

Supplementary Material for: Genes2Reg – Constructing a Biologically-Faithful Gene Coupling Prior via Deep Adaptive Fusion of Heterogeneous Knowledge Sources

This document provides supplementary tables and figures supporting the main manuscript. Section [S1](#) details the encoding of the biological knowledge sources (main text § 3.1–3.2). Section [S2](#) reports the embedding-health diagnostics and held-out regulatory generalization analyses underlying main text § 4.2. Section [S3](#) reports the coupling-structure diagnostics and sparsification sensitivity supporting the choice of $k = 50$ (main text § 4.1–4.2). Section [S4](#) reports the full synthetic benchmark summarized in main text § 4.3. Section [S5](#) documents the end-to-end workflow implemented in the released source code.

S1 Encoding of biological knowledge sources

Table S1 lists the sparse adjacency and bipartite incidence matrices constructed from the five knowledge sources of main text Table 1, in the notation of main text §3.1.2–3.2. These matrices are the inputs to the heterogeneous graph $\mathcal{G}$ consumed by the HGAT encoder.

Table S1: Source adjacency matrices. Let G denote the set of n genes, and let P , T , R , C denote pathway terms, transcription factors, miRNAs, and protein complexes. Directed regulatory relations (TF activation, TF repression, miRNA) are kept as separate bipartite adjacencies to preserve regulatory asymmetry (main text §3.2.1).

Matrix	Dimensions	Description
$A^{(\text{ppi})}$	$\mathbb{R}_{\geq 0}^{n \times n}$	Symmetric gene–gene PPI
$B^{(p)}$	$\mathbb{R}_{\geq 0}^{n \times p}$	Gene–pathway (Reactome)
$B^{(t_+)}$	$\mathbb{R}_{\geq 0}^{n \times t}$	TF activation (DoRothEA A–C)
$B^{(t_-)}$	$\mathbb{R}_{\geq 0}^{n \times t}$	TF repression (DoRothEA A–C)
$B^{(r)}$	$\mathbb{R}_{\geq 0}^{n \times r}$	miRNA–gene (OmniPath)
$B^{(c)}$	$\mathbb{R}_{\geq 0}^{n \times c}$	Gene–complex (CORUM)

S2 Embedding diagnostics and held-out generalization

Table S2 lists the benchmark models compared in main text §4.2 together with the embedding-health diagnostics referenced there: our prior is the only embedding that is simultaneously non-saturated ($\text{Sat.} = 0.00$), low-anisotropy, and high-coverage, whereas the diffusion-based methods saturate and the text- and integration-based embeddings are anisotropic.

Table S2: Benchmark models and embedding-health diagnostics. Models compared in the gene representation benchmark (main text §4.2), with embedding-health diagnostics. *Saturation* is the fraction of nearest-neighbor cosine similarities exceeding 0.99: connected genes collapse onto near-identical vectors. *Aniso.* is the norm of the mean (L2-normalized) embedding vector. *Iso.* is the fraction of genes with no embedding.

Model	Emb. size	Sat.↓	Aniso.↓	Iso.↓	Description
Ours	128	0.00	0.30	0.09	HGAT with reconstruction and contrastive objectives.
Gene2Vec	200	0.00	0.34	0.19	Skip-gram embeddings from gene co-expression patterns.
GenePT	1,536	0.00	0.63	0.09	LLM embeddings of NCBI gene summary texts.
Mashup	128	0.03	0.85	0.28	Diffusion-based integration of multiple gene networks.
SVD	128	0.14	0.31	0.15	Truncated SVD over bipartite incidence matrices.
GG-kernel	128	0.16	0.49	0.15	Spectral decomposition of summed co-membership matrix.
SNF	128	0.34	0.35	0.24	Iterative cross-diffusion fusion of source similarity matrices.
Diffusion	128	0.30	0.43	0.15	Random-walk multi-hop diffusion over source-graph union.

Table S3 compares all priors on the three regulatory sources withheld from training (TF activation, TF repression, miRNA), complementing the per-hop analysis of main text Fig. 3 and Table 3. At matched sparsity (every prior trimmed to a common edge budget so the comparison is not confounded by density), Genes2Reg attains the largest separation, roughly twice the next-best prior.

Table S3: Held-out regulatory generalization by prior. Mean Cohen’s- d separation Δ on the three regulatory sources held out from training (TF activation, TF repression, miRNA), for the dense embedding and at matched sparsity (every prior trimmed to a common edge budget so the comparison is not confounded by density). Genes2Reg attains the largest matched-sparsity separation, roughly twice the next-best prior.

Prior	Δ (matched sparsity)	Δ (dense embedding)
Genes2Reg	0.430	1.520
Diffusion	0.215	1.438
GenePT	0.191	0.869
SNF	0.143	0.620
SVD	0.083	0.838
Gene2Vec	0.053	0.519
GG-kernel	0.047	1.154
Mashup	−0.009	0.714

Table S4 verifies that the held-out separation is not an artifact of the sparsification level: the signal remains positive at every $k \in \{10, 25, 50, 100\}$.

Table S4: Sensitivity of held-out separation to sparsification. Hop-1 $\Delta_{s,1}$ measured on the sparsified coupling S_k for the three regulatory sources withheld from training. The signal is positive at every k ; values increase with k because held-out pairs pruned at aggressive sparsification score zero. The primary analysis (Fig. 3 of the main text) uses the dense embedding and is k -independent.

Held-out source	$k=10$	$k=25$	$k=50$	$k=100$
TF activator	0.051	0.088	0.152	0.210
TF repressor	0.070	0.141	0.173	0.276
miRNA	0.041	0.059	0.083	0.123

S3 Coupling-structure diagnostics

Table S5 reports the topological diagnostics of the coupling graph across sparsification levels, supporting two claims of the main text: the choice of $k = 50$ as the knee where connectivity saturates at minimal density (main text § 4.1), and the markedly higher robustness of our prior’s connectivity under aggressive sparsification relative to spectral and diffusion baselines (main text § 5).

Table S5: Coupling-structure diagnostics across sparsification levels k (top- k neighbors retained per gene). For each k we report the number of nonzero edges (nnz), the fraction of genes in the largest connected component (CC↑; higher indicates a globally coherent, non-fragmented prior), and the degree skew (max/median degree, Skew↓; higher indicates hub domination). These metrics are topological and invariant to edge-weight normalization. Our prior maintains the highest connectivity across all k and remains robust under aggressive sparsification (largest CC 0.73 at $k=10$ versus 0.20–0.43 for spectral and diffusion baselines), while Mashup exhibits monotonically worsening hub domination as k grows.

Model	$k=10$			$k=25$			$k=50$			$k=100$		
	nnz	CC↑	Skew↓	nnz	CC↑	Skew↓	nnz	CC↑	Skew↓	nnz	CC↑	Skew↓
Ours	84,399	0.73	2.8	193,939	0.88	3.2	389,817	0.90	3.0	808,983	0.91	2.8
SVD	87,053	0.25	3.7	217,505	0.57	3.7	490,837	0.73	2.6	1,136,199	0.81	1.6
GG-kernel	92,703	0.43	3.7	215,299	0.69	3.2	431,751	0.78	3.0	884,665	0.81	2.7
Diffusion	83,419	0.20	5.5	204,229	0.54	4.3	422,783	0.65	3.2	870,967	0.74	2.7
SNF	104,767	0.31	2.8	242,005	0.60	2.6	475,891	0.72	2.3	956,401	0.75	2.2
Gene2Vec	92,287	0.71	2.8	220,703	0.80	2.6	452,931	0.81	2.3	950,121	0.81	2.1
Mashup	71,939	0.36	5.5	122,667	0.58	8.7	170,239	0.62	12.8	234,023	0.67	16.8
GenePT	132,959	0.72	1.6	308,829	0.87	1.6	609,845	0.89	1.6	1,232,121	0.90	1.6

S4 Synthetic benchmark

Table S6 reports the full per-method results of the controlled simulation of main text §4.3, from which the paired across-seed contrasts of main text Table 4 are computed. Consistent with the $p \gg N$ regime, AUC is saturated across all methods and the prior’s value surfaces in coefficient quality (recovery by hop distance, alignment with β^* , smoothness, and stability), bracketed by the oracle G^* and the shuffled control.

Table S6: Synthetic benchmark, mean over 10 seeds ($p = 400$, $n_{\text{train}} = 120$). $\text{rec}@d$: coefficient recovery at hop distance d (higher better); smoothness $\beta^\top L_{\text{gt}} \beta$ lower-is-better. AUC is saturated; the prior’s value is in coefficient quality. Paired across-seed contrasts between these rows are reported in Table 4 of the main text.

Method	AUC	$\cos(\beta, \beta^*)$	TPR	rec@d1	rec@d2	rec@d3	smooth↓	stab
Lasso	0.735	0.441	0.335	0.620	0.508	0.430	0.843	0.333
ElasticNet	0.744	0.463	0.352	0.664	0.554	0.471	0.799	0.385
Lap: shuffled (control)	0.711	0.386	0.368	0.696	0.568	0.481	0.790	0.545
Lap: hop-1 edges (1 view)	0.726	0.415	0.429	0.649	0.543	0.503	0.506	0.625
Lap: diffusion (1 view)	0.734	0.433	0.409	0.693	0.591	0.526	0.569	0.600
Lap: diffusion (union views)	0.737	0.442	0.388	0.716	0.607	0.545	0.610	0.599
Lap: adaptive fuse + diffuse	0.746	0.457	0.429	0.700	0.601	0.545	0.518	0.616
Lap: oracle G^* (true L)	0.742	0.457	0.502	0.629	0.519	0.508	0.281	0.692

S5 End-to-end workflow

Figure S1 summarizes the end-to-end Genes2Reg workflow as four functional stages, from knowledge acquisition to downstream inference. The released implementation (<https://github.com/JMOriggi/Genes2Reg>) follows the same four stages and reproduces all results reported in the main text and in this document.

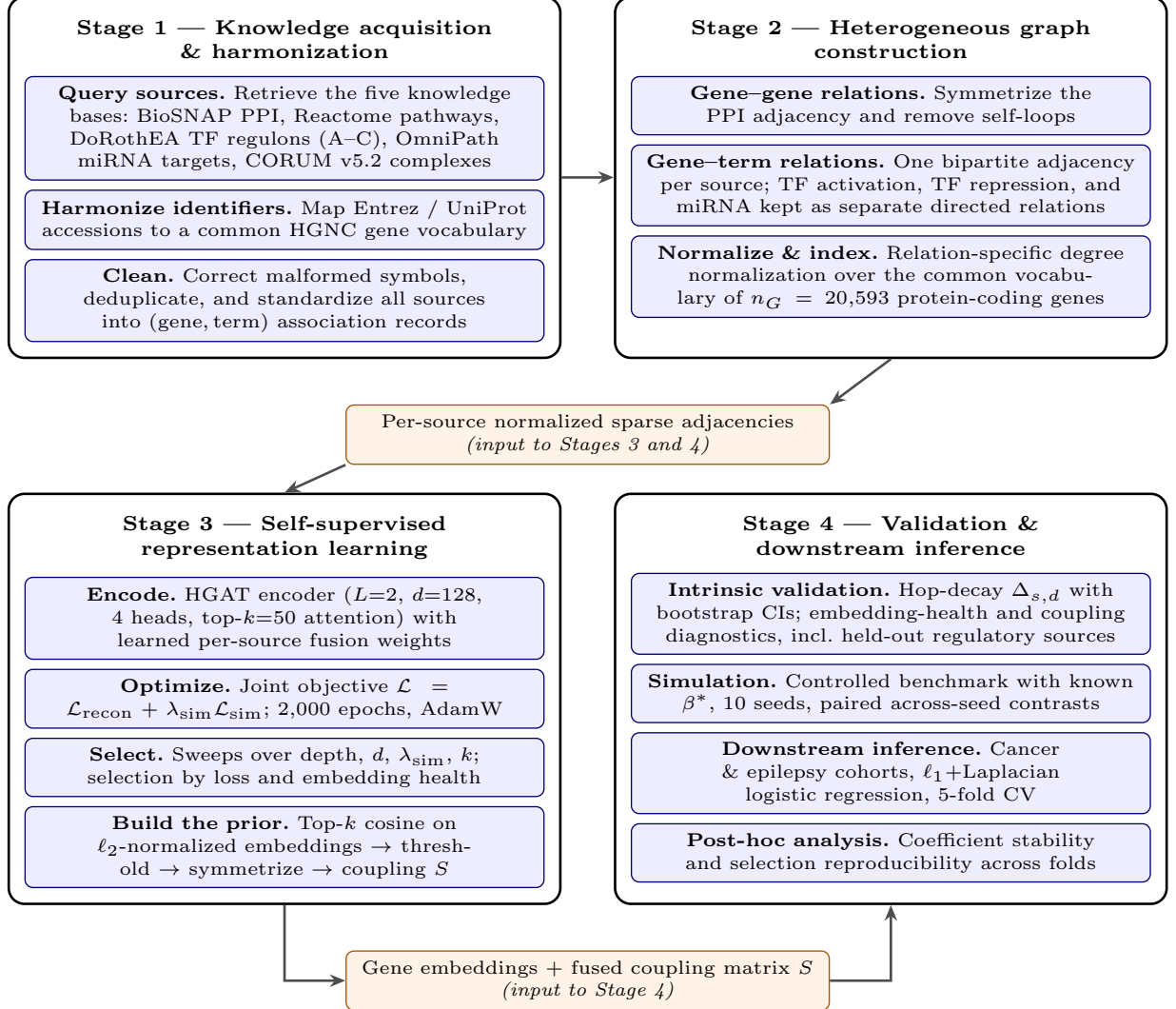


Fig. S1: Genes2Reg end-to-end workflow. The framework proceeds through four functional stages arranged in two rows; blue sub-blocks list the principal steps of each stage, orange blocks the intermediate outputs passed between stages. Stage 1 acquires the five biological knowledge bases (main text Table 1) and harmonizes all identifiers into a common gene vocabulary. Stage 2 encodes each source in its natural relational form, producing the per-source sparse adjacency and bipartite matrices of Table S1 with relation-specific normalization (main text § 3.2). Stage 3 trains the HGAT encoder under the self-supervised objective (main text § 3.2.3), performs the hyperparameter sweeps of main text § 4.1, and projects the learned embeddings into the fused symmetric coupling matrix S . Stage 4 validates the prior intrinsically, on simulated data with known ground truth, and in downstream Laplacian-regularized inference on the cancer and epilepsy cohorts, followed by stability and reproducibility analyses.