GENEBRIDGE: HOW TO BRIDGE GENE SPACES FOR TRANSDUCTIVE ZERO-SHOT GRN INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Gene regulatory network (GRN) inference faces a fundamental precision–generality trade-off: supervised methods achieve high accuracy on labeled cell types but cannot generalize to unseen ones, whereas unsupervised and foundation model-based approaches apply broadly at the cost of predictive accuracy. We propose GeneBridge, a transductive zero-shot framework that transfers regulatory knowledge across cell types through a unified semantic gene graph. The central challenge is gene space misalignment: different cell types express different gene subsets, making expression-based representations non-transferable. GeneBridge addresses this by constructing a transferable semantic graph from foundation model embeddings, where hybrid node representations—contextual for expressed genes, frozen pretrained for absent ones—allow all genes to coexist in a shared semantic space. Experiments on seven BEELINE datasets evaluate regulatory knowledge transfer under leave-one-cell-type-out evaluation. Ablation studies further demonstrate that semantic gene selection is foundational: replacing it with standard HVG filtering reduces Top-100 precision to near zero. Together, these results demonstrate that semantic graph learning provides a principled mechanism for transferring regulatory knowledge across heterogeneous biological contexts.

1 INTRODUCTION

Gene regulatory network (GRN) inference identifies which transcription factors (TFs) regulate which target genes from single-cell RNA sequencing data. It is important for understanding cell identity, development, and disease mechanisms (Pratapa et al., 2020). It therefore requires both accurate predictions and the ability to generalize to target cell types without regulatory annotations. Existing methods fall into two categories. The first includes methods based on statistical associations, expression-based inference, or pretrained foundation model representations. These methods use known gene expression in known cells to infer unknown relationships without regulatory labels from the target cell type. Examples include GENIE3 (Huynh-Thu et al., 2010), GRNBoost2 (Moerman et al., 2019), DeepSEM (Shu et al., 2021), and methods based on foundation model embeddings (Cui et al., 2024; Yang et al., 2024). These methods are flexible and thus widely applicable, but their accuracy in identifying regulatory interactions is very low (Pratapa et al., 2020). The second category is based on supervised learning. Methods such as GNNLink (Mao et al., 2023), GENELink (Chen & Liu, 2022), and GCLink (Yu et al., 2025b) often achieve strong predictive performance when regulatory annotations are available. However, their reliance on known regulatory networks limits their use in unannotated cell types.

We identify two reasons for this gap between generalization to target cell types without regulatory annotations and prediction accuracy in GRN inference. First, cells contain a large number of genes, while actual biological regulatory networks are sparse. Existing methods select training data using only simple highly variable gene (HVG) criteria. As a result, data from many irrelevant genes affect learning and training outcomes. Second, gene expression spaces are misaligned when models are trained directly on annotated source cell types and transferred across cell types (CTs) to predict regulatory edges in an unannotated target cell type. Current methods learn from gene expression data using only traditional statistics or large pretrained models. This makes it difficult to extract

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.

the high-dimensional biological relationships that connect the regulatory network as a whole. These methods cannot make good use of the biological causal effects and overall relationships in gene regulation across cell types. This severely limits their ability to generalize.

To address these issues, we propose a framework called GeneBridge. It first uses an attention mechanism, which can better capture relationships, to select data in expression space. It then uses supervised graph learning suited to complex gene regulatory networks to capture regulatory relationships. Genes from different spaces are mapped into a shared biological semantic space to study their regulatory relationships and biological connections. This produces a transferable semantic graph that learns regulatory knowledge from annotated source cell types and applies it to regulatory prediction in target cells. No annotations from the target cell type are needed. GeneBridge has two stages: (1) an unsupervised stage that jointly selects regulatory genes and builds a unified semantic graph across all cell types; and (2) a supervised stage that uses a graph Transformer trained across cell types with soft labels to perform zero-shot inference on target cell types without regulatory annotations (Figure 1). By decoupling gene identity from expression, GeneBridge effectively connects gene spaces, cell types, and species.

Experiments on seven BEELINE datasets evaluate zero-shot GRN inference, with Precision@100 reaching 89.0% on hESC. Ablations highlight the importance of both semantic gene selection and edge-aware graph learning: HVG filtering reduces T100 to 0–1, while GraphTransformer achieves higher average T100 than the tested GCN and GAT configurations.

Our contributions are centered on a unified transferable semantic graph learning framework for zero-shot GRN inference:

1. We identify two core barriers to cross-cell-type GRN inference: selecting data for genes with true regulatory relationships from large amounts of gene expression data, and extracting biological relationships across expression spaces. This perspective shifts the focus from learning cell-specific regulatory patterns to learning shared, transferable gene semantics across biological contexts.
2. We propose GeneBridge, a framework based on foundation models. It uses an attention mechanism to select biologically meaningful genes more efficiently. It also aligns heterogeneous gene spaces through semantic representations to build a transferable semantic graph. On this graph, an edge-aware graph Transformer uses both the constructed semantic relationships and expression dynamics to achieve highly accurate regulatory inference.
3. Our experiments show that data selection and semantic graph learning are key to generalization, rather than expression similarity alone. Extensive experiments on seven BEELINE datasets and comprehensive ablation studies show that GeneBridge consistently outperforms existing methods on target cell types without regulatory annotations, both within and across species.

2 Related Work

Expression-based methods. Classical GRN inference uses statistical associations (PCC, MI), tree-based feature importance (GENIE3 (Huynh-Thu et al., 2010), GRNBoost2 (Moerman et al., 2019)), or deep generative models (DeepSEM (Shu et al., 2021)). These methods learn expression-to-regulation mappings within a single cell type and cannot transfer when the gene vocabulary changes.

Graph-based methods. Graph learning has been applied to GRN inference through methods such as GRNformer (Hegde & Cheng, 2026), GT-GRN (Teji & Roy, 2025), KINDLE (Peng et al., 2026), and KEGNI (Li et al., 2025). Meta-TGLink (Yu et al., 2025a) further explores few-shot inference. These methods use expression information, network structure, or biological priors to learn gene representations and regulatory relationships. For example, KEGNI jointly learns representations from expression data and a biological knowledge graph. This shows how external knowledge can complement expression-based features. GeneBridge focuses on transferring regulatory knowledge across different gene spaces in heterogeneous cell types. Pretrained gene embeddings form a shared semantic graph. This graph is combined with cell-type-specific expression information and source regulatory labels to train a predictor. The predictor is then applied to target data without target regulatory supervision.

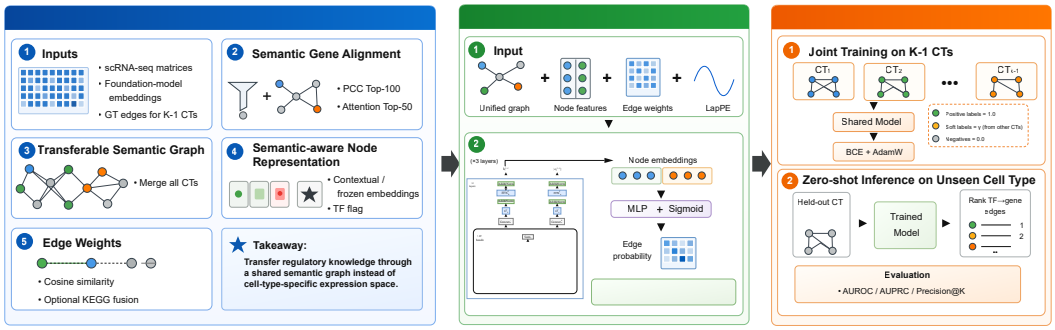

Figure 1: Overview of GeneBridge. GeneBridge addresses gene space misalignment by transferring regulatory knowledge through a shared semantic graph rather than cell-type-specific expression space. (a) constructs a transferable semantic graph from foundation-model-derived gene representations; (b) performs edge-aware message passing for GRN prediction; and (c) enables zero-shot inference on unseen cell types through joint training with soft labels.

Foundation model-based methods. scGPT (Cui et al., 2024), GeneCompass (Yang et al., 2024), and related models learn gene representations from large transcriptomic corpora. Benchmarking across gene- and cell-level tasks highlights that their utility depends on the downstream task and evaluation setting (Wu et al., 2025). Here, we investigate how these representations can support regulatory edge prediction across cell types. The critical gap is that **no prior work constructs an explicit transferable graph from these embeddings for zero-shot cross-cell-type GRN inference**.

Evaluation frameworks. BEELINE (Pratapa et al., 2020) and geneRNIB (Nourisa et al., 2025) standardize intra-CT evaluation; Stock et al. (Stock et al., 2025) identify cross-context generalization as a bottleneck. We extend these by establishing a cross-CT, cross-species zero-shot evaluation framework.

3 Method

GeneBridge is motivated by the observation that regulatory knowledge transfer becomes feasible once genes from heterogeneous cell types are projected into a shared semantic graph. We operationalize this insight through three integrated modules.

3.1 Problem Formulation and Key Insight

We consider K single-cell transcriptomic datasets $\{\mathcal{D}_k\}_{k=1}^{K}$ from distinct cell types, where $\mathcal{D}_k = (\mathbf{X}_k, \mathcal{G}_k)$. For cell type k, let $\mathcal{V}_k$ denote the gene set, $\mathcal{T}_k \subseteq \mathcal{V}_k$ the transcription factor set, and C_k the number of cells. The expression matrix is $\mathbf{X}_k \in \mathbb{R}^{|\mathcal{V}_k| \times C_k}$, and the reference regulatory edge set satisfies

$$\mathcal{G}_k \subseteq \{(t, g) \in \mathcal{T}_k \times \mathcal{V}_k \mid t \neq g\}.$$

This candidate space allows TF-to-TF interactions while excluding self-loops. We study transductive zero-shot transfer: regulatory labels are available only for source cell types during learning, while unlabeled expression data from the target cell type are available for self-supervised foundation-model adaptation, gene selection, and graph construction. The task is to predict regulatory edges in the target cell type without using its regulatory labels for training or fine-tuning.

Why existing methods fail. For any two cell types A and B, their gene sets $\mathcal{V}_A$ and $\mathcal{V}_B$ differ substantially. A standard approach learns a model $f(\mathbf{X}_A; \mathcal{G}_A)$ that maps from expression-based features to edge predictions. This model cannot be applied to B because: (1) many genes in $\mathcal{V}_B$ are absent from $\mathbf{X}_A$, so the model has never seen their features; (2) even for shared genes, expression patterns differ between cell types, violating the i.i.d. assumption. This is the gene space misalignment problem.

Our solution. We combine shared gene representations with cell-type-specific expression information. Pretrained gene embeddings $\mathbf{E}^{\text{pt}}$ provide a common representation space that can encode functional relationships useful across cell types. We use their semantic similarities to construct a transferable graph, while contextual embeddings provide information about the expression context of each cell type. The graph serves as a shared scaffold for learning regulatory relationships from source labels; it does not assume that the same regulatory edges are active in every cell type.

Formally, GeneBridge learns a zero-shot edge predictor by jointly training on source cell types over the constructed semantic graph:

$$\min_{\theta} \sum_{k \in \mathcal{K}_{\text{train}}} \mathcal{L}_{\text{BCE}}\big(f_\theta(\mathcal{G}(\mathcal{V}^{\text{sel}}, \mathcal{E}, \mathbf{E}^{\text{pt}}); \mathbf{X}_k),\ \tilde{\mathbf{y}}^{(k)}\big), \tag{1}$$

where $\mathcal{K}_{\text{train}}$ indexes the source cell types and excludes the held-out target cell type. $\mathcal{V}^{\text{sel}}$ and $\mathcal{E}$ denote the node set and edge set of the semantic graph, both determined once during preprocessing (Modules I–II); only the model parameters θ are optimized during training, while $\tilde{\mathbf{y}}^{(k)}$ are soft labels incorporating regulatory knowledge from all training CTs. Each component is detailed below.

3.2 Module I: Semantic Gene Selection

Standard BEELINE pipelines select candidate genes using highly variable genes (HVGs) filtered by expression variance. While computationally convenient, this strategy is fundamentally misaligned with the goal of GRN inference: a gene's expression variance across cells within a CT does not indicate whether it serves as a TF target. As our ablation demonstrates (§4.3.1), replacing our gene selection with HVG causes Top-100 precision to drop from 89.0 to near zero—the model can distinguish trivial negatives but cannot identify true regulatory edges among its highest-confidence predictions. This result motivates selecting candidate genes for regulatory prediction rather than expression variability alone.

We therefore use foundation-model signals to select candidate genes for regulatory prediction.

3.3 Foundation Model-Driven Dual-Pathway Gene Selection

We propose a sequential two-step strategy: a coarse statistical filter followed by a fine attention-based filter. This dual-pathway design is motivated by a practical consideration: attention computation over the full gene set is quadratic in the number of genes, but co-expression provides a computationally cheap first pass that reduces the attention search space by approximately two orders of magnitude.

Co-expression coarse filter. For each TF $t \in \mathcal{T}_k$, compute Pearson correlation with all non-TF genes g across cells, retaining the top n_{co}=100 by $|r_{t,g}|$ as the coarse candidate set $\mathcal{C}_t$:

$$r_{t,g} = \frac{\sum_c (x_{t,c} - \bar{x}_t)(x_{g,c} - \bar{x}_g)}{\sqrt{\sum_c (x_{t,c} - \bar{x}_t)^2}\sqrt{\sum_c (x_{g,c} - \bar{x}_g)^2}} \tag{2}$$

Co-expression provides a rough but computationally cheap regulatory signal—TF-target gene pairs with strong co-expression are enriched for true regulatory relationships, making it an effective coarse filter.

Attention fine filter. We first fine-tune GeneCompass (Yang et al., 2024) (12-layer BERT-style Transformer, d=768, pretrained on human+mouse ∼50K genes) via self-supervised MLM on pooled expression data from all K CTs, freezing the bottom 9 layers and training the top 3 plus token embeddings (1,000 steps, lr $= 5\times10^{-6}$). This fine-tuning adapts the gene embedding space using only expression data, without accessing any ground truth regulatory annotations from any cell type. The fine-tuned model f_{GC} is then applied to $\mathbf{X}_k$. We average final-layer attention across cells and heads, then use each gene's most frequent token position to obtain a gene-pair attention proxy.

We retain the top n_{att}=50 genes from $\mathcal{C}_t$, forming $\mathcal{A}_t$. The coarse filter reduces the attention computation cost by ∼200× (from ∼20K to 100 genes per TF). This formulation is motivated by the observation that attention weights reflect non-linear, context-dependent gene-gene relationships learned during pretraining on cross-species data—information not captured by co-expression alone. The final gene set per CT is $\mathcal{V}_k^{\text{sel}} = \mathcal{T}_k \cup \bigcup_{t \in \mathcal{T}_k} \mathcal{A}_t$.

3.4 MODULE II: SEMANTIC GRAPH CONSTRUCTION

Unified graph. We merge gene subgraphs from all K cell types into a single graph: $\mathcal{V} = \bigcup_{k=1}^{K} \mathcal{V}_k^{\text{sel}}$, $N = |\mathcal{V}|$. Edge set: $\mathcal{E} = \bigcup_{k=1}^{K} \{(t, g) \mid t \in \mathcal{T}_k, g \in \mathcal{V}_k^{\text{sel}} \setminus \{t\}\}$. Gene selection and graph construction include the held-out CT and use no regulatory labels.

Hybrid node embeddings. Different CTs express different gene subsets, so expression-derived features are not directly transferable across CTs. We resolve this via hybrid embeddings that combine cell-type-specific and universal representations: for genes expressed in a given CT, we use contextual embeddings from the fine-tuned foundation model (capturing CT-specific expression context); for genes absent from that CT, we fall back to the frozen pretrained embedding (ensuring every gene retains a stable position in semantic space). Formally, for gene $i \in \mathcal{V}$ in CT k:

$$\mathbf{h}_i^{(k)} = \begin{cases} f_{\text{GC}}(i; \mathbf{X}_k), & i \in \mathcal{V}_k \text{ (expressed)} \\ \mathbf{E}_i^{\text{pt}}, & \text{otherwise (absent)} \end{cases} \tag{3}$$

where $\mathbf{E}_i^{\text{pt}} \in \mathbb{R}^d$ is the frozen pretrained token embedding. Contextual embeddings are averaged over cells in which the gene appears as an input token. The final node feature is $\mathbf{x}_i^{(k)} = \mathbf{h}_i^{(k)} \parallel \not\Vdash_{\text{TF}}(i) \in \mathbb{R}^{d+1}$, with $d_{\text{in}} = 769$. The frozen embedding acts as a fixed "coordinate" in semantic space, ensuring that every gene—regardless of expression status—has a stable representation that the Graph Transformer can learn to associate with regulatory roles.

Edge weights. We define edge weights using pretrained embedding cosine similarity, motivated by the hypothesis that genes with similar functional roles occupy nearby positions in the foundation model's embedding space. The cosine similarity is normalized to $[0, 1]$:

$$\tilde{w}_{ij} = \frac{1}{2}\left(\frac{\langle \mathbf{E}_i^{\text{pt}}, \mathbf{E}_j^{\text{pt}} \rangle}{\|\mathbf{E}_i^{\text{pt}}\| \|\mathbf{E}_j^{\text{pt}}\|} + 1\right) \tag{4}$$

Optionally fused with KEGG pathway Jaccard coefficient: $w_{ij} = \lambda \cdot \text{Jaccard}(i, j) + (1-\lambda) \cdot \tilde{w}_{ij}$, where λ is species-adaptive (λ=0.3 for human, λ=0 for mouse; see §4.4).

3.5 MODULE III: CROSS-CELL-TYPE KNOWLEDGE TRANSFER

Graph Transformer encoder. We use a Graph Transformer with edge-feature-modulated multi-head attention and Laplacian positional encoding. Attention logits are modulated by learned edge features: $e_{uv}^h = (\mathbf{Q}^h \mathbf{h}_u^{(l)})^\top (\mathbf{K}^h \mathbf{h}_v^{(l)}) / \sqrt{d_{\text{head}}} \cdot s_{uv}^h$, where s_{uv}^h is the sum of the projected edge-feature components for head h. **EdgePredictor:** an MLP scoring node pairs from concatenated encoder embeddings. Complete architecture specifications are provided in Appendix B.

Soft-label multi-CT joint training. Let $\mathcal{K}_{\text{train}}$ denote the $K-1$ cell types with known GT and $\mathcal{K}_{\text{test}}$ the single held-out cell type. Training iterates exclusively over $\mathcal{K}_{\text{train}}$; the held-out CT's ground truth $\mathcal{G}_{\text{test}}$ is never accessed. For a focal training CT $k \in \mathcal{K}_{\text{train}}$, soft labels leverage regulatory knowledge from other training CTs without treating their annotations as definitive:

$$\ell_{uv}^{(k)} = \begin{cases} 1.0, & (u, v) \in \mathcal{G}_k \\ \gamma, & (u, v) \in \bigcup_{j \in \mathcal{K}_{\text{train}} \setminus \{k\}} \mathcal{G}_j \setminus \mathcal{G}_k \\ 0.0, & \text{randomly sampled non-GT} \end{cases} \tag{5}$$

with γ=0.3 and negative sampling ratio NR=7. The union in the second case is taken only over other training cell types—not the held-out CT. This formulation is motivated by the observation that edges present in other training CTs' ground truth are more likely to represent conserved regulatory relationships than randomly sampled pairs, but should not be treated as definitively true for the focal CT—hence the soft label γ rather than a hard 1.0. This design reflects the assumption that regulatory interactions conserved across multiple training cell types have a higher prior probability than randomly sampled negatives, while remaining cell-type-dependent (hence γ<1). Training uses BCE

loss with AdamW (lr $= 10^{-3}$, wd $= 10^{-4}$), batch size 4096, early stopping patience 3. The held-out CT participates structurally in the unified graph—its genes are nodes and its edges carry semantic similarity weights—but its regulatory labels remain entirely unseen during training. Inference is performed on the held-out CT's subgraph over the full candidate edge space.

3.6 Complexity Analysis

The graph predictor has approximately 0.84M trainable parameters. Attention operates only on graph edges. Including node and edge projections and feed-forward networks, each encoder layer requires $O((N + |\mathcal{E}|)d_h^2)$ time and $O((N + |\mathcal{E}|)d_h)$ activation memory, where N is the number of nodes and d_h is the hidden dimension.

3.7 Evaluation Protocol

Unless otherwise specified, metrics are computed over the full TF→gene candidate edge space without evaluation-time sampling. We report **AUROC**, **AUPRC**, and **Top-K Precision**. Let $\mathcal{E}_{\text{test}}$ denote the complete candidate edge set for the evaluated method and configuration, and let $\mathcal{G}_{\text{test}}$ denote the held-out reference edge set. Define $\mathcal{P}_{\text{test}} = \mathcal{G}_{\text{test}} \cap \mathcal{E}_{\text{test}}$ as the reference positives within this candidate space. Then

$$\text{Precision@}K = \frac{|\text{Top}_K(\hat{y}; \mathcal{E}_{\text{test}}) \cap \mathcal{P}_{\text{test}}|}{K}, \tag{6}$$

where $\text{Top}_K(\hat{y}; \mathcal{E}_{\text{test}})$ contains the K highest-scoring candidate edges. Under a uniformly random ranking of the same candidate space, the expected precision is

$$p_{\text{rand}} = \frac{|\mathcal{P}_{\text{test}}|}{|\mathcal{E}_{\text{test}}|}.$$

We report enrichment over this random baseline as

$$\times\text{Random} = \frac{\text{Precision@}K}{p_{\text{rand}}}, \qquad p_{\text{rand}} > 0.$$

In the tables, T100 and T500 denote the numbers of reference positives among the top 100 and 500 predictions, respectively; non-integer values arise when counts are averaged across runs or datasets. Thus, $\text{T}K = K\,\text{Precision@}K$, and Rand T100 denotes the expected random count $100p_{\text{rand}}$. T100 is numerically equal to Precision@100 expressed as a percentage, whereas Precision@500 (%) equals T500 divided by five. Top-K precision measures agreement with the reference network among the highest-ranked predictions.

Three LOO settings: Human 2CT (γ=0), Mouse 5CT (γ=0.3), Cross-species 4CT (γ=0.3). Complete hyperparameters in Appendix B.

4 Experiments

4.1 Setup

Like the UGRN work (Qi et al., 2026), we evaluate on seven BEELINE datasets (Pratapa et al., 2020) spanning two human and five mouse cell types, using Combined ChIP-seq reference networks. Baselines cover statistical, tree-based, deep learning, and foundation-model approaches; dataset statistics and baseline settings are provided in Appendix A. The main comparison uses TF+500 HVG for baselines and semantic selection for GeneBridge, each evaluated over its full candidate space. GeneCompass+DeepSEM is reported separately under its own protocol. Main LOO results average 10 seeds, with standard deviations in Appendix C; implementation details and hyperparameters are in Appendix B.

4.2 Main Results

Same-species zero-shot prediction. Table 1 shows results within species. For human data, the held-out CT name serves as the row label. For example, "hHep" means training on hESC and predicting

Table 1: Same-species zero-shot results. Row = held-out CT. 10-seed mean.

CT	AUROC	AUPRC	Rand T100	T100	×Rand
Human 2CT					
hHep	0.791	0.253	3.2	82.0	26×
hESC	0.790	0.242	2.7	89.0	33×
Mouse 5CT					
mESC	0.537	0.083	7.1	14.3	2.0×
mDC	0.650	0.038	1.7	32.5	19×
mHSC-E	0.783	0.398	7.5	93.8	13×
mHSC-GM	0.783	0.398	7.4	91.1	12×
mHSC-L	0.774	0.388	7.4	90.8	12×

Table 2: Cross-species 4CT LOO results. 10-seed mean.

Holdout	Species	AUROC	AUPRC	Rand T100	T100
hESC	Human	0.600	0.051	2.7	35.4
hHep	Human	0.621	0.058	3.2	18.5
mESC	Mouse	0.549	0.092	7.1	44.8
mDC	Mouse	0.655	0.032	1.7	5.2

on hHep. Both directions achieve strong T100 scores (82.0 and 89.0, respectively; ×Random = 26–33×). For mouse data, HSC subtypes perform best (T100 = 90.8–93.8, ×Random = 12–13×), reflecting highly conserved hematopoietic regulatory networks. mESC is the most challenging (T100 = 14.3). The mESC dataset contains over 250,000 ground-truth (GT) regulatory edges, more than the combined total of the other four mouse training CTs. Three of these CTs are HSC subtypes with highly similar GT annotations from the same ChIP-seq source, and their regulatory patterns are highly HSC-specific. As a result, the constructed graph contains far less useful information about regulatory relationships than the raw edge count suggests. This shows that the amount of useful information in the graph is crucial for prediction across cell types. It highlights the importance of expression information from known samples and relationship information in the graph structure for prediction generalization. Despite its sparse GT data, mDC maintains strong discrimination (T100 = 32.5, ×Random = 19×).

Cross-species transfer. T100 improves for mESC (14.3 to 44.8) but declines for hESC (89.0 to 35.4), hHep (82.0 to 18.5), and mDC (32.5 to 5.2). One possible explanation is that hESC provides regulatory information more relevant to mESC than the HSC-dominated mouse training set, whereas the altered source composition reduces compatibility with the human targets and lacks immune-lineage context for mDC. These patterns suggest that source–target regulatory compatibility matters beyond species membership.

Baseline comparison. Table 3 compares complete pipelines with different candidate spaces; the results therefore reflect both gene selection and regulatory prediction. GeneBridge achieves AUROC 0.790 and T100 89.0 on hESC, and AUROC 0.774, AUPRC 0.388, and T100 90.8 on mHSC-L. On mESC, scGPT's AUROC 0.684 (no cross-CT constraint) exceeds our 0.537 under strict LOO. We note GRNformer (Hegde & Cheng, 2026) reports AUROC 0.96–0.99 under subgraph evaluation; under our full-graph protocol this drops to 0.532–0.632 with T100 8–36%.

4.3 Main Ablation Experiments

4.3.1 The Foundational Role of Semantic Gene Selection

This is the central experiment of our paper. We replace the original dual-pathway gene selection strategy with BEELINE-standard TF+500 and TF+1000 HVG filtering. All other components remain unchanged, and we retrain the model from scratch. Human 2CT cross-validation results (Table 4) show that replacing semantic selection with HVG reduces T100 to 0–1, even though AUROC remains at 0.71–0.78. This shows that gene selection is crucial for GeneBridge's top-ranked pre-

Table 3: Method-specific candidate spaces: TF+500 HVG for baselines and semantic selection for GeneBridge. Bold marks column maxima of the reported values; protocols differ across methods. GeneCompass+DeepSEM is a separate reference using its retained gene sets and Combined ChIP-seq/STRING/LOF-GOF networks; AUROC and AUPRC are separately selected epoch maxima, and T100 is from the best-AUPRC epoch.

	hESC			mESC			mHSC-L		
Method	AUROC	AUPRC	T100	AUROC	AUPRC	T100	AUROC	AUPRC	T100
Random	0.500	0.033	1	0.500	0.030	1	0.500	0.060	1
PCC	0.514	0.036	19	0.610	0.049	11	0.550	0.082	22
MI	0.518	0.036	29	0.588	0.046	10	0.460	0.055	25
GRNBoost2	0.499	0.034	23	0.502	0.032	17	0.508	0.064	12
GENIE3	0.562	0.060	4	0.579	0.070	8	0.639	0.180	17
DeepSEM	0.525	0.033	20	0.551	0.035	19	0.536	0.061	67
scGPT	0.581	0.047	20	**0.684**	0.058	10	0.532	0.048	20
GRNformer	0.532	0.051	8	0.614	0.080	17	0.632	0.225	36
GeneCompass+DeepSEM	0.501	0.205	12	0.431	0.081	**22**	0.463	0.088	31
GeneBridge	**0.790**	**0.242**	**89.0**	0.537	**0.083**	14.3	**0.774**	**0.388**	**90.8**

Table 4: Gene selection ablation. HVG reduces T100 to near zero.

Gene Selection	Held-out	T100	T500	AUROC
TF+500	hHep	0	1	0.730
TF+500	hESC	1	4	0.715
TF+1000	hHep	0	5	0.781
TF+1000	hESC	0	2	0.737
GeneBridge	hHep	**82.0**	**367**	**0.791**
GeneBridge	hESC	**89.0**	**387**	**0.790**

dictions. High-quality data selection is the foundation of effective graph construction and model training. Training with traditional HVG selection also fails to produce good results. These findings demonstrate the important role of the semantic selection module in the overall framework.

4.3.2 Encoder Architecture Ablation

To test how well GraphTransformer captures regulatory relationships in the graph structure and semantic information in gene expression, we replace it with GCN and GAT. These are two representative models in graph neural networks with architectures similar to GraphTransformer. We compare their T100 scores. GraphTransformer achieves a higher average T100 (75.3) than the tested GCN (25.1) and GAT (20.4) configurations (Table 5). This demonstrates the better overall performance of GraphTransformer on this task.

4.4 Additional Ablations

Ablation averages come from separate runs under the settings specified in the appendices and need not match the multi-seed main results.

Soft-label weight γ. We evaluate $\gamma \in \{0, 0.1, 0.3, 0.5, 0.7\}$ on mouse 5CT LOO. γ=0.3 achieves the best Avg T100 of 68.8 (vs. 59.0 for γ=0), a +16.6% gain that far exceeds the corresponding AUROC improvement (+2.8%), indicating that soft labels primarily enhance high-confidence edge identification. Full results are provided in Appendix Section D.2.

Edge weight: Cosine vs. KEGG. Pure cosine similarity gives the highest average T100 on mouse datasets (71.2), while adding KEGG information improves human T100 from 79.5 to 90.0. Full results are in Appendix Section D.4.

HSC merged validation. Merging mHSC-E/GM/L into one CT and performing 3-fold LOO: mHSC T100 = 92 remains robust (Table 6), confirming HSC performance stems from intrinsic pre-

Table 5: Encoder architecture ablation. Averages over 7 datasets (5 mouse + 2 human LOO). All other components (hybrid embeddings, soft-label training, evaluation protocol) held constant.

Encoder	Avg T100	Avg AUROC	Avg AUPRC
GAT	20.4	0.589	0.084
GCN	25.1	0.579	0.071
GraphTransformer	**75.3**	**0.732**	**0.266**

Table 6: HSC merged 3-fold LOO.

Holdout	AUROC	AUPRC	T100
mESC	0.548	0.078	8
mDC	0.672	0.044	29
mHSC (merged)	0.718	0.186	**92**

dictability, not soft-label leakage among subtypes. Notably, mESC T100 drops from 14.3 (in 5CT LOO) to 8 in this merged setting. This is because the merged HSC contributes $\sim$140K GT edges as training supervision—roughly half of mESC's own 257K edges—and merging further reduces the already limited diversity of non-mESC training signals. This observation reinforces our earlier analysis that mESC's difficulty stems from insufficient quantity and diversity of non-mESC training supervision, rather than from any deficiency of the semantic graph.

5 Discussion and Conclusion

We presented GeneBridge, a transferable semantic graph learning framework for zero-shot cross-cell-type GRN inference. Using foundation model embeddings as a universal semantic coordinate system, GeneBridge bridges gene spaces across cell types and species.

Our ablations highlight the importance of both **semantic gene selection and edge-aware graph learning**. Replacing semantic selection with HVG filtering reduces T100 to 0–1, while GraphTransformer outperforms the tested GCN and GAT configurations. Together, these components support regulatory prediction across heterogeneous gene spaces.

Why does the semantic graph transfer? Pretrained gene embeddings define a shared semantic graph through which the predictor learns regulatory patterns from source annotations. Hybrid node representations combine this shared foundation with cell-type-specific expression context, allowing the same predictor to score regulatory edges across heterogeneous gene spaces. Transfer effectiveness depends on the compatibility of source and target regulatory programs.

Limitations and future work. mESC performance remains modest when training CTs lack pluripotency-related supervision. The fixed gene vocabulary of the pretrained model excludes genes absent from the pretraining corpus. Future work includes incorporating multi-modal data (chromatin accessibility, histone modifications), modeling CT-specific transferability, and extending the semantic graph principle to other biological network prediction tasks such as cross-species protein interaction and drug-target prediction. This work is designed for basic biological research and does not involve clinical decision-making or patient data; no ethical concerns are raised. Code and pretrained models will be publicly released upon publication.

AI use statement

We used generative AI tools to assist with drafting and revising the manuscript, reviewing code for potential errors, and checking the consistency of result interpretations and evaluation descriptions. The authors reviewed the AI-assisted text and code-review suggestions and take full responsibility for the final manuscript, implementation, and reported results.

REFERENCES

Guangyi Chen and Zhi-Ping Liu. Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data. *Bioinformatics*, 38(19):4522–4529, 2022. doi: 10.1093/bioinformatics/btac559.

Haotian Cui, Chloe Wang, Hassaan Maan, et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024.

Akshata Hegde and Jianlin Cheng. Grnformer: a graph transformer framework for gene regulatory network inference from single-cell transcriptomics. *Bioinformatics*, 42(4):btag144, 2026.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, et al. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 5(9), 2010.

Pengxiao Li, Lin Li, Jingminjie Nan, et al. Kegni: knowledge graph enhanced framework for gene regulatory network inference. *Genome Biology*, 26, 2025.

Guo Mao, Zhengbin Pang, Ke Zuo, Qinglin Wang, Xiangdong Pei, Xinhai Chen, and Jie Liu. Predicting gene regulatory links from single-cell RNA-seq data using graph neural networks. *Briefings in Bioinformatics*, 24(6):bbad414, 2023. doi: 10.1093/bib/bbad414.

Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, et al. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

Jalil Nourisa, Antoine Passemiers, Marco Stock, et al. genernib: a living benchmark for gene regulatory network inference. *bioRxiv*, pp. 2025.02.25.640181, 2025.

Rui Peng, Yuchen Lu, Qichen Sun, et al. Kindle: knowledge-guided distillation for prior-free gene regulatory network inference. In *Advances in Neural Information Processing Systems*, volume 38, pp. 125189–125217, 2026.

Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17:147–154, 2020.

Jiaxin Qi, Hang Li, Yan Cui, Yuhua Zheng, and Jianqiang Huang. Towards universal gene regulatory network inference: Unlocking generalizable regulatory knowledge in single-cell foundation models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Hantao Shu, Jing Zhou, Qiuyu Lian, et al. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1:491–501, 2021.

Marco Stock, Corinna Losert, Matteo Zambon, et al. Leveraging prior knowledge to infer gene regulatory networks from single-cell rna-sequencing data. *Molecular Systems Biology*, 21(3):214–230, 2025.

Binon Teji and Swarup Roy. Gt-grn: a novel graph transformer framework for gene regulatory network inference. *arXiv preprint arXiv:2504.16961*, 2025.

Jialu Wu, Qing Ye, Yilin Wang, et al. Biology-driven insights into the power of single-cell foundation models. *Genome Biology*, 26(1):334, 2025.

Xiaodong Yang, Guohao Liu, Guihai Feng, et al. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 34:830–845, 2024.

Weiming Yu, Zhuobin Chen, Yaohua Hu, et al. Structure-enhanced graph meta learning for few-shot gene regulatory network inference. *Genome Biology*, 26(1):394, 2025a.

Weiming Yu, Zerun Lin, Miaofang Lan, and Le Ou-Yang. GCLink: a graph contrastive link prediction framework for gene regulatory network inference. *Bioinformatics*, 41(3):btaf074, 2025b. doi: 10.1093/bioinformatics/btaf074.

A Datasets and Baseline Settings

Dataset statistics and baseline categories are listed in Tables 7 and 8. Candidate spaces and the separate GeneCompass+DeepSEM protocol are specified in Table 3.

Table 7: Datasets. #GT Pairs = Combined GT edges in our gene selection subgraph.

Species	CT	#Cells	#Genes	#GT Pairs
Human	hESC	758	17,735	322,689
Human	hHep	425	11,515	315,615
Mouse	mESC	421	18,385	256,910
Mouse	mDC	383	7,371	43,601
Mouse	mHSC-E	1,071	4,762	67,599
Mouse	mHSC-GM	889	4,762	70,328
Mouse	mHSC-L	847	4,762	71,978

Table 8: Baselines. †Pretrained embeddings without CT-specific GT training.

Category	Method	Cross-CT
Statistical	Random, PCC, MI	Intra
Tree	GENIE3, GRNBoost2	Intra
DL	DeepSEM	Intra
FM	scGPT, GeneCompass+DeepSEM	Zero-shot†
	GRNformer	Cross-CT
Ours	**GeneBridge**	**Cross-CT**

B Implementation and Algorithm

B.1 GeneCompass Fine-Tuning

We fine-tune GeneCompass (12-layer BERT-style Transformer, d=768, pretrained on human+mouse ∼50K genes) via self-supervised masked language modeling (MLM) on pooled expression data from all K cell types. The bottom 9 Transformer layers are frozen; only the top 3 layers and token embeddings are trained. Fine-tuning runs for 1,000 steps with learning rate 5×10^{-6}, batch size 8, and MLM probability 0.15. No ground truth regulatory annotations are accessed during fine-tuning.

B.2 Gene Selection Parameters

Co-expression coarse filter: n_{co}=100 genes per TF (Pearson correlation, absolute value). Attention fine filter: n_{att}=50 genes per TF, extracted from the final Transformer layer averaged over all H=12 attention heads.

B.3 Graph Transformer Encoder

The encoder consists of L=3 layers with H=4 attention heads, hidden dimension d_h=128, key dimension d_{head}=32, and dropout rate 0.1. We use eight eigenvectors of the symmetrized graph Laplacian after discarding the first eigenvector. For head h at layer l, attention logits are modulated by learned edge features:

$$e^h_{uv} = (\mathbf{Q}^h\mathbf{h}^{(l)}_u)^\top(\mathbf{K}^h\mathbf{h}^{(l)}_v)/\sqrt{d_{\text{head}}}\cdot s^h_{uv}, \tag{7}$$

where s^h_{uv} is the sum of the projected edge-feature components for head h. This is followed by softmax, residual aggregation, BatchNorm, and FFN (d_h→$2d_h$→d_h, ReLU).

The EdgePredictor decoder is a 3-layer MLP ($2d_h \to d_h \to d_h/2 \to 1$) with BatchNorm, ReLU, Dropout(0.1), and sigmoid output, scoring node pairs from concatenated encoder embeddings $\mathbf{Z} \in \mathbb{R}^{N\times d_h}$. Total model size is approximately 0.84M trainable parameters.

B.4 Training Hyperparameters

BCE loss, AdamW optimizer (lr $= 10^{-3}$, weight decay $= 10^{-4}$), batch size 4096, negative sampling ratio NR=7, early stopping patience 3. Soft-label weight γ=0.3 for mouse and cross-species settings, γ=0 for human. KEGG fusion weight λ=0.3 for human and cross-species settings, λ=0 for mouse. All experiments use a single NVIDIA RTX 5090 GPU.

B.5 Inference and Evaluation Protocol

Target expression data are used for MLM adaptation, gene selection, graph construction, and contextual node features; target regulatory labels are excluded from training. Inference scores the full candidate space of the held-out CT subgraph.

Main LOO results use full-space evaluation and report mean ± standard deviation over ten seeds: {42, 123, 234, 789, 55, 100, 200, 345, 567, 890}. T100 and T500 count reference positives among the top-ranked predictions; Rand T100 is $100p_{\text{rand}}$. Ablation-specific seeds, edge weights, and evaluation sampling ratios are stated in Appendix D. Sampled AUROC/AUPRC should not be directly compared with full-space main results; T100 uses the full candidate space throughout.

B.6 Algorithm Pseudocode

Stage 1: Unsupervised Semantic Graph Construction
for each cell type $k \in \{1,\dots,K\}$ **do**
// Gene Selection (Module I)
$\mathcal{C}_t \leftarrow$ CoarseFilter$(t, \mathbf{X}_k, n_{\text{co}}$=100) // Pearson corr., per TF
$\mathcal{A}_t \leftarrow$ AttentionFilter$(t, \mathbf{X}_k, \mathcal{C}_t, f_{\text{GC}}, n_{\text{att}}$=50)
$\mathcal{V}^{\text{sel}}_k \leftarrow \mathcal{T}_k \cup \bigcup_{t\in\mathcal{T}_k}\mathcal{A}_t$

end for

// Graph Construction (Module II)

$\mathcal{V} \leftarrow \bigcup_{k=1}^{K} \mathcal{V}_k^{\text{sel}}$ // unified node set

$\mathcal{E} \leftarrow \bigcup_{k=1}^{K} \{(t, g) \mid t \in \mathcal{T}_k, g \in \mathcal{V}_k^{\text{sel}} \setminus \{t\}\}$

$w_{ij} \leftarrow \lambda \cdot \text{Jaccard}(i, j) + (1-\lambda)\frac{1}{2}(\cos(\mathbf{E}_i^{\text{pt}}, \mathbf{E}_j^{\text{pt}}) + 1)$

// λ=0.3 human, λ=0 mouse

// Hybrid Node Embeddings

for each CT k **do**

$\mathbf{h}_i^{(k)} \leftarrow f_{\text{GC}}(i; \mathbf{X}_k)$ if $i \in \mathcal{V}_k$, else $\mathbf{E}_i^{\text{pt}}$

$\mathbf{x}_i^{(k)} \leftarrow \mathbf{h}_i^{(k)} \parallel \mathbb{1}_{\text{TF}}(i)$

end for

Stage 2: Supervised Zero-Shot Training

for each training CT $k \in \mathcal{K}_{\text{train}}$ **do**

// Soft-label construction (soft-label equation in main text)

$\ell_{uv}^{(k)} \leftarrow 1.0$ if $(u, v) \in \mathcal{G}_k$

$\ell_{uv}^{(k)} \leftarrow \gamma$ if $(u, v) \in \bigcup_{j \in \mathcal{K}_{\text{train}} \setminus \{k\}} \mathcal{G}_j \setminus \mathcal{G}_k$

$\ell_{uv}^{(k)} \leftarrow 0.0$ otherwise (randomly sampled negatives, NR=7)

end for

Train f_θ on $\{(\mathbf{X}_k, \boldsymbol{\ell}^{(k)})\}_{k \in \mathcal{K}_{\text{train}}}$ with BCE loss,

AdamW (lr=10^{-3}, wd=10^{-4}), batch size 4096, early stopping

Stage 3: Zero-Shot Inference

$\hat{\mathbf{y}} \leftarrow f_\theta(\mathbf{X}_{\text{test}})$ on held-out CT subgraph

Evaluate on full candidate edge space $\mathcal{E}_{\text{test}}$

C Full Leave-One-Cell-Type-Out Results

All results report mean ± standard deviation over ten seeds using full-space evaluation (Appendix B). Each held-out CT is evaluated after training on the remaining CTs in the specified setting.

C.1 Human 2CT LOO

Table 9: Human 2CT LOO (γ=0, KEGG:Cosine=0.3:0.7). Arrows indicate source and target CTs. Each setting lists the mean followed by ±std.

Setting	AUROC	AUPRC	Rand T100	T100	T500
hESC→hHep	0.791 ±0.008	0.253 ±0.020	3.2	82.0 ±4.4	367 ±16
hHep→hESC	0.790 ±0.016	0.242 ±0.020	2.7	89.0 ±4.4	387 ±21

C.2 Mouse 5CT LOO

Table 10: Mouse 5CT LOO (γ=0.3, KEGG:Cosine=0:1). Rows identify the held-out CT; each mean is followed by ±std.

Holdout	AUROC	AUPRC	Rand T100	T100	T500
mESC	0.537 ±0.021	0.083 ±0.006	7.1	14.3 ±5.8	74 ±26
mDC	0.650 ±0.011	0.038 ±0.002	1.7	32.5 ±4.5	110 ±17
mHSC-E	0.783 ±0.016	0.398 ±0.032	7.5	93.8 ±5.7	408 ±12
mHSC-GM	0.783 ±0.028	0.398 ±0.026	7.4	91.1 ±4.3	408 ±9
mHSC-L	0.774 ±0.027	0.388 ±0.032	7.4	90.8 ±4.5	411 ±12

C.3 Cross-Species 4CT LOO

Table 11: Cross-species 4CT LOO (γ=0.3, KEGG:Cosine=0.3:0.7). Training uses the other three CTs; each mean is followed by ±std.

Holdout	Species	AUROC	AUPRC	Rand T100	T100	T500
hESC	Human	0.600 ±0.017	0.051 ±0.007	2.7	35.4 ±13.0	142 ±65
hHep	Human	0.621 ±0.025	0.058 ±0.009	3.2	18.5 ±5.0	78 ±30
mESC	Mouse	0.549 ±0.022	0.092 ±0.010	7.1	44.8 ±24.1	183 ±99
mDC	Mouse	0.655 ±0.009	0.032 ±0.002	1.7	5.2 ±2.4	33 ±8

D Ablation Studies

These experiments use separate runs from the multi-seed main results. Each subsection specifies its settings; differences across these settings should not be interpreted as controlled comparisons.

D.1 Encoder Architecture

The encoder comparison uses seed 42 and full-space evaluation. Hybrid node embeddings, soft-label training, and evaluation protocol are held constant. GraphTransformer uses three layers, four heads, and edge-feature-modulated attention; GCN uses three layers with edge-weighted aggregation; GAT uses three layers with 4/4/1 heads and no edge weights. The comparison therefore evaluates architecture and edge-weight handling together.

Table 12: Per-dataset encoder comparison (seed 42, full-space evaluation).

Dataset	Architecture	AUROC	AUPRC	T100
mESC	GraphTransformer	0.508	0.077	29
	GCN	0.473	0.066	9
	GAT	0.519	0.072	0
mDC	GraphTransformer	0.627	0.037	34
	GCN	0.612	0.026	2
	GAT	0.617	0.028	6
mHSC-E	GraphTransformer	0.789	0.418	100
	GCN	0.553	0.089	3
	GAT	0.719	0.169	55
mHSC-GM	GraphTransformer	0.823	0.426	92
	GCN	0.564	0.099	70
	GAT	0.592	0.098	5
mHSC-L	GraphTransformer	0.807	0.392	98
	GCN	0.569	0.096	67
	GAT	0.631	0.153	69
hESC→hHep	GraphTransformer	0.789	0.260	86
	GCN	0.625	0.066	12
	GAT	0.521	0.034	0
hHep→hESC	GraphTransformer	0.778	0.249	88
	GCN	0.655	0.053	13
	GAT	0.529	0.032	8

D.2 SOFT-LABEL WEIGHT

We vary $\gamma \in \{0, 0.1, 0.3, 0.5, 0.7\}$ on mouse 5CT LOO with NR=7 and KEGG:Cosine=0.5:0.5 (single seed). AUROC/AUPRC use Eval NR=7; T100 uses the full candidate space. The highest average T100 is 68.8 at γ=0.3, compared with 59.0 at γ=0.

Table 13: Soft-label weight sweep: averages over five mouse holdouts (single seed, Eval NR=7).

γ	Avg AUROC	Avg AUPRC	Avg T100
0 (ablation)	0.834	0.517	59.0
0.1	0.849	0.542	66.0
0.3	**0.858**	**0.563**	**68.8**
0.5	0.843	0.547	67.6
0.7	0.839	0.540	67.6

Table 14: Per-CT soft-label comparison: γ=0 versus γ=0.3 under the same settings.

Holdout	γ=0 T100	γ=0.3 T100	γ=0 AUROC	γ=0.3 AUROC
mESC	12	27	0.680	0.651
mDC	25	35	0.820	0.828
mHSC-E	85	95	0.886	0.917
mHSC-GM	88	94	0.917	0.948
mHSC-L	85	93	0.868	0.944
Avg	59.0	**68.8**	0.834	0.858

D.3 Negative Sampling Ratio

We vary training NR $\in \{1, 3, 5, 7, 10\}$ for mouse 5CT LOO (γ=0.3) and human 2CT LOO (γ=0), both with pure cosine edge weights. These are single-seed runs with Eval NR=1 for AUROC and full-space T100. NR=7 gives the highest human T100 (91.0) and competitive mouse T100 (65.4).

Table 15: Negative sampling sweep: averages across holdout CTs. Bold marks column maxima.

	Mouse 5CT		Human 2CT	
NR	Avg AUROC	Avg T100	Avg AUROC	Avg T100
1	0.702	**69.4**	0.780	75.5
3	0.688	63.4	0.796	73.0
5	0.711	62.0	0.779	79.0
7	0.699	65.4	0.773	**91.0**
10	**0.718**	64.4	**0.796**	87.5

Table 16: Mouse per-CT negative sampling sweep. Each holdout has a T100 row followed by an AUROC row.

Holdout	NR=1	NR=3	NR=5	NR=7	NR=10
mESC	63	12	8	43	5
	0.584	0.513	0.514	0.509	0.576
mDC	27	35	41	15	33
	0.641	0.644	0.649	0.656	0.663
mHSC-E	91	94	89	94	94
	0.771	0.788	0.807	0.817	0.771
mHSC-GM	74	92	89	85	93
	0.729	0.738	0.762	0.735	0.781
mHSC-L	92	84	83	90	97
	0.785	0.757	0.825	0.780	0.797

D.4 SEMANTIC EDGE WEIGHTS

We vary the KEGG fusion weight $\lambda \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ with NR=7 and γ=0.3 (mouse) or 0 (human). These single-seed runs use Eval NR=7 for AUROC/AUPRC and full-space T100. Table 17 summarizes four settings; Tables 18 and 19 cover all six. Pure cosine gives the highest reported mouse average T100; for human, λ=0.3 gives the highest AUPRC and near-optimal T100.

Table 17: Edge-weight summary for $\lambda \in \{0, 0.1, 0.3, 0.5\}$. Bold marks maxima within the displayed settings for each species.

Mouse 5CT			
λ	Avg AUROC	Avg AUPRC	Avg T100
0	0.833	0.543	**71.2**
0.1	0.854	0.555	66.4
0.3	0.842	0.549	66.4
0.5	0.855	0.564	66.4
Human 2CT			
λ	Avg AUROC	Avg AUPRC	Avg T100
0	0.920	0.695	79.5
0.1	0.914	0.693	**90.0**
0.3	**0.924**	**0.704**	89.0
0.5	0.911	0.690	79.0

Table 18: Mouse per-CT edge-weight sweep. Each holdout has a T100 row followed by an AUROC row; bold marks its highest T100.

	λ=0	λ=0.1	λ=0.3	λ=0.5	λ=0.7	λ=0.9
Holdout	0:1	0.1:0.9	0.3:0.7	0.5:0.5	0.7:0.3	0.9:0.1
mESC	**37**	13	17	17	9	9
	0.595	0.684	0.573	0.623	0.584	0.670
mDC	35	30	33	32	30	**39**
	0.770	0.781	0.829	0.841	0.850	0.825
mHSC-E	97	95	89	**99**	98	95
	0.950	0.937	0.926	0.924	0.931	0.947
mHSC-GM	95	97	**98**	90	89	92
	0.914	0.925	0.933	0.944	0.913	0.947
mHSC-L	92	**97**	95	94	96	94
	0.938	0.945	0.950	0.944	0.940	0.893

Table 19: Human edge-weight sweep. The selected setting is λ=0.3; bold metric values mark column maxima for the three average metrics.

λ	Avg AUROC	Avg AUPRC	Avg T100	hHep T100	hESC T100
0	0.920	0.695	79.5	79	80
0.1	0.914	0.693	**90.0**	84	96
0.3	0.924	**0.704**	89.0	80	98
0.5	0.911	0.690	79.0	67	91
0.7	0.914	0.693	80.0	84	76
0.9	**0.925**	0.697	88.0	81	95