14, since it uses a different, smaller protein pool and is not directly comparable on the same P-H@50 scale (see Appendix H.13). In the following paragraphs, we provide more details for each split policy.

Random-edge split. Edges are partitioned uniformly at random (80/10/10). Both proteins and reactions appear in training, validation, and test, so the model evaluates on nodes it has directly optimised over. This split is provided as a sanity-check upper bound. Strong performance confirms the model can fit the task, but does not imply cross-species or cross-pathway generalisation.

Pathway holdout (unseen pathways). Entire pathways are withheld for validation (47 pathways) and test (34 pathways), partitioned proportional to pathway size. Reactions from held-out pathways are never seen during training (Conversion Tanimoto = 0.00), but proteins *can* appear in both training and test via their participation in other pathways (Protein Tanimoto train–test = 0.13). This tests whether the model generalises to novel biochemical contexts (unseen reaction–pathway combinations) while retaining access to the protein embeddings it was trained on.

Table 13: Statistics of the four PlantMetBench splits. Tanimoto measures the node-set overlap between training and test. The taxa holdout leaves species unseen, the pathway holdout leaves pathways unseen, and the *A. thaliana* reaction holdout leaves reactions unseen within one species; the random-edge split holds out random edges.

Split	Train	Val	Test	Tanimoto (train ∩ test)	
	edges	edges	edges	Protein	Reaction
Random-edge split	9,584	1,198	1,198	0.22	0.16
Pathway holdout	9,612	1,191	1,177	0.13	0.00
Taxa holdout	9,569	1,211	1,200	0.00	0.10
A. thaliana reaction holdout	3,079	402	864	0.36	0.00

Taxa holdout (unseen species) – Default setting. Entire species are withheld for validation (11 organisms) and test (69 organisms), selected with a minimum embedding-coverage gate ($\geq$10% of the organism's proteins must have ESM-C embeddings) to avoid evaluating solely on zero-feature proteins. Protein Tanimoto between training and test is 0.00 under the taxa holdout, confirming complete protein disjointness (no protein seen during training appears in the test set). Reaction Tanimoto is not zero (train–test 0.10, train–val 0.13, val–test 0.15). Because the curated input data annotates many reactions across multiple species, some test-set reactions are also positive training examples via a different organism's catalyst. Appendix D.8 quantifies this directly and shows performance on reactions absent from training entirely. Because of the disjoint nature of proteins, the model must transfer enzyme–reaction associations to organisms it has never encountered, though the same appendix shows part of this transfer is transductive rather than purely cross-species. This is nonetheless the biologically meaningful setting for cross-species biosynthesis prediction and is the default evaluation split.

***A. thaliana* reaction holdout (unseen reactions in *A. thaliana*).** Reactions belonging to the one training organism with near-complete (97.6%) ESM-C coverage are withheld directly for validation (402 edges) and test (864 edges), stratified by EC class so that no class appears exclusively in test (Appendix D.6). **Unlike the taxa holdout, this split does not achieve protein disjointness by design.** Table 13 shows Reaction Tanimoto at exactly 0.00, matching the pathway holdout, but Protein Tanimoto at 0.36 (the highest of any split) because every candidate protein is intentionally kept visible across train, validation, and test; only the reactions, not the organism, are held out. This isolates reaction-level generalisation within a single, densely-characterised species from the harder cross-species transfer the taxa holdout targets, at the cost of testing a closed-pool retrieval problem (no held-out possible). Results are in Appendix H.13.

Table 14: P-H@50 under the three splits evaluated on the 8,445-protein pool (HeteroSAGE+Res&Jump, 2 layers, organism nodes removed, bidirectional message passing, mean $\pm$ std, 5 seeds, up to 650 epochs with early stopping). Difficulty increases from the random-edge split to the pathway holdout (unseen pathways) to the taxa holdout (unseen species) as measured by test performance.

Split	Val P-H@50	Test P-H@50	Val–Test gap
Random-edge split	71.0±2.5%	69.0±1.4%	+2.0 pp
Pathway holdout	73.6±0.9%	58.0±5.2%	+15.6 pp
Taxa holdout	49.9±3.4%	48.0±2.9%	+1.9 pp

The ordering confirms the intended difficulty gradient as measured by the test P-H@50: random-edge split (69.0%) > pathway holdout (58.0%) > taxa holdout (48.0%). The val–test gap in the random split shows a negligible gap (+2.0 pp, within noise), confirming that when train and test proteins are identical there is no meaningful distribution shift to overfit to. The pathway holdout produces the largest gap (+15.6 pp): the model has seen the proteins but must generalise to novel biochemical contexts, and validation pathways happen to be easier than test pathways. The taxa holdout gap is small (+1.9 pp) despite being the hardest split by absolute performance, indicating that whatever the model learns transfers comparably to both held-out sets rather than favouring validation organisms

specifically. For the main part of this manuscript and all subsequent experiments, *we therefore use the taxa holdout as the default splitting policy*.

D.8 Held-Out Organism Neighbourhood Isolation

The taxa holdout masks held-out organisms' `catalyzes` edges from message passing and supervision, but their Protein and GeneProduct nodes remain in the graph with every other edge type intact. We check two distinct questions this raises: (i) whether this allows information to leak from training organisms into held-out ones, and (ii) how much held-out performance depends on a held-out organism's own within-species graph context, independent of any leak.

No cross-organism leak. Organism nodes – the one node type that could route information between organisms directly – are removed entirely in the default configuration. Of the two remaining edge types connecting Protein and GeneProduct nodes to each other, (Protein, `is_about`, Protein) is 99.9% same-organism (12 of 9,932 edges cross an organism boundary) and (GeneProduct, `is_about`, GeneProduct) is 95.2% same-organism (219 of 4,580, raw graph); (GeneProduct, `encodes`, Protein) is same-organism by construction, since a gene encodes a protein in its own genome. No structural path therefore exists for a training organism's proteins to reach a held-out protein's embedding through message passing.

Isolation setup. We instead isolate a held-out organism's *own* neighbourhood – its other proteins and gene products, linked only to each other, never to training organisms. Using the existing best checkpoints (HeteroSAGE+Res&Jump, 5 seeds, no retraining), we construct a modified validation/test graph in which every edge of any type except the target (Protein, `catalyzes`, Conversion) edge is removed wherever a held-out organism's Protein or GeneProduct node appears as either endpoint. This isolates 2,596 Protein and 575 GeneProduct nodes (belonging to the 80 held-out organisms: 11 validation, 69 test) to their own 0-hop feature, propagated through each layer's self-transform with no neighbour aggregation, while every other node in the graph is untouched.

Table 15: P-H@K with held-out organisms' own neighbourhood intact (baseline) vs. isolated to their 0-hop feature only (HeteroSAGE+Res&Jump, taxa holdout, mean $\pm$ std over 5 seeds, no retraining). Baseline P-H@10/P-H@50 reproduce Table 1; P-H@1/P-H@5 baseline and all Isolated columns are measured directly on this experiment's own checkpoints.

	Baseline		Isolated	
Metric	**Val**	**Test**	**Val**	**Test**
P-H@1	0.8±0.8%	1.7±0.6%	0.0±0.0%	0.0±0.0%
P-H@5	11.8±1.3%	14.6±1.5%	0.7±0.2%	0.6±0.6%
P-H@10	21.6±3.7%	20.7±2.4%	3.4±1.2%	5.5±1.3%
P-H@50	49.9±3.4%	48.0±2.9%	32.5±2.9%	26.8±1.5%

Interpretation. Isolating a held-out organism's own neighbourhood costs 21.2 pp test P-H@50 (48.0% to 26.8%), a 44% relative drop, and collapses P-H@1 to exactly 0.0% in every seed on both splits. However, this is not evidence of a train/test leak, as the removed edges connect a held-out organism's proteins only to each other, never to training organisms (no label-adjacent information from training crosses the split boundary). Thus, a substantial share of taxa-holdout performance is transductive rather than purely cross-species transfer. The model benefits from seeing a held-out organism's other proteins and gene products alongside the one being scored, not from the query protein's own feature and conserved biochemistry alone. This is a capability under PlantMetBench's transductive design, but it needs to be accounted for when interpreting cross-species transfer and generalisation claims.

Reaction overlap. A held-out organism's own neighbourhood is not the only source of non-zero-shot signal: 64.7% of val edges and 62.9% of test edges target a Conversion node that already has a catalysing edge from a *different*, training-organism protein (Appendix D.7 quantifies this as a train–test reaction Tanimoto of 0.10, versus 0.00 for proteins). We stratify P-H@K by this property, using the same checkpoints and no retraining.

Bridging reactions are much easier than novel ones: test P-H@50 is 66.0% versus 12.6%, a 53.4 pp gap. This means a substantial share of taxa-holdout performance again reflects information already

Table 16: P-H@K stratified by whether the target reaction already has a training-organism catalyst (*bridging*) or not (*novel*) (HeteroSAGE+Res&Jump, taxa holdout, mean ± std over 5 seeds; same checkpoints as Table 15).

Metric	Bridging		Novel	
	Val	Test	Val	Test
P-H@1	1.1±1.1%	2.4±0.9%	0.2±0.2%	0.5±0.7%
P-H@5	16.2±2.0%	21.6±2.1%	2.1±1.4%	1.7±1.1%
P-H@10	31.3±3.8%	32.4±2.7%	4.3±1.2%	3.3±1.3%
P-H@50	63.3±2.9%	66.0±5.6%	18.7±5.4%	12.6±1.7%

present in the graph instead of testing true generalisation. Performance on the novel subset (12.6% test P-H@50, still ≈21× the 0.59% random baseline) is the more conservative estimate of the model's ability to generalise to biochemistry absent from training in any form.

E Cofactor / Currency Metabolite Removal

Currency metabolites (also called cofactors or hub metabolites) are small molecules that participate in a very large number of biochemical reactions (e.g., ATP participates in most energy-transfer reactions regardless of their specific chemistry). In the PlantMetWiki graph, hub metabolites appear as shared Metabolite nodes with extremely high degree, connected to hundreds of Conversion nodes. Their shared embeddings can therefore act as a *reaction-identity shortcut*: two reactions involving ATP share the same MAP4-embedded ATP neighbour, making it possible for the GNN to distinguish reactions simply by the presence or absence of cofactors, without learning the specific enzyme–substrate chemistry.

E.1 Currency Metabolite List

Table 17 shows 22 metabolite nodes, that are classified as currency metabolites and can be optionally excluded from the message-passing graph (`--remove-currency-metabolites`).

Note that in the default baseline configuration, currency metabolite removal is **disabled**. When enabled (`--remove-currency-metabolites`), edges connecting these metabolites to Interaction nodes are removed from the MP graph; the metabolite nodes themselves are retained. Alternatively, `--remove-all-metabolites` removes the entire Metabolite node store along with all Metabolite-touching edges, relying solely on conversion fingerprints for reaction-chemistry information.

F Shortcut Analysis

A *shortcut* is a structural pattern in the training graph from which a model can infer supervision labels without learning the task-relevant biochemical features. We identify four shortcuts in the raw PlantMetWiki graph and remove each in the default configuration. Table 18 summarises their effect on validation and test P-H@50.

F.1 Pathway Co-membership

PlantMetWiki encodes pathway membership via `(Interaction, is_part_of, Pathway)` and `(Protein, is_part_of, Pathway)` edges. If these edges are present in the MP graph, the model can infer that a protein p and a conversion c are likely to be associated simply because they co-occur in the same pathway node — without learning any enzyme–substrate compatibility. This is a near-trivial shortcut: two-thirds of positive pairs share at least one pathway.

We remove all `is_part_of` edges and the Pathway node store (2,478 nodes, 54,800 pathway-membership edges) from the MP graph. The pathway-membership relation is preserved in the raw `edges.tsv` file for use in hard-negative mining and post-hoc analysis.

Table 17: Currency metabolites excluded by `--remove-currency-metabolites`. Nodes are identified by their InChIKey suffix or ChEBI ID in the graph.

Metabolite	InChIKey / ChEBI ID	Role
ATP	GPRLSGONYQIRFK-UHFFFAOYSA-N	Energy carrier
ATP (alt form)	GVVPGTZRZFNKDS-JXMROGBWSA-K	Energy carrier
ADP	XPPKVPWEQAFLFU-UHFFFAOYSA-K	Energy carrier
H_2O	XLYOFNOQVPJJNP-UHFFFAOYSA-N	Solvent / hydrolysis
CO_2	CURLTUGMZLYLDI-UHFFFAOYSA-N	C1 carrier
NAD^+	XCCTYIAWTASOJW-XVFCMESISA-K	Redox carrier
NADH	RGJOEKWQDUBAIZ-IBOSZNHHSA-J	Redox carrier
$NADP^+$	XTWYTFMLZFPYCI-KQYNXXCUSA-K	Redox carrier
NADPH	ACFIXJIJDZMPPO-NNYOXOHSSA-J	Redox carrier
NADPH (ChEBI)	CHEBI:17499	Redox carrier
Phosphate (Pi)	NBIIXXVUZAFLBC-UHFFFAOYSA-L	Phosphate carrier
CoA-SH	BOPGDPNILDQYTO-NNYOXOHSSA-L	Acyl carrier
CoA (alt)	ZJUKTBDSGOFHSH-WFMPWKQPSA-N	Acyl carrier
CoA (ChEBI)	CHEBI:13392	Acyl carrier
Acetyl-CoA	VWFJDQUYCIWHTN-YFVJMOTDSA-K	Acyl carrier
Acetyl-CoA (ChEBI)	CHEBI:15339	Acyl carrier
FAD	UDMBCSSLTHHNCD-KQYNXXCUSA-L	Flavin carrier
$FADH_2$	XJLXINKUBYWONI-NNYOXOHSSA-K	Flavin carrier
SAM	WHUUTDBJXJRKMK-VKHMYHEASA-M	Methyl donor
SAH	HSCJRCZFDFQWRP-JZMIEXBBSA-L	Methyl donor product
H_2O_2	MHAJPDPJQMAIIY-UHFFFAOYSA-N	Reactive oxygen
NH_4^+	QGZKDVFQNNGYKY-UHFFFAOYSA-O	Nitrogen source

F.2 Reverse Catalysis Edge

Including `(Interaction, catalyzed_by, Protein)` in the MP graph creates a 2-hop shortcut: for a training pair (p, c), the model can traverse $p \xrightarrow{\texttt{catalyzes}} c \xrightarrow{\texttt{catalyzed_by}} p$ and learn to score this neighbourhood pattern, independent from biochemical compatibility. We remove `catalyzed_by` entirely from the MP graph; the forward `catalyzes` direction is never reversed in message passing, eliminating the 2-hop retrieval shortcut while preserving multi-hop biochemical context propagation.

F.3 1-Hop Training Supervision

If all positive training edges are in the MP graph, a 1-layer GNN can construct a node embedding for p that directly encodes its neighbours in c-space, making training label recovery trivial. The disjoint training ratio (Appendix D.4) prevents this by holding out 20% of training positives as supervision-only pairs, absent from the MP graph.

F.4 Enzyme Class (EC) Feature

The reaction EC number (e.g. EC 1.1.1.1) is a coarse functional label that directly identifies the enzyme class responsible for a reaction. Appending EC one-hot features to Conversion node features lets the model memorise the mapping from EC class to training-organism proteins rather than learning biochemical compatibility (Appendix C.5), causing a ≈80% relative drop from validation to test P-H@50 in the taxa holdout. EC features are therefore **disabled by default** and provided only as an ablation flag (`--ec-features`).

F.5 Shortcut results

Reverse catalysis edge is the most damaging shortcut, and the only one whose effect clearly exceeds seed-to-seed noise in both directions: re-introducing

Table 18: Effect of reintroducing individual training shortcuts on top of the organism-removed baseline (HeteroSAGE+Res&Jump, L2). All configurations share the same architecture and training protocol; only the graph construction differs per row. Mean ± std over 5 seeds (42, 0, 1, 2, 3); full baseline reproduces Table 1.

Configuration	Val P-H@50	Test P-H@50
Full baseline	$49.9\pm3.4\%$	$48.0\pm2.9\%$
+ Pathway co-membership edges (`--keep-pathways`)	$50.2\pm2.7\%$	$47.6\pm4.8\%$
+ Reverse catalysis edge (`--keep-catalyzed-by`)	$37.6\pm4.2\%$	$36.3\pm3.0\%$
+ No disjoint-train split (`--disjoint-train-ratio 0`)	$59.3\pm2.9\%$	$58.1\pm2.0\%$
+ EC-number Conversion features (`--ec-features`)	$49.0\pm3.9\%$	$47.7\pm3.2\%$

`(Conversion, catalyzed_by, Protein)` lowers val P-H@50 to $37.6\pm4.2\%$ and test P-H@50 to $36.3\pm3.0\%$, roughly 12 percentage points below the $49.9\pm3.4\%/48.0\pm2.9\%$ baseline on both splits. The GNN exploits the 2-hop path $p^* \to c^* \to p^*$ during training, learning a retrieval shortcut through the Conversion node itself and not through its substrate and product chemistry. Residual connections in the message-passing retain enough of the underlying biochemical signal, resulting in a partial degradation.

Pathway co-membership and EC one-hot features show no effect that exceeds baseline noise once averaged over 5 seeds. Pathway co-membership reaches val $50.2\pm2.7\%$ / test $47.6\pm4.8\%$, and EC features reach val $49.0\pm3.9\%$ / test $47.7\pm3.2\%$ – both configurations sit within one standard deviation of the baseline itself. Under a single seed (42), both had appeared to inflate validation by 1–2 pp; that apparent effect does not survive averaging over 5 seeds and was within-noise seed variance, not a real shortcut effect. We still exclude both from the default configuration on mechanistic grounds – pathway co-membership introduces a topological correlation unrelated to catalytic compatibility, and EC annotations are derived from the same catalytic labels that define the positive edges, so the feature could in principle let the model read class-level information directly from node features – but neither currently demonstrates a measurable shortcut effect at this seed count.

Disabling the disjoint split (ratio = 0) reuses training supervision edges as message-passing edges, raising val P-H@50 to $59.3\pm2.9\%$ (+9.4 pp) and test P-H@50 to $58.1\pm2.0\%$ (+10.1 pp) in near-equal measure – the second effect that clearly exceeds baseline noise. Because a large share of test reactions recur in training through a different organism's catalyst (Appendix D.8), the additional connectivity this shortcut introduces around shared reaction nodes that benefits both validation and test, instead of only inflating validation.

Taken together, only two of the four shortcuts show an effect that exceeds seed-to-seed noise at this sample size: reverse catalysis (damaging) and disabling the disjoint split (inflating), both moving validation and test in the same direction and by comparable magnitude. Pathway co-membership and EC features remain excluded from the default configuration on mechanistic grounds even though their measured effect here is not distinguishable from noise.

F.6 Organism Modality Ablation

PlantMetWiki contains 424 plant species, each connected to its proteins and genes via `(Protein, organism, Organism)` and `(GeneProduct, organism, Organism)` edges, with reverse edges routing messages back from Organism to both node types (Appendix B). A Protein therefore both contributes to, and receives from, its host species' aggregate representation during message passing. We ask whether this organism-level context helps or hurts generalisation to species absent from training, and whether the content of the Organism representation matters.

We test three conditions on HeteroSAGE+Res&Jump, 2 layers (Table 19):

1. **No organism nodes** (`--remove-organism-nodes`): all Organism nodes and their incident edges are removed. The model relies solely on biochemical information — reaction fingerprints, protein sequence embeddings, and metabolite graph structure.

2. **Random features**: Organism nodes are present with random 64-dimensional features. This tests whether species topology alone, without any phylogenetic content, influences generalisation.
3. **Taxonomy MDS embeddings** (`--organism-embeddings`): Organism nodes receive 64-dimensional coordinates from Classical Multidimensional Scaling (PCoA) on the pairwise taxonomic-hop distance matrix computed from NCBI lineages of all 424 species (construction details in Appendix C.6). This tests whether phylogenetic proximity provides a useful inductive bias for cross-species generalisation.

Two layers are the minimum depth at which organism context can round-trip back to a protein: a protein's contribution reaches its host Organism node at the first layer, and only returns to proteins of that organism (including others sharing the same host) at the second. HeteroSAGE+Res&Jump therefore isolates the effect of the Organism representation itself.

Table 19: Effect of organism node representation on HeteroSAGE+Res&Jump, 2 layers (mean ± std, 5 seeds, bidirectional message passing, taxa holdout). Removing organism nodes entirely yields the best test generalisation.

Organism representation	Val P-H@50	Test P-H@50
None (pure biochemistry)	$\mathbf{49.4 \pm 2.5\%}$	$\mathbf{48.1 \pm 4.1\%}$
Random features (bidirectional)	$38.3 \pm 4.1\%$	$41.9 \pm 2.8\%$
Taxonomy MDS, 64-dim (bidirectional)	$41.4 \pm 5.0\%$	$36.9 \pm 4.7\%$

Results. Including Organism nodes costs 6–11 percentage points of test P-H@50 relative to excluding them, regardless of what feature the node carries (Table 19). Taxonomy MDS coordinates, despite encoding genuine phylogenetic distance, do not outperform random features: the two conditions' confidence ranges overlap on both splits, and MDS trends lower on test. Because organisms are strictly disjoint across train, validation, and test, any adjustment a protein's representation picks up from its host Organism node reflects that node's position within the *training* species set. A held-out organism's Organism node occupies a comparable position structurally, but the model has no training signal tying that position to reaction outcomes for a species it has never seen, so the pathway contributes noise instead of a transferable prior. Supplying real taxonomic distance does not fix this: the model still has no example, during training, of using phylogenetic position to predict a novel species' enzyme–reaction associations, so it cannot be expected to extrapolate that usage correctly at test time. Excluding Organism nodes removes this pathway entirely and lets the model rely on protein and reaction features that transfer directly across species.

G Evaluation Metrics

We use three metrics. The primary metric directly measures the research question; the secondary metrics provide complementary diagnostic information.

G.1 Protein Hits at K (P-H@K)

Definition. Let $\mathcal{P}_{\text{emb}} \subset \mathcal{P}$ be the set of proteins with non-zero ESM-C embeddings (the *ranking pool*, $|\mathcal{P}_{\text{emb}}| = 8{,}445$). For each held-out positive edge $(p^*, c) \in \mathcal{E}_{\text{eval}}$ with $p^* \in \mathcal{P}_{\text{emb}}$, let

$$\text{rank}(p^*, c) = \left|\left\{q \in \mathcal{P}_{\text{emb}} : s(q, c) \geq s(p^*, c)\right\}\right|, \tag{3}$$

where $s(p, c) = \mathbf{z}_p \cdot \mathbf{z}_c$ is the dot-product score of the model embeddings. Protein Hits at K is then:

$$\text{P-H@}K = \frac{1}{|\mathcal{E}_{\text{eval}}|} \sum_{(p^*, c) \in \mathcal{E}_{\text{eval}}} \mathbf{1}\big[\text{rank}(p^*, c) \leq K\big]. \tag{4}$$

Design rationale. P-H@K follows the same full-pool, rank-based retrieval protocol adopted for structure identification from mass spectra in MassSpecGym (Bushuiev et al., 2024). Instead of reducing evaluation to a small set of sampled negatives, the true match is ranked against the entire candidate pool, so the metric reflects the actual deployment setting (retrieving the correct catalyst out of every protein the model could propose) beyond an easier pairwise discrimination task.

Ranking pool. Only proteins in $\mathcal{P}_{\text{emb}}$ are ranked. Positive edges whose true catalyst $p^* \notin \mathcal{P}_{\text{emb}}$ are excluded from evaluation and not counted as misses, since the model has no feature information about these proteins and their rank would reflect embedding coverage, not model quality.

Random baseline. Under a uniform-random ranking over $|\mathcal{P}_{\text{emb}}| = 8{,}445$ proteins, the expected P-H@K is $K/8{,}445$. For $K = 50$: random P-H@50 $\approx 0.59\%$.

Reported values. We report P-H@K for $K \in \{1, 5, 10, 50\}$. The primary metric is P-H@50; P-H@10 is a secondary more challenging ranking metric.

G.2 Any-Catalyst P-H@K (anyP-H@K) — Isoenzyme-Aware Variant

Many plant biochemical reactions are catalysed by multiple isoenzymes (proteins with the same function encoded by different genes). Standard P-H@K penalises the model if it ranks one correct isoenzyme in the top K but not the specific one in the evaluation edge. anyP-H@K removes this false penalty:

$$\text{anyP-H@}K = \frac{1}{|\mathcal{C}_{\text{eval}}|} \sum_{c \in \mathcal{C}_{\text{eval}}} \mathbf{1}\big[\min_{p^* \in \Pi(c)} \text{rank}(p^*, c) \leq K\big], \tag{5}$$

where $\mathcal{C}_{\text{eval}}$ is the set of unique Conversion nodes in the evaluation split and $\Pi(c)$ is the set of all known catalysts of reaction c across all splits. anyP-H@K is computed per unique reaction. Table 20 compares both metrics directly.

Catalyst multiplicity. Of the 4,572 unique reactions in the dataset, 2,357 (51.6%) have exactly one known catalyst and 2,215 (48.4%) have two or more, with a mean of 2.62 catalysts per reaction and a maximum of 61. anyP-H@K therefore differs from the standard metric for nearly half of all evaluation queries.

Table 20: Standard P-H@K vs. anyP-H@K (HeteroSAGE+Res&Jump, 2 layers, organism nodes removed, bidirectional message passing, taxa holdout). Mean ± std over 5 seeds (42, 0, 1, 2, 3); P-H@10/P-H@50 reproduce Table 1, all other cells measured directly. anyP counts any known catalyst across all splits as a positive.

	P-H@K		anyP-H@K	
K	Val	Test	Val	Test
1	0.8±0.8%	1.7±0.6%	28.8±3.0%	25.0±2.2%
5	11.8±1.3%	14.6±1.5%	41.3±2.5%	35.5±2.7%
10	21.6±3.7%	20.7±2.4%	45.0±2.6%	38.8±1.1%
50	49.9±3.4%	48.0±2.9%	56.0±2.4%	48.1±2.5%

Interpretation. anyP-H@K substantially exceeds the exact-match metric at low K: test P-H@1 rises from 1.7±0.6% to 25.0±2.2%. The model's top-ranked candidate is therefore usually a genuine catalyst of the reaction, but rarely the specific isoform present in the held-out organism. At $K = 50$ the two metrics nearly converge (48.0±2.9% versus 48.1±2.5% on test, closer than under the earlier single-seed estimate), so reactions for which any correct catalyst is retrievable within the top 50 almost always have their exact catalyst retrieved there as well. Isoenzyme ambiguity therefore accounts for most of the gap between the two metrics at $K = 1$ and little of it by $K = 50$. We report standard P-H@K as the primary metric throughout: once K reaches 50, anyP-H@K adds little beyond what standard P-H@K already captures.

G.3 Corrupt-Protein AUC and Average Precision (CP-AUC, CP-AP)

Definition. For each positive edge $(p^*, c) \in \mathcal{E}_{\text{eval}}$, we sample one *corrupt protein* $\tilde{p}$ uniformly from $\mathcal{P}_{\text{emb}} \setminus \{p^*\}$, ensuring $(\tilde{p}, c)$ is not a known positive. This yields a balanced set of positive and

negative pairs. We compute:

$$\text{CP-AUC} = \text{AUC}\big(\{(s(p^*, c), 1)\} \cup \{(s(\tilde{p}, c), 0)\}\big), \tag{6}$$

$$\text{CP-AP} = \text{AP}\big(\{(s(p^*, c), 1)\} \cup \{(s(\tilde{p}, c), 0)\}\big), \tag{7}$$

where AUC and AP are the standard area under the ROC curve and the average precision respectively.

Interpretation and Random Baselines. CP-AUC and CP-AP measure whether the model scores the true enzyme above a random protein *for the same reaction*. This is the research question in a binary, pairwise form. The random baseline is CP-AUC $= 0.50$, CP-AP $= 0.50$.

Corrupt-protein pool. Corrupt proteins are sampled from $\mathcal{P}_{\text{emb}}$ (not from all proteins) for the same reason as in P-H@K: proteins without embeddings are trivially easy negatives that inflate the metric without providing information about model quality.

G.4 Why Not Standard Link-Prediction AUC?

The standard link-prediction evaluation samples random (p, c) pairs as negatives, where both endpoints are chosen uniformly at random. In the PlantMetWiki graph, Pathway co-membership provides a near-trivial signal: two nodes drawn at random are very unlikely to share a pathway. A model that learns only pathway membership achieves AUC ≈ 0.97 without any biochemical generalisation, rendering this metric uninformative. We log the random-negative AUC for historical comparability only; it **must not be used as the primary optimisation target**.

H Extended Numerical evaluation

In the following, we extend the numerical evaluation analysis of Section 4 in the main manuscript. We start by providing more details regarding the setup (e.g., the models used, hyperparameters, etc.). Then we provide additional evaluation results.

H.1 GNN baselines

GNN encoders. We evaluate four heterogeneous GNN families. HeteroSAGE (Hamilton et al., 2017) applies one `SAGEConv` per edge type with mean aggregation and sums the outputs at each destination node. R-GCN (Schlichtkrull et al., 2018) uses a per-relation weight matrix but omits the self-loop/neighbourhood concatenation, while GAT (Veličković et al., 2018) and HGT (Hu et al., 2020b) replace mean aggregation with per-edge attention weights. For the attention-based models we apply gradient clipping (max norm 1.0) for training stability. All other hyperparameters are shared (Table 22).

Decoder. The default decoder is a parameter-free dot product: $\hat{y}(p, c) = \mathbf{z}_p \cdot \mathbf{z}_c$, where $\mathbf{z}_p, \mathbf{z}_c \in \mathbb{R}^d$ are the final-layer embeddings. An MLP decoder ($\hat{y} = \text{MLP}([\mathbf{z}_p \| \mathbf{z}_c])$, one hidden layer of dim d, ReLU) is also available as an ablation option.

H.2 HeteroSAGE with Residual & Jumping Knowledge

Encoder (residual variant). A second encoder, `HeteroSAGE+Res`, extends the baseline with residual connections, per-node-type LayerNorm, and multi-scale aggregation across layers, while keeping the same per-edge-type `SAGEConv` message function. Raw node features are first projected into a common d-dimensional space per node type $\tau(v)$:

$$\mathbf{h}_v^{(0)} = \text{LN}\big(\text{ReLU}\big(\mathbf{W}_{\text{in},\tau(v)}\, \mathbf{x}_v\big)\big)\,. \tag{8}$$

At layer ℓ, per-edge-type messages are computed as in the baseline, combined across relations by *mean*, and folded in as a residual update, not replacing the previous state:

$$\mathbf{h}_{v,r}^{(\ell+1)} = \mathbf{W}_r^{(\ell)} \left[\mathbf{h}_v^{(\ell)} \,\|\, \text{MEAN}_{u \in \mathcal{N}_r(v)}\, \mathbf{h}_u^{(\ell)}\right], \qquad \tilde{\mathbf{h}}_v^{(\ell+1)} = \text{MEAN}_r\, \mathbf{h}_{v,r}^{(\ell+1)}, \tag{9}$$

$$\mathbf{h}_v^{(\ell+1)} = \text{LN}\Big(\mathbf{h}_v^{(\ell)} + \text{Dropout}\Big(\tilde{\mathbf{h}}_v^{(\ell+1)}\Big)\Big)\,. \tag{10}$$

ReLU is applied between layers as before, but not after the final layer, so the last-layer embedding is unconstrained in sign for the dot-product decoder. The outputs of all L layers are combined via Jumping Knowledge Xu et al. (2018), concatenating each layer's state and projecting back to d dimensions:

$$\mathbf{z}_v = \mathbf{W}_{\mathrm{out},\tau(v)} \left[\mathbf{h}_v^{(1)} \,\|\, \mathbf{h}_v^{(2)} \,\|\, \cdots \,\|\, \mathbf{h}_v^{(L)} \right]. \tag{11}$$

This lets each node type draw on whichever hop-distance carries the most useful signal, rather than fixing all nodes to a single receptive field determined by L. The decoder (dot product or MLP ablation) is unchanged and operates on $\mathbf{z}_p, \mathbf{z}_c$ in place of the baseline's final-layer embeddings.

Parameter count. The residual variant adds a per-node-type input LayerNorm (negligible) and a per-node-type output projection $\mathbf{W}_{\mathrm{out},\tau(v)} \in \mathbb{R}^{d \times Ld}$ for the JK concatenation. For d=128, L=2, the total number of parameters is on the order of 1–2M parameters over the baseline's 4.2M, while the 1-layer model has approximately 1.3M parameters.

H.3 Dual-Encoder Retrieval

We evaluate the dual-encoder retrieval architecture that has become standard for enzyme–reaction association Mikhael et al. (2024); Zhang et al. (2026); Rocks et al. (2026) . Reactions and enzymes are encoded independently and projected into a shared embedding space; cosine similarity between a reaction and an enzyme is interpreted as the likelihood of catalysis. Reactions are represented by their DRFP fingerprints; enzymes by ESM-C sequence embeddings. Each representation is fed to a separate two-layer MLP (4,098 hidden dimensions) and L2-normalised. The model is trained with a full-batch maximum-likelihood NCE (MLNCE) objective, setting β =10 as the inverse temperature. Each epoch contrasts every training reaction against the entire enzyme pool via a global partition function, admitting multiple catalysing enzymes per reaction without a one-positive-per-query assumption. At inference, it ranks by cosine similarity to the query.

This design follows the strongest published model Rocks et al. (2026), which uses a composite of ECFP6 and DRFP fingerprints paired with ProtT5 embeddings. We substitute DRFP alone and ESM-C to ensure the only variable relative to the GNN baselines is the architecture (no graph structure used), not the molecular representations. The dual encoder serves as a non-graph reference point: underperformance relative to the GNN baselines would indicate that graph context provides a meaningful inductive bias for cross-species enzyme function prediction.

H.4 MIX model

As an alternative architecture, we propose MIX, a heterogeneous GNN that combines GraphSAGE and graph attention (GAT) convolutions within a single message-passing layer. The two operators aggregate neighbourhood information differently. GraphSAGE uses mean aggregation, so every neighbour contributes equally to a node's updated representation. GAT learns attention coefficients over the incoming edges, so the model can put more weight on the most informative neighbours and less on the rest.

MIX assigns one operator to each edge type, based on the node types it connects. We split node types into primary types, the two node types of the link being predicted (Conversion and Protein), and secondary types, all other node types. Edge types with at least one primary endpoint use `SAGEConv`. Edge types that connect only secondary node types use `GATConv`. Neighbours linked directly to a target node are assumed to be relevant to the prediction, so aggregating them uniformly preserves their signal and avoids the variance of learned attention. Relations among secondary node types, by contrast, are more numerous and more heterogeneous in how much they matter for the prediction, so attention lets the model select useful context and down-weight noisy neighbours. In each layer, the messages from the different edge types that arrive at the same node type are combined by summation.

H.5 BLASTp Sequence-Retrieval Baseline

We implement a reaction-conditioned sequence-retrieval baseline following the protocol of ReactZyme Hua et al. (2024). For each query reaction c, we define a score for any candidate protein p:

$$\mathrm{score}(p, c) = \max_{t \in \mathcal{T}(c)} \mathrm{bitscore}(p, t), \tag{12}$$

where $\mathcal{T}(c)$ is the set of training catalysts of c and $\mathrm{bitscore}(p, t)$ is the BLASTp bitscore between the sequences of p and t (0 if no hit passes an E-value threshold of 10^{-3}). Proteins are ranked by this score; ties are broken randomly (seed 42).

Database. We run BLAST+ (Camacho et al., 2009)'s `makeblastdb` on the 1,458 training catalyst sequences to build a BLAST database. We then run `blastp` with `-evalue 1e-3` and output format `"6 qseqid sseqid bitscore"`, querying all 8,445 ranked-pool proteins against this database. The search yields 165,317 high-confidence hits.

H.6 Structure-Only Baseline

To measure how much of the model's performance is attributable to graph topology alone, independent of the precomputed node features, we train the best-performing architecture (HeteroSAGE+Res&Jump, L2) after replacing every node type's input features with independent Gaussian noise of the same dimensionality (`torch.randn`, seeded identically to the run's training seed). Graph structure (all edges), the training protocol, negative sampling, and the 8,445-protein evaluation pool used by every other model in Table 1 and Table 23 are left unchanged, so the comparison isolates topology's own contribution under an otherwise identical setup.

Results. Across 5 seeds, the structure-only model reaches $2.1\pm0.3\%$ test P-H@50 ($2.4\pm0.3\%$ validation), roughly $3.6\times$ the 0.59% random baseline (Table 1) but far below the $48.0\pm2.9\%$ reached with real features. Graph connectivity is therefore mildly informative on its own, with proteins linked through shared reaction neighbourhoods are somewhat more predictable than chance. The large majority of the model's performance therefore depends on the biochemical content of the ESM-C, PlantCaduceus, and MAP4 features.

H.7 Training Objective and Negative Sampling

The model is trained with binary cross-entropy loss over sampled positive and negative edge pairs:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{(p,c,y)\in\mathcal{B}} \big[y \log \sigma(\hat{y}) + (1-y)\log(1-\sigma(\hat{y}))\big], \tag{13}$$

where σ is the sigmoid function and $\mathcal{B}$ is a mini-batch of labelled edges.

Negative sampling. For each positive training edge (p^*, c^*), we sample:

- k_C *random-conversion* negatives (p^*, c_{rand}): same protein, random Conversion from the full 19,927-node pool. These train the model to distinguish the true reaction above random reactions.
- k_P *random-protein* (corrupt-protein) negatives (p_{rand}, c^*): random protein, same conversion. These train the model in the protein-ranking direction — the research question.

Both negative types exclude known positive pairs.

We set $k_C = k_P = 5$ by default. Table 21 reports val and test P-H@50 for $k \in \{1, 3, 5, 10, 20\}$ on a single seed. Val P-H@50 declines gradually and monotonically across the full range tested (53.3% at $k = 1$ to 50.3% at $k = 20$), a 3.0 pp spread. Test P-H@50 does not track this decline: it varies within a narrow band (44.7–46.6%) with no consistent direction, and its **highest** value in the sweep occurs at $k = 20$. CP-AUC is more sensitive to k than P-H@50 on test, falling from 0.934 at $k = 3$ to 0.920 at $k = 20$, consistent with the pairwise discrimination task becoming harder as negatives increasingly outnumber positives within a batch. We adopt $k = 5$ as a default that sits within the stable region on both metrics, without the higher variance-per-batch cost of larger k.

H.8 Hyperparameters

Table 22 lists the full hyperparameter configuration used for all GNN experiments; dual encoder hyperparameters are described in Appendix H.11.

Table 21: Effect of negative sampling ratio k (both k_C and k_P set equally) on HeteroSAGE+Res&Jump, 2 layers (single seed 42, organism nodes removed, bidirectional message passing).

k	Val P-H@50	Test P-H@50	Val CP-AUC	Test CP-AUC
1	53.3%	45.9%	0.947	0.932
3	52.0%	46.0%	0.948	0.934
5	51.3%	45.5%	0.950	0.932
10	51.1%	44.7%	0.949	0.927
20	50.3%	46.6%	0.947	0.920

Table 22: Full hyperparameter configuration for the PlantMetBench baseline. GAT and HGT use the same settings but with gradient clipping (max norm 1.0).

Group	Hyperparameter	Value
Architecture	GNN type	HeteroSAGE (`SAGEConv`, `aggr=sum`)
	Hidden dimension d	128
	Number of layers	2
	Decoder	Dot product
	Dropout	0.3 (between layers only)
	Lazy init	Yes (SAGEConv $(-1, -1)$)
Training	Optimiser	AdamW
	Learning rate	1×10^{-4}
	Weight decay	1×10^{-5}
	Max epochs	200
	Early stop patience	50 (by val P-H@50)
	Neg. ratio k_C	5
	Neg. ratio k_P	5
Data	Disjoint train ratio	0.2
	Pathway edges	Removed
	Reverse catalysis	Removed
	EC features	Off
	Currency metabolites	Kept (not removed)
	Protein features	ESM-C, 960-dim
	Reaction features	MAP4 DRFP-approx, 3,072-dim
	Metabolite features	MAP4, 1,024-dim
	Gene features	PlantCaduceus, 1,024-dim
	Organism nodes	Removed
	Random seed	42, 0, 1, 2, 3

H.9 RESULTS

Table 23 provides the full multi-model results and HeteroSAGE depth comparison. The main paper (Table 1) reports only the best-depth entry per architecture; the 2-layer and 3-layer HeteroSAGE variants here quantify the effect of message-passing depth, which is not otherwise covered in the main text. All GNN results use organism nodes excluded; see §F.6 for the organism ablation.

BLASTp Sequence-Retrieval. On the taxa holdout test split, BLASTp achieves P-H@10 $= 10.17\%$ and P-H@50 $= 32.44\%$, outperforming GAT (11.6%) but trailing R-GCN (39.6%), HeteroSAGE (42.0%), and HGT (35.5%). CP-AUC and CP-AP are not reported since BLASTp does not produce a protein-space embedding. These results establish sequence homology as a strong non-trivial baseline that a pure architecture-based signal must clear, and motivate the need for reaction-conditioned multimodal representations to achieve further improvements.

Attention-based GNNs. GAT is sharply depth-dependent: at L1 (1.5±0.9%) and L3 (1.6±1.2%) it stays near the 0.59% random baseline, but at L2 it reaches 11.6±3.6% test P-H@50 (CP-AUC

Table 23: Full results on PlantMetBench (organism nodes excluded for all GNNs, mean ± std, 5 seeds: 42, 0, 1, 2, 3, taxa holdout). P-H@K: fraction of reactions where the true catalyst ranks in the top K of 8,445 proteins. CP-AUC/AP: corrupt-protein AUC and average precision.

Model	Split	P-H@1 (%)	P-H@5 (%)	P-H@10 (%)	P-H@50 (%)	CP-AUC	CP-AP
Random	—	0.01	0.06	0.12	0.59	0.500	0.500
Structure-only (Res&Jump L2)	Val	0.0±0.1	0.1±0.1	0.3±0.2	2.4±0.3	0.70±0.02	0.66±0.02
	Test	0.0±0.1	0.2±0.1	0.4±0.2	2.1±0.3	0.70±0.03	0.67±0.02
BLASTp	Test	0.00	2.03	10.17	32.44	—	—
HeteroSAGE (L1)	Val	0.4±0.4	3.9±1.3	11.0±2.4	40.5±1.3	**0.952±0.004**	0.946±0.006
	Test	2.1±0.7	10.9±1.6	17.5±2.2	42.0±5.0	**0.948±0.001**	0.943±0.002
HeteroSAGE (L2)	Val	0.4±0.3	7.0±2.0	15.1±4.2	44.1±4.7	0.947±0.003	0.941±0.004
	Test	1.1±0.7	12.1±1.5	17.9±1.3	41.9±6.1	0.947±0.003	0.941±0.003
HeteroSAGE (L3)	Val	0.3±0.3	4.8±2.0	11.9±3.3	40.0±3.0	0.941±0.006	0.932±0.006
	Test	1.4±0.9	10.2±1.4	15.6±2.2	40.3±5.1	0.943±0.003	0.936±0.002
GAT (L1)	Val	0.1±0.2	0.4±0.4	0.6±0.4	3.8±1.9	0.865±0.013	0.788±0.020
	Test	0.2±0.3	0.2±0.4	0.3±0.5	1.5±0.9	0.861±0.017	0.766±0.041
GAT (L2)	Val	0.2±0.4	1.3±0.6	3.5±1.2	21.9±3.2	0.902±0.009	0.875±0.015
	Test	0.1±0.1	1.0±0.7	1.6±1.0	11.6±3.6	0.901±0.008	0.871±0.013
GAT (L3)	Val	0.1±0.2	0.4±0.1	0.5±0.1	5.5±4.2	0.875±0.008	0.807±0.013
	Test	0.1±0.2	0.1±0.2	0.2±0.3	1.6±1.2	0.872±0.009	0.798±0.018
MIX (L1)	Val	0.6±0.7	4.4±1.6	9.9±2.4	31.4±3.7	0.940±0.003	0.928±0.004
	Test	1.1±0.8	5.5±2.1	10.4±3.7	32.5±6.8	0.942±0.005	0.932±0.006
MIX (L2)	Val	0.3±0.6	3.2±1.3	7.6±1.5	34.6±2.2	0.918±0.010	0.908±0.012
	Test	0.8±0.9	4.2±1.8	7.4±2.0	29.4±5.2	0.913±0.009	0.903±0.012
MIX (L3)	Val	0.1±0.2	2.5±0.9	6.2±2.2	32.7±2.8	0.888±0.037	0.887±0.029
	Test	0.4±0.3	2.1±0.8	4.9±1.	25.2±3.6	0.888±0.022	0.885±0.017
R-GCN (L1)	Val	0.0±0.1	4.7±1.8	11.5±1.6	37.7±2.1	0.931±0.005	0.923±0.006
	Test	0.7±0.4	8.0±2.4	15.5±2.4	39.6±4.5	0.932±0.002	0.925±0.002
R-GCN (L2)	Val	0.7±0.3	4.4±1.0	9.1±1.7	33.7±2.3	0.938±0.003	0.927±0.004
	Test	0.8±0.6	7.7±1.7	12.2±2.0	30.1±2.3	0.940±0.005	0.929±0.005
R-GCN (L3)	Val	0.2±0.2	3.0±1.8	6.5±1.4	28.9±2.2	0.925±0.004	0.910±0.005
	Test	0.5±0.7	5.2±1.9	9.3±1.5	23.7±2.7	0.926±0.002	0.911±0.004
HGT (L1)	Val	0.4±0.5	5.8±2.9	12.3±1.7	36.3±3.2	0.905±0.009	0.896±0.011
	Test	1.2±1.1	6.8±2.8	12.2±2.4	35.5±5.5	0.904±0.006	0.894±0.006
HGT (L2)	Val	0.2±0.2	2.5±1.4	6.6±1.4	30.2±1.7	0.915±0.006	0.897±0.008
	Test	0.6±0.6	4.9±1.0	8.8±1.7	27.6±3.5	0.916±0.002	0.897±0.002
HGT (L3)	Val	0.1±0.1	1.2±0.5	3.9±1.0	22.9±3.9	0.903±0.012	0.875±0.022
	Test	0.2±0.3	2.1±1.4	3.7±1.9	15.6±4.7	0.903±0.010	0.871±0.020
HeteroSAGE+Res (L1)	Val	0.0±0.1	1.9±1.3	9.4±3.0	34.6±3.3	0.930±0.003	0.918±0.005
	Test	1.0±0.4	5.8±1.9	10.7±2.7	34.0±5.1	0.929±0.002	0.918±0.003
HeteroSAGE+Res (L2)	Val	0.1±0.1	2.0±1.6	5.9±4.3	28.4±4.0	0.915±0.005	0.899±0.008
	Test	0.4±0.4	5.2±2.5	8.9±2.8	26.0±4.4	0.915±0.003	0.897±0.006
HeteroSAGE+Res (L3)	Val	0.2±0.4	1.1±1.6	3.1±3.2	16.6±9.7	0.888±0.018	0.845±0.051
	Test	0.4±0.7	1.8±2.4	2.7±2.8	10.7±6.1	0.890±0.014	0.847±0.046
HeteroSAGE+Res&Jump (L1)	Val	0.6±0.5	**12.4±2.8**	**25.1±3.9**	49.1±3.9	0.926±0.009	0.931±0.005
	Test	1.6±0.9	13.6±1.0	21.3±2.0	47.6±3.2	0.919±0.009	0.926±0.008
HeteroSAGE+Res&Jump (L2)	Val	0.5±0.2	10.5±2.7	21.6±3.7	**49.9±3.4**	0.943±0.009	0.944±0.006
	Test	1.7±0.6	**13.9±1.8**	20.7±2.4	**48.0±2.9**	0.937±0.006	0.940±0.005
HeteroSAGE+Res&Jump (L3)	Val	**0.7±0.6**	11.6±3.4	21.5±3.7	48.8±2.9	0.946±0.009	0.945±0.009
	Test	**1.6±0.8**	13.8±2.6	**21.4±2.2**	47.5±3.2	0.940±0.007	0.941±0.006
Dual encoder (L1)	Val	0.0±0.0	0.8±0.5	4.6±1.3	27.3±3.4	0.915±0.003	**0.957±0.002**
	Test	0.0±0.0	1.0±0.5	5.3±0.7	29.6±1.8	0.932±0.010	0.966±0.005
Dual encoder (L2)	Val	0.0±0.1	0.6±0.5	4.1±1.2	24.2±3.1	0.914±0.007	0.956±0.004
	Test	0.0±0.0	1.5±0.5	6.8±1.2	29.3±1.4	0.929±0.003	0.964±0.001
Dual encoder (L3)	Val	0.0±0.0	0.9±0.6	5.6±0.8	27.4±2.0	0.915±0.011	0.957±0.006
	Test	0.0±0.0	1.5±0.7	6.2±1.2	29.7±1.4	0.936±0.006	**0.968±0.003**

0.901±0.008), an order-of-magnitude jump over its other two depths. HGT is more consistent, peaking at L1 (35.5±5.5%, CP-AUC 0.904±0.006) and declining monotonically with depth (27.6% at L2, 15.6% at L3). R-GCN performs comparably to HeteroSAGE, reaching 39.6±4.5% test P-H@50 at L1 (CP-AUC 0.932±0.002), also declining with depth.

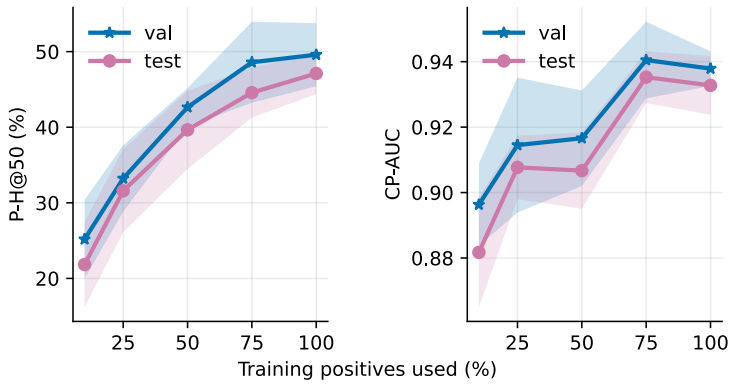

Figure 12: Effect of training-data fraction on link-prediction performance, using the best-performing architecture (HeteroSAGE+Res&Jump, L2, 128-dim hidden, bidirectional message passing). The model is trained on increasing fractions of the training positive edges (10–100%) and evaluated on held-out validation and test sets across P-H@50 and CP-AUC (mean $\pm$ std over 5 seeds; shaded bands show ± 1 std).

Dual encoder. The dual encoder (L1, 29.6±1.8% test P-H@50, CP-AUC 0.932±0.010, CP-AP 0.966±0.005) sits below every architecture's best-performing depth except GAT, showing that independently encoding reaction and enzyme representations with domain-adapted embeddings and MLNCE training does not by itself match heterogeneous message passing's ability to exploit graph structure under the taxa holdout. CP-AP tracks CP-AUC closely, confirming the model reliably ranks the true catalyst above a random negative once evaluated with the same one-negative protocol used for the GNN baselines (Appendix G.3).

H.10 Model saturation discussion

Figure 12 presents how the mean test accuracy of the P-H@50 score changes as we alter the fraction of training positives. This setting evaluates whether the model learning process is saturated for the proposed model (HeteroSAGE+Res&Jump - L2). Both metrics are still rising at 100% of the available training data, indicating the model has not yet saturated and would likely benefit from additional optimisations.

H.11 Results on the Deduplicated 2,232-Protein Pool

Pool-size background. The 8,445 proteins in the main ranking pool include many with identical amino-acid sequences (and therefore identical ESM-C embeddings), arising from the same enzyme occurring across multiple plant species. After collapsing all proteins that share an identical ESM-C vector, 2,232 distinct representatives remain. The random P-H@50 baseline rises from $50/8{,}445 \approx 0.59\%$ to $50/2{,}232 \approx 2.24\%$, and any model whose score depends only on ESM-C embeddings—including both the dual encoder and BLASTp under tie-breaking, cannot distinguish proteins within the same duplicate group. Following the Horizyn dual encoder evaluation protocol (Rocks et al., 2026), we collapse each group to its representative before ranking, so that identical-sequence proteins do not compete for the same rank. We evaluate the dual encoder (all three depths) and BLASTp on this pool. GNN architectures use graph neighbourhood features that distinguish same-sequence proteins in different species; re-evaluating them on the 2,232 pool is left to future work.

Discussion. On the 2,232-protein pool (Table 24), the dual encoder (L1: 64.5% test P-H@50) outperforms BLASTp (52.7%) by nearly 12 percentage points, demonstrating that reaction-conditioned encoding provides a meaningful gain beyond sequence retrieval even when both methods rank the same set of distinct protein sequences. BLASTp achieves 0% P-H@1 on this pool: no reaction has its correct catalyst ranked first by sequence similarity alone. By contrast, the dual encoder achieves 0.3% P-H@1 at L1, rising to 1.5% at L3. This improvement points to DRFP occasionally providing enough specificity to rank the true catalyst at the top of the pool.

Table 24: Results on the **2,232 deduplicated protein pool** (taxa holdout; random P-H@50 ≈ 2.24%). Dual encoder: mean ± std over 5 seeds. BLASTp: deterministic (test split only). CP-AUC/AP not computed for BLASTp on this pool (marked —).

Model	Split	P-H@1	P-H@5	P-H@10	P-H@50	CP-AUC	CP-AP
Random	—	0.04%	0.22%	0.45%	2.24%	0.500	0.500
BLASTp	Test	0.00%	26.23%	37.79%	52.68%	—	—
Dual encoder (L1)	Val	0.5±0.4%	17.6±0.9%	36.3±1.2%	65.4±0.9%	0.902±0.006	0.951±0.003
	Test	0.3±0.2%	23.2±0.8%	36.0±0.4%	**64.5±0.7%**	0.914±0.003	0.957±0.002
Dual encoder (L2)	Val	1.1±0.7%	17.0±1.5%	33.7±1.2%	63.7±0.4%	0.902±0.006	0.951±0.003
	Test	1.3±0.8%	24.3±2.1%	34.7±1.6%	62.2±1.1%	0.912±0.009	0.956±0.004
Dual encoder (L3)	Val	1.5±1.0%	16.9±2.0%	35.1±2.3%	61.9±1.2%	0.895±0.007	0.948±0.004
	Test	1.5±0.6%	22.2±0.6%	32.7±1.7%	61.3±1.4%	0.911±0.011	0.955±0.006

Depth hurts on the deduplicated pool as it does on the full pool: L1 (64.5%) > L2 (62.2%) > L3 (61.3%), consistent with the 8,445-protein results (Table 23). This points to the performance decay with depth being a genuine overfitting tendency of deeper encoders on this benchmark, not an artefact of the larger pool.

Comparing the two pool sizes, the dual encoder gains $64.5 - 29.6 = 34.9$ percentage points by switching from the 8,445 to the 2,232 pool, while BLASTp gains $52.7 - 32.4 = 20.3$ points. The larger gain for the dual encoder is consistent with it benefiting more from collapsing duplicate sequences, likely because MLNCE training is distracted by proteins that have identical embeddings but different sequence labels in the full pool.

H.12 Diagnostics

In this section, we examine how model performance evolves throughout training and how the models use the different input modalities. We first track validation P-H@50 and CP-AUC across training and graph depth (Fig. 13), highlighting a substantial difference in how the two metrics distinguish model performance. We then use input-feature attribution to investigate which node modalities contribute most strongly to the model's predictions and how this attribution changes with depth. Together, these analyses provide complementary views of model behavior. The training curves characterize optimization and predictive performance, while the attribution analysis reveals which inputs the models respond to.

CP-AUC does not separate models. Figure 13 shows validation P-H@50 and CP-AUC over training at each depth. P-H@50 ranges from GAT around 3% to Residual+JK Sage around 49%, while CP-AUC converges to 0.86-0.95 for every model at every depth. GAT reaches a CP-AUC of roughly 0.89 while ranking the true catalyst in the top 50 for around 3% of reactions. Scoring the true catalyst above a single random protein is therefore a weak test. The curves also show that Residual+JK SAGE converges fastest and is the only model whose accuracy holds across depth.

Input-feature attribution. To identify which modality the model responds to, we attribute the loss to each node's input features using gradient × input (Bach et al., 2015), which is equivalent up to a scaling factor to layerwise relevance propagation (Shrikumar et al., 2017). For node i of type t with input $x_i \in \mathbb{R}^{(d_t)}$ and gradient $g_{id} = \partial\mathcal{L}/\partial x_{id}$, the attribution is

$$a_i = \sum_{d=1}^{d_t} g_{id}\, x_{id}, \tag{14}$$

computed in evaluation mode over the supervised edges. We prefer this to the gradient norm for two reasons. It is invariant to feature scale and dimension, so modalities whose embeddings change from 960 to 3,072 dimensions stay comparable, and it is a first-order estimate of the change in loss from setting a node's features to zero, which is the state unresolved nodes are in in our graph. We restrict the average to valid nodes, $V_t = \{i : x_i \neq 0 \text{ and } g_i \neq 0\}$, and report the share of total attribution per type, $A_t = \sum_{i \in V_t} |a_i|$, together with the mean per-node attribution $\bar{a}_t = |V_t|^{-1} \sum_{i \in V_t} |a_i|$. For the

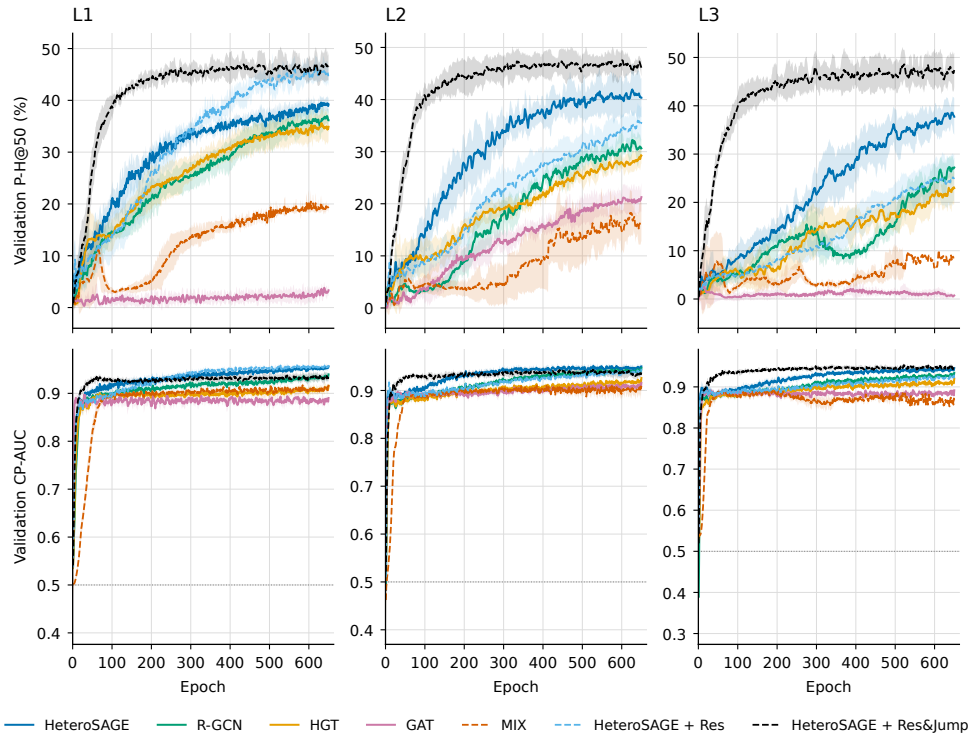

Figure 13: Validation P-H@50 (top) and CP-AUC (bottom) over training, for all models at one to three layers (mean over 5 seeds, shaded band = ± one standard deviation). The two metrics separate the models very differently. P-H@50 spans nearly the full range from 0 to 48%, while CP-AUC converges to 0.88–0.94 for every model at every depth.

dimension-level view, we instead sum $|g_{id}x_{id}|$ over nodes for each d, taking the absolute value before summing so that contributions cannot cancel across dimensions. Two caveats apply. The measure is local, so a feature with small attribution at convergence may still be informative (Shrikumar et al., 2017), and attribution is shaped by graph directionality, since metabolite features reach only the reaction side of the protein-reaction score and never the protein representations being ranked. We therefore read these values as indicating which inputs the model responds to rather than how much it depends on a modality.

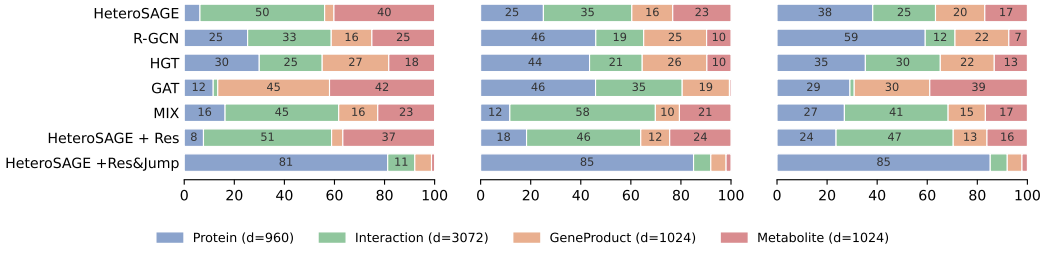

Figure 14: Share of input-feature attribution per node type, for one, two, and three layers (left to right), averaged over 5 seeds. Bars show each modality's percentage of $\sum_i |a_i|$, so they sum to 100% within a panel.

Attribution shifts toward Protein with depth. Figure 14 shows how input-feature attribution is distributed across modalities. Most models move attribution toward Protein as layers are added, with R-GCN rising from 30% to 59% and MIX from 8% to 27% between one and three layers.

HeteroSAGE+Res&Jump is the exception and the strongest model, placing over 80% of attribution on Protein at every depth, consistent with jumping-knowledge connections preserving direct access to the input features, and it is also the only model whose accuracy is unchanged across depth (47.5–48.0% test P-H@50).

H.13 *A. thaliana* Reaction Holdout Results

Table 25 reports performance on the *A. thaliana* split for HeteroSAGE+Res&Jump, GAT and HGT (5 seeds each), the dual encoder (5 seeds) and BLASTp. The random baseline is $50/5{,}289 = 0.94\%$. We report both P-H@50 and anyP-H@50; the distinction is explained below.

Table 25: Performance on the *A. thaliana* reaction holdout vs. the standard taxa holdout (HeteroSAGE+Res&Jump, 2 layers, dot decoder). Mean ± std across 5 seeds unless noted. P-H@50 averages over all test edges (864); anyP-H@50 averages over unique test reactions (454), counting a hit if any known catalyst for that reaction appears in the top-50. CP-AUC is the corrupt-protein AUC (see Appendix G.3). All rows share the same 5,289-protein A. thaliana ranking pool, except the taxa-holdout reference row (8,445-protein pool), whose P-H@50 and CP-AUC reproduce Table 1.

Model	Pool	Val P-H@50	Test P-H@50	Test anyP-H@50	Test CP-AUC
Random baseline	5,289	0.94%	0.94%	0.94%	50.0%
A. thaliana reaction holdout					
HeteroSAGE+Res&Jump	5,289	55.7±0.8%	58.9±1.7%	50.4±1.2%	93.0±0.7%
GAT	5,289	0.0±0.0%	0.0±0.0%	0.0±0.0%	28.6±2.7%
HGT	5,289	3.5±1.0%	2.6±1.0%	3.9±0.9%	86.2±1.4%
Dual encoder	5,289	68.6±0.9%	64.6±0.3%	72.6±0.4%	93.7±0.3%
BLASTp[†]	5,289	5.97%	4.75%	7.27%	—
Taxa holdout (reference)					
HeteroSAGE+Res&Jump	8,445	49.9±3.4%	48.0±2.9%	—	93.7±0.6%

[†] BLASTp score is the maximum bitscore to a training catalyst sharing the query reaction's top-level EC class (the split's own stratification key). Scoring by exact-reaction identity is not applicable here since the split holds out reactions themselves, guaranteeing zero train/test reaction overlap.

Table 26: BLASTp and dual-encoder performance on the deduplicated *A. thaliana* pool (1,451 candidates: the 5,289 A. thaliana ESM-C proteins collapsed to their distinct embedding vectors, following the dual encoder's default pool construction). Reported for reference alongside Table 25's like-for-like 5,289-protein comparison. Dual encoder: mean ± std over 5 seeds; BLASTp is deterministic.

Model	Val P-H@50	Test P-H@50	Test anyP-H@50	Test CP-AUC
Random baseline	3.45%	3.45%	3.45%	50.0%
Dual encoder	75.7±0.5%	80.5±0.4%	77.4±0.5%	93.7±0.5%
BLASTp[†]	21.14%	20.83%	22.25%	—

Deduplicated pool. Table 26 repeats this comparison on the deduplicated 1,451-candidate *A. thaliana* pool: the dual encoder reaches 80.5% test P-H@50 against BLASTp's 20.83%, the same qualitative gap seen on the full 8,445-protein taxa-holdout pool (Table 24).

P-H@50 vs. anyP-H@50. The standard P-H@50 averages over all 864 test catalysis edges, meaning reactions with multiple known catalysts (isoenzymes) are counted once per edge. Since 35.5% of test reactions have more than one known catalyst (max 24, mean 1.9), well-characterised multi-enzyme reactions are up-weighted relative to rare single-catalyst reactions. anyP-H@50 corrects for this by averaging over the 454 *unique* test reactions, counting a hit if any known catalyst for that reaction is ranked in the top-50. HeteroSAGE+Res&Jump's lower anyP-H@50 (50.4% vs. 58.9% P-H@50) reflects performance concentrated on a subset of common, multi-isoenzyme reactions that P-H@50 counts multiple times. The dual encoder shows the opposite pattern (anyP-H@50 72.6% vs. P-H@50 64.6%). Its hits are spread more evenly across distinct reactions and not concentrated on a few highly

promiscuous ones. We therefore consider anyP-H@50 the more informative per-reaction metric for the *A. thaliana* reaction holdout and recommend it for future multi-catalyst benchmarks.

HeteroSAGE+Res&Jump. anyP-H@50 rises from $\approx 0.94\%$ (random) to 50.4% on the *A. thaliana* reaction holdout, a **54× improvement** over random. Test P-H@50 (58.9%) is only 1.23× the taxa-holdout reference (48.0%), a substantially narrower gap than embedding coverage alone would predict: residual connections and jumping-knowledge aggregation recover most of the headroom that missing ESM-C coverage and unseen-species generalisation cost a shallower architecture. The task remains hard: the model correctly identifies at least one catalyst for only 50.4% of distinct test reactions, and 18.8% of pathways are not reconstructed at all (Appendix I.2).

Dual encoder. The dual encoder reaches 64.6% test P-H@50 on the *A. thaliana* split, outperforming HeteroSAGE+Res&Jump (58.9%) by 5.7 percentage points, a reversal of the taxa holdout on the full dataset, where the same dual encoder tops out at 29.6% against HeteroSAGE+Res&Jump's 48.0% (Table 23). The two splits stress opposite capabilities. Under the taxa holdout, catalysts in the test organism are never seen attached to any reaction during training, so a model with no graph structure to fall back on has no way to place their embeddings correctly; HeteroSAGE+Res&Jump's message passing supplies that missing context through homology and pathway edges shared across species. Under the *A. thaliana* reaction holdout, every candidate protein is already densely represented in training, only the reactions are held out, so the task reduces to learning a chemistry-to-catalyst mapping for a fixed, already-observed pool, which is exactly what a pure similarity encoder is built for and does not require cross-species structural transfer.

HGT: discrimination without ranking. HGT achieves 86.2% CP-AUC, close to HeteroSAGE+Res&Jump (93.0%), yet its test P-H@50 is only 2.6%, just 2.8× above random. This dissociation reveals a qualitatively distinct failure mode: HGT learns to discriminate the true catalyst from a random protein (high CP-AUC), but fails to rank it tightly enough to appear in the top-50 of the full 5,289-protein pool. The anyP-H@50 (3.9%) exceeds P-H@50 (2.6%) for HGT, the opposite direction from HeteroSAGE+Res&Jump: this reflects that HGT's marginal hits tend to fall on reactions with a single known catalyst, which are counted equally under anyP-H@50 but down-weighted under P-H@50 relative to multi-catalyst reactions. Inspection of training curves shows CP-AUC remaining below 0.50 (worse than random) until epoch 75–80, then rising steeply; P-H@50 never leaves 0% throughout training, with early stopping selecting very late checkpoints (epoch $\approx$170). This behaviour suggests HGT's attention mechanism learns coarse catalytic compatibility but cannot resolve fine-grained ranking under the heterogeneous graph topology and default hyperparameters.

GAT failure (5 seeds). GAT achieves *zero* P-H@50 and anyP-H@50 across all 5 seeds and all 200 training epochs, with early stopping selecting epoch 1 in every seed. The mean test CP-AUC (28.6%) is well below the 50% random baseline, confirming the model actively mis-ranks catalysts instead of just failing to learn. This represents an additional distinct failure mode (rank inversion) distinct from both HeteroSAGE+Res&Jump (learns and ranks) and HGT (learns discrimination but not ranking). Potential causes include attention-weight instability across the 16 diverse edge types and a more complex loss landscape than the mean-aggregation in HeteroSAGE.

CP-AUC as a training diagnostic. The *A. thaliana* reaction holdout further illustrates the diagnostic value of CP-AUC (see Appendix G.3). HeteroSAGE+Res&Jump reaches $93.0 \pm 0.7\%$ CP-AUC on the *A. thaliana* reaction holdout, closely matching the $93.7 \pm 0.6\%$ reached under the taxa holdout: unlike P-H@50, the model's ability to discriminate a true catalyst from a random protein no longer depends strongly on organism coverage, even though ranking it tightly enough within the full pool still does. The GAT failure is immediately visible in CP-AUC terms ($28.6 \pm 2.7\%$, well below the 50% random baseline), while HGT's $86.2 \pm 1.4\%$ CP-AUC correctly signals that the model has learned meaningful signal, just not enough to resolve full-pool ranking. P-H@50 alone cannot distinguish these three regimes.

I Pathway Reconstruction

I.1 Pathway Reconstruction Analysis (Taxa Holdout)

P-H@K aggregates over individual reactions, but the biologically meaningful unit is a *pathway*: a coordinated sequence of reactions. We ask how many pathways HeteroSAGE+Res&Jump can

reconstruct from the standard cross-species taxa-holdout test set (5 seeds, 8,445-protein pool), where the target organism itself, not just the individual reaction, is entirely unseen during training. For each pathway containing at least one evaluable test reaction (true catalyst in the ESM-C pool), we compute the *pathway reconstruction rate*: the fraction of its test reactions for which the model ranks the true catalyst in the top-50. A pathway is *fully reconstructed* if all its test reactions are hits in every seed; *not reconstructed* if none are, in any seed. Results are summarised in Table 27 and Figure 15.

Table 27: Pathway reconstruction at top-50 across the 267 pathways with at least one evaluable test reaction under the taxa holdout (HeteroSAGE+Res&Jump, 5 seeds). "Fully reconstructed" requires all test reactions to be hits in all seeds.

Category	Pathways	%
Fully reconstructed (100%)	25	9.4%
≥75% of test reactions correct	63	23.6%
≥50%	110	41.2%
≥25%	157	58.8%
Not reconstructed (0%)	61	22.8%
Mean reconstruction rate	41.6%	
Median reconstruction rate	40.0%	

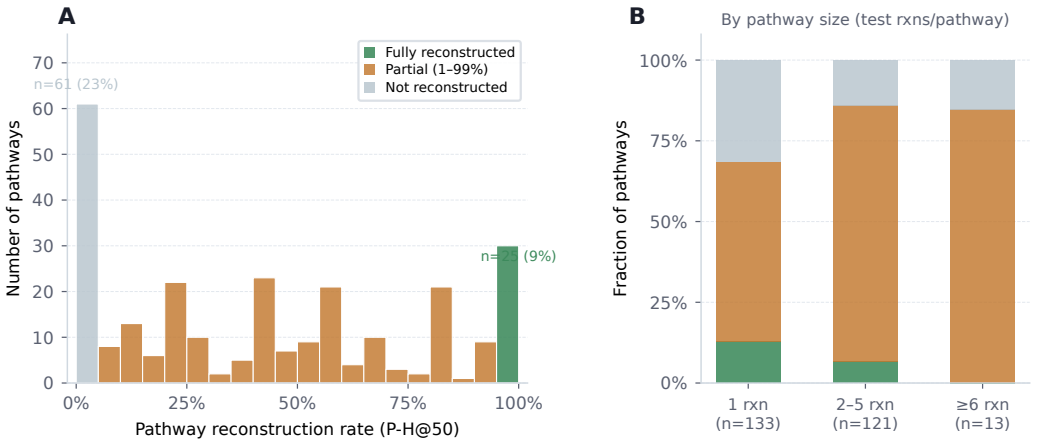

Figure 15: **Pathway reconstruction at top-50 under the taxa holdout (HeteroSAGE+Res&Jump, 5 seeds). (A)** Distribution of per-pathway reconstruction rates across 267 pathways with ≥1 evaluable test reaction. 23% of pathways are entirely missed (0%), while 9% are fully reconstructed (100%); the remainder are partially reconstructed. **(B)** Breakdown by the number of test reactions per pathway. Pathways with multiple test reactions are substantially more likely to be at least partially reconstructed (86% for the 2–5 reaction group and 85% for ≥6, vs. 68% for single-reaction pathways).

Comparison to the *A. thaliana* reaction holdout. Full reconstruction is substantially rarer under cross-species generalisation than within-species generalisation (9.4% vs. 13.7%), consistent with unseen-organism catalysts being the harder generalisation regime. Mean reconstruction rates are comparable across splits (41.6% vs. 44.9%), and the medians are now identical (40.0% both). The not-reconstructed rate, however, is *higher* under the taxa holdout (22.8% vs. 18.8%). This seems to conflict with pathway size alone, as the taxa holdout's test pathways carry more test reactions per pathway on average (median 2, mean 2.1, vs. median 1 for *A. thaliana*) This buffering effect does not translate into a lower not-reconstructed rate, suggesting that cross-species transfer failures are more often total-pathway failures, rather than a simple consequence of test-pathway size. The model most often misses every reaction in a pathway rather than a mix of hits and misses. This may point to the importance of pathway context in the biosynthesis graph.

Methodological note on pathway multiplicity. The 25 fully reconstructed pathways draw on 35 distinct reactions, and as with the *A. thaliana* reaction holdout, none of the 35 recurs in more than one fully reconstructed pathway.

What makes a pathway fully reconstructable? Two patterns emerge, both in line with the corresponding *A. thaliana* finding. **(i) Very high isoenzyme count.** The 35 reactions underlying fully reconstructed pathways have a median of 6 known catalysts (mean 7.3, max 61), far above the median of 3 (mean 4.9) across all 563 evaluable test reactions. Under cross-species generalisation specifically, a reaction with many known catalysts across the training organisms is more likely to already include a close homologue of the true, unseen-organism catalyst, giving the model a usable sequence-similarity anchor through the ESM-C embedding space even though the exact protein was never seen during training. **(ii) EC 2 (transferases) reconstructs disproportionately well.** 13 of 147 EC 2 reactions (8.8%) are fully reconstructed, against 1 of 112 for EC 1 (0.9%) and zero for EC 3, EC 4, and EC 5 despite non-trivial representation (12, 33, and 10 reactions respectively). EC 2.4.1.91, a glycosyltransferase code, alone contributes 7 independent fully reconstructed reactions (the largest single cluster) suggesting this enzyme family's sequence signature is unusually well captured by ESM-C.

Interpretation. The isoenzyme and EC-class patterns found here reinforce, and sharpen, the *A. thaliana* result: reconstruction succeeds where many biochemically-confirmed catalysts already anchor the sequence-to-function mapping, and specific enzyme families (glycosyltransferases in particular) reconstruct disproportionately well under both splits. That the median catalyst count needed for full reconstruction is substantially higher here (6, vs. 2 under the *A. thaliana* reaction holdout) is consistent with cross-species transfer requiring a denser web of homologous catalysts to succeed, since the target organism itself is never seen during training.

I.2 Pathway Reconstruction Analysis (*A. thaliana* Reaction Holdout)

P-H@K aggregates over individual reactions, but the biologically meaningful unit is a *pathway*: a coordinated sequence of reactions. We therefore ask how many pathways (or fractions thereof) the model can reconstruct from the test set.

Setup. The PlantMetBench knowledge graph encodes 2,476 PlantMetWiki pathways (median 5 reactions each, max 200) via `Conversion` $\xrightarrow{\texttt{is_part_of}}$ `Pathway` edges. For each pathway containing at least one evaluable test reaction (true catalyst in the ESM-C pool), we compute the *pathway reconstruction rate*: the fraction of its test reactions for which HeteroSAGE+Res&Jump ranks the true catalyst in the top-50. A pathway is *fully reconstructed* if all its test reactions are correct across all 5 seeds; *not reconstructed* if none are correct in any seed. Results are summarised in Table 28 and Figure 16.

Table 28: Pathway reconstruction at top-50 across the 335 pathways with at least one evaluable test reaction (HeteroSAGE+Res&Jump, 5 seeds). "Fully reconstructed" requires all test reactions to be hits in all seeds.

Category	Pathways	%
Fully reconstructed (100%)	46	13.7%
$\geq$75% of test reactions correct	89	26.6%
$\geq$50%	149	44.5%
$\geq$25%	207	61.8%
Not reconstructed (0%)	63	18.8%
Mean reconstruction rate	44.9%	
Median reconstruction rate	40.0%	

Methodological note on pathway multiplicity. A single reaction node can belong to multiple pathways via `is_part_of` edges, so the 46 fully reconstructed pathways could in principle share a much smaller set of underlying reactions. In practice they draw on 59 distinct reactions, and none of the 59 recurs in more than one fully reconstructed pathway: the fully reconstructed set is not being inflated by one well-predicted reaction propagating across several of its pathway memberships.

What makes a pathway fully reconstructable? Inspection of the 59 reactions underlying the 46 fully reconstructed pathways points to two patterns, one of which reverses the naive expectation. **(i) Higher, not lower, isoenzyme count.** These reactions have a median of 2 known *A. thaliana* catalysts (mean 2.53, max 6), compared with a median of 1 (mean 1.93, max 24) across all 454 evaluable

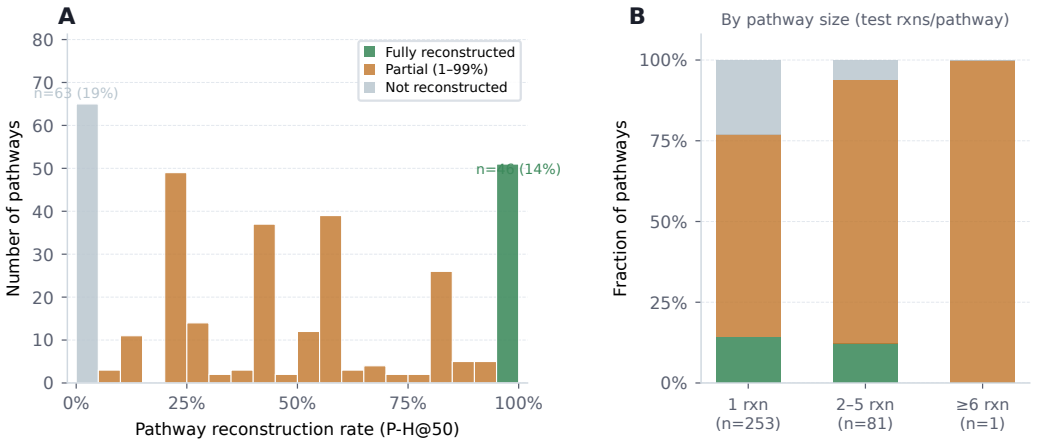

Figure 16: **Pathway reconstruction at top-50 (HeteroSAGE+Res&Jump, 5 seeds). (A)** Distribution of per-pathway reconstruction rates across 335 pathways with ≥1 evaluable test reaction. 19% of pathways are entirely missed (0%), while 14% are fully reconstructed (100%); the remainder are partially reconstructed. **(B)** Breakdown by the number of test reactions per pathway. Pathways with multiple test reactions are more likely to be at least partially reconstructed (94% for pathways with ≥2 test reactions vs. 77% for single-reaction pathways).

test reactions. Reactions with more annotated catalysts appear easier to reconstruct: additional isoenzymes give the model more than one biochemically valid sequence consistent with the reaction, widening the set of proteins whose embedding the top-50 cutoff can land on. **(ii) One enzyme class stands out sharply.** EC 6 (ligases) reactions are reconstructed at a far higher rate than any other class: 10 of the 15 EC 6 reactions in the test set are fully reconstructed (67%), driven largely by EC 6.3 (10 of 10 fully reconstructed reactions with that code). EC 1 (4%), EC 2 (8%), EC 4 (12%), and EC 5 (18%) reconstruct at much lower, comparable rates, and EC 3 (hydrolases, 0 of 32) and EC 7 (translocases, 0 of 1) contribute no fully reconstructed reactions at all. Diacylglycerol kinase (EC 2.7.1.151) contributes 3 independent fully reconstructed reactions and EC 2.4.1.91 contributes 2, consistent with distinctive sequence families being easier for ESM-C to separate, but neither dominates the set the way EC 6 does. EC 2.4.1.91 is the largest single cluster under the taxa holdout (Appendix I.1, 7 reactions), suggesting this glycosyltransferase family's sequence signature is captured reasonably well by ESM-C across the benchmark as a whole, though the signal is stronger under the taxa holdout than within *A. thaliana* alone.

Interpretation. Pathway-level reconstruction is currently limited primarily by pathway granularity: 75% of pathways contribute only a single test reaction (a consequence of the EC-stratified split), which makes their reconstruction binary (hit-or-miss) and sensitive to per-reaction variance across seeds. Isoenzyme multiplicity is not a sizeable limiting factor, as reactions with more known catalysts reconstruct *more* reliably. The harder cases are concentrated among single-catalyst, single-test-reaction pathways outside EC 6. A pathway-holdout split (holding out entire pathways) would be a more direct probe of pathway reconstruction capability and is a natural direction for future work.

I.3 Rhea-Based Independent Validation (Taxa Holdout)

PlantMetWiki's GPML pathway files embed Rhea cross-references for a subset of reactions (Bansal et al., 2022), an expert-curated, independently maintained biochemical reaction database whose enzyme annotations derive from UniProt (The UniProt Consortium, 2023) SwissProt evidence, curated independently of PlantCyc. We ask whether proteins that HeteroSAGE+Res&Jump ranks highly on the taxa holdout (5 seeds, 8,445-protein pool), but that PlantMetWiki does *not* label as the true catalyst, are nonetheless supported by Rhea, providing evidence that the benchmark undercounts true positives even when the target organism itself is entirely unseen during training. The underlying Rhea/UniProt cross-reference coverage is a property of the whole knowledge graph: 591 reactions graph-wide have ≥1 Rhea-annotated protein among the 1,788 protein nodes with a UniProt

accession. What differs by split is how many of those reactions fall in the evaluated val/test set with a pool-covered Rhea-annotated protein: 203 (val) and 132 (test) reactions here.

Result. HeteroSAGE+Res&Jump reaches Rhea anyP-H@50 of $0.99\pm0.92\%$ (val) and $2.42\pm1.64\%$ (test) across 5 seeds, against a random baseline of 0.59% for the 8,445-protein pool, roughly $4\times$ random on test, and statistically indistinguishable from random on val (two of five seeds scored exactly 0%).

Interpretation. The under-counted-positives signal survives cross-species generalisation on test, but is far weaker and noisier than under the *A. thaliana* split ($4.1\times$ vs. $5.9\times$ random on test) and does not clearly survive on val. This is consistent with Rhea's SwissProt annotations being concentrated on well-characterised reference organisms: *A. thaliana*, as the most extensively characterised plant species in both UniProt and Rhea, benefits disproportionately from this cross-validation, while the taxa holdout's 69 held-out test species (less represented in reference databases) offer far fewer Rhea-annotated catalysts for the model to be credited against in the first place. The taxa-holdout result should therefore be read as a weak positive consistent with the *A. thaliana* finding, not as independent confirmation at comparable strength.

I.4 Rhea-Based Independent Validation (*A. thaliana* Reaction Holdout)

PlantMetWiki's GPML pathway files embed Rhea cross-references for a subset of reactions (Bansal et al., 2022). Rhea is an expert-curated, independently maintained biochemical reaction database whose enzyme annotations derive from UniProt SwissProt evidence, distinct from the PlantCyc curation underlying PlantMetWiki. We therefore ask whether proteins that HeteroSAGE+Res&Jump ranks highly, but that PlantMetWiki does *not* label as the true catalyst, are nonetheless supported by Rhea, providing evidence that the benchmark under-counts true positives.

Coverage. Of the 454 unique test reactions, 131 (28.9%) carry at least one Rhea identifier via PlantMetWiki's embedded GPML xrefs. Mapping these Rhea IDs through `rhea2uniprot_sprot.tsv` (SwissProt; 14,163 Rhea→UniProt entries) and then through PlantMetWiki's `properties_extra.ttl` UniProt cross-references yields 79 reactions with at least one Rhea-annotated protein in the 5,289-protein A. thaliana ranking pool (Table 29). The remaining 52 reactions with Rhea IDs but no matching pool protein are catalysed by enzymes whose Rhea-curated UniProt accession belongs to a non-*A. thaliana* reference organism (human, *E. coli*, yeast) absent from the A. thaliana pool. This 79-reaction (test) and 40-reaction (val) evaluable set is smaller than the 132 and 203 reactions evaluable under the taxa holdout (Appendix I.3), even though the underlying Rhea/UniProt cross-reference coverage is the same graph-wide property in both cases: the taxa holdout's much larger 8,445-protein pool simply captures more of those graph-wide annotated proteins than the 5,289-protein *A. thaliana* pool does.

Table 29: Rhea cross-reference coverage for the 454 unique *A. thaliana* test reactions (seed 42). UniProt mapping via `rhea2uniprot_sprot.tsv` (SwissProt only). "In pool" means at least one Rhea-annotated UniProt accession maps to a protein in the 5,289-protein *A. thaliana* ranking pool.

Criterion	Reactions	% of test set
With Rhea xref (GPML source)	131	28.9%
of which: $\geq$1 annotated pool protein	79	17.4%
of which: no matching pool protein	52	11.5%
Without Rhea annotation	323	71.1%

Result. For the 79-reaction (test) and 40-reaction (val) evaluable subsets, HeteroSAGE+Res&Jump reaches Rhea anyP-H@50 of $5.57\pm4.79\%$ (test) and $8.50\pm8.40\%$ (val) across 5 seeds, roughly 6–9$\times$ the 0.94% random baseline on average, though with substantial seed-to-seed variance (individual seeds range from 0% to 17.5%).

Interpretation. The model ranks Rhea-annotated proteins highly at a rate that is, on average, well above chance, even though these proteins are not labelled as true catalysts under PlantMetWiki's own annotations for that reaction. This is consistent with the benchmark under-counting true positives, instead of hallucinating spurious associations, as an independently curated database from SwissProt evidence (with no direct dependency on PlantCyc curation) corroborates a meaningful fraction of the

model's top-ranked candidates. The wide seed-to-seed spread is itself informative: with only 40–79 evaluable reactions per split, a handful of additional hits shifts the rate substantially, so this result should be read as evidence of a real but modest enrichment. PlantMetBench encodes substantial *original* biochemical knowledge not already present in Rhea: the benchmark draws on PlantCyc's literature curation of plant metabolism, which is largely orthogonal to the reference-organism focus of SwissProt, and this analysis provides positive support for that claim. Future experiments such as restricting to *A. thaliana* SwissProt entries from UniProt proteome `UP000006548` (The UniProt Consortium, 2023) remain an open avenue for future work.

Since *A. thaliana* is the most extensively characterised plant species in both UniProt and Rhea, this modest enrichment is likely close to an upper bound on what cross-validation against Rhea can currently detect for the remaining 403 species in the benchmark, the majority of which are far less represented in reference databases.

J Computational Resources

Table 30: Number of parameters of each model trained and the average runtime per epoch.

Model	Layers	Hidden	# Params	Avg. time/epoch (s)
Dual encoder	L1	2048	10,368,000	0.02 ± 0.001
HeteroSAGE	L1	128	3,847,808	0.084 ± 0.002
GAT	L1	128	1,080,448	0.082 ± 0.001
HGT	L1	128	1,191,446	0.095 ± 0.001
R-GCN	L1	128	1,075,840	0.077 ± 0.002
HeteroSAGE+Res	L1	128	4,409,984	0.086 ± 0.001
HeteroSAGE+Res&Jump	L1	128	1,142,912	0.078 ± 0.001
Mix	L1	128	1,108,096	0.078 ± 0.001
Dual encoder	L2	2048	18,768,896	0.038 ± 0.001
HeteroSAGE	L2	128	4,062,976	0.089 ± 0.001
GAT	L2	128	1,382,144	0.098 ± 0.001
HGT	L2	128	1,604,140	0.115 ± 0.002
R-GCN	L2	128	1,481,728	0.085 ± 0.001
HeteroSAGE+Res	L2	128	4,707,072	0.092 ± 0.001
HeteroSAGE+Res&Jump	L2	128	1,505,536	0.086 ± 0.001
Mix	L2	128	1,437,440	0.085 ± 0.001
Dual encoder	L3	2048	27,169,792	0.055 ± 0.001
HeteroSAGE	L3	128	4,547,456	0.095 ± 0.002
GAT	L3	128	1,683,840	0.113 ± 0.003
HGT	L3	128	2,016,834	0.141 ± 0.001
R-GCN	L3	128	1,670,016	0.088 ± 0.001
HeteroSAGE+Res	L3	128	5,004,160	0.096 ± 0.002
HeteroSAGE+Res&Jump	L3	128	1,868,160	0.094 ± 0.002
Mix	L3	128	1,766,784	0.092 ± 0.001

Some of the work underlying this manuscript was part of an internal hackathon and communal effort, so the models were run across different machines.

GNNs All models are lightweight and fast to train as shown in Table 30. Values are computed on NVIDIA A100 GPUs (40 GiB HBM2). Graph architectures range from $\approx$ 1.08M parameters (GAT, R-GCN) to $\approx$ 5.00M parameters (HeteroSAGE+Res), growing with the number of layers. Per-epoch time remains low, ranging from 0.07–0.14 s across the evaluated GNN architectures and depths.

Dual encoders. For runs with deduplicated protein values, the average per-epoch runtimes are 0.011 $\pm$ 0.001 s, 0.019 $\pm$ 0.001 s, and 0.027 $\pm$ 0.001 s for L1, L2, and L3, respectively.

All experiments are implemented in PyTorch.