Incorporating Contextual Information into KGWAS for Interpretable GWAS Discovery

Cheng Jiang[1§], Brady Ryan[1§], Megan Crow[2], Kipper Fletez-Brant[2], Kashish Doshi[2], Sandra Melo-Carlos[2], Kexin Huang[3], Heming Yao[2#], Burkhard Hoeckendorf[2#], and David Richmond[2#]

[1]University of Michigan [2] gRED, Genentech [3] Stanford University
[§] Work conducted during an internship at Genentech [#]{yao.heming, hoeckendorf.burkhard, richmond.david}@gene.com

Biology Research | AI Development

We present **context-aware KGWAS** and show that **cell-type-specific functional genomics** from Perturb-seq can substantially **improve both the power and interpretability of GWAS discovery.** As Perturb-seq datasets become increasingly available, this framework can be extended to diverse tissues by matching disease traits to corresponding cell-type-specific Perturb-seq atlases, moving genetic discovery toward a more context-driven paradigm.

Overview

Genome-wide association studies (GWAS) identify disease-associated variants. **KGWAS** is a geometric deep learning method that improves detection power by leveraging prior biological knowledge, and provides mechanistic insights that link variants to genes and programs.

Perturbation screens provide cell type-specific insights into gene-gene relationships with direct experimental evidence.

A key question is **whether gene interactions from perturbation screens of relevant cell lines improve the performance and interpretability of KGWAS.**

In this study, we **sparsify** the Knowledge Graph (KG) constructed in KGWAS and further integrate perturbation data as **contextual information** to make the general-purpose KG more trait-specific. The resulting model, which we refer to as **context-aware KGWAS**, yields more consistent disease-critical networks, **providing deeper, actionable insights into the mechanisms underlying complex traits.**

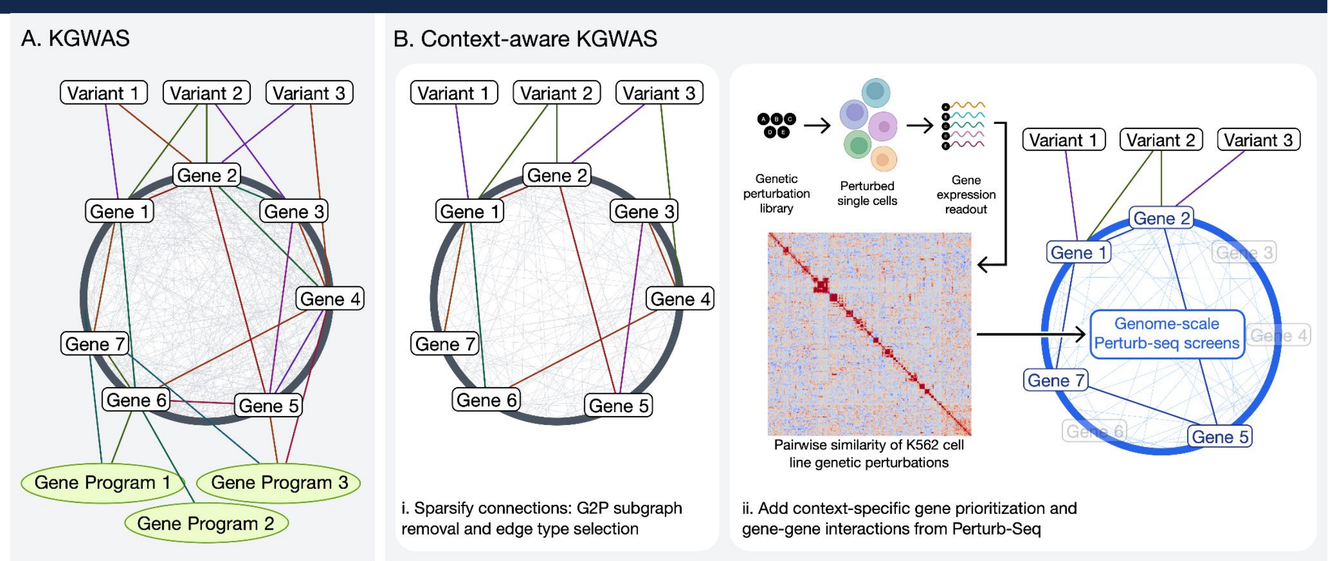

FIgure 1: Knowledge graph construction. A. The original KG in KGWAS; B. Our extension to KGWAS: removing gene-program edges, sparsifying remaining connections, and replacing gene-gene edges with contextually relevant relationships derived from Perturb-seq.

Experimental Setting

Perturb-seq data: We use data from a genome-scale experimental Perturb-seq screen targeting all essential and expressed genes in the human leukemia-derived **K562 cell line** using CRISPR interference (Replogle et al., 2022). For each perturbation, we compute log-fold changes (LFCs) and apply independent component analysis (ICA) for dimensionality reduction. We then calculate cosine similarity between each pair of target genes. Pairs with extreme similarity values are used as context-specific G2G edges.

GWAS data: We focus on three traits: Mean Corpuscular Hemoglobin (MCH), Immature Reticulocyte Fraction (IRF), and Red Cell Distribution Width (RDW). These traits were previously selected by Ota et al. (2025) because open chromatin regions in **K562** exhibit significant heritability enrichment for these traits, consistent with the erythroid lineage of K562.

Model training: Each trait is subsampled to include 1000, 2500, 5000, 7500, 10000, and 50000 individuals and the corresponding χ^2 association statistics are computed. For each trait and cohort size combination, a heterogeneous graph attention network (GAT) is trained to predict GWAS χ^2 association statistics from the learned variant embeddings. The resulting predicted χ^2 statistics are then used to generate a new set of association p-values.

Evaluation Criterion: We define the ground-truth associations as the 100 most significant loci from GWAS performed on the full UK Biobank cohort. We evaluate the ability of KGWAS to correctly identify ground-truth associations from sub-sampled cohorts. We report the number of loci in the top 100 independent risk loci identified by each method that overlap with loci discovered in the full cohort.

Knowledge Graph Sparsification

While KGWAS achieves good performance, the constructed KG is large and highly redundant, which may limit the generalizability and interpretability of disease-critical networks inferred from edge attention scores. To evaluate the contribution of specific edge types, we perform the following ablation studies:

A. **Disentangling gene-to-gene (G2G) and gene-to-program (G2P) contributions**
B. **Edge type selection**

We hypothesize that low-specificity edges dilute the message-passing signal. To test this, we define a high-confidence V2G subset restricted to cis-regulatory relationships with strong functional evidence, and a G2G subset that retains edge types with more than 10,000 connections.

We show that a two-layer GAT performs well without G2P connections, and that restricting the graph to the high-confidence V2G and G2G subsets maintains performance while substantially reducing graph size.

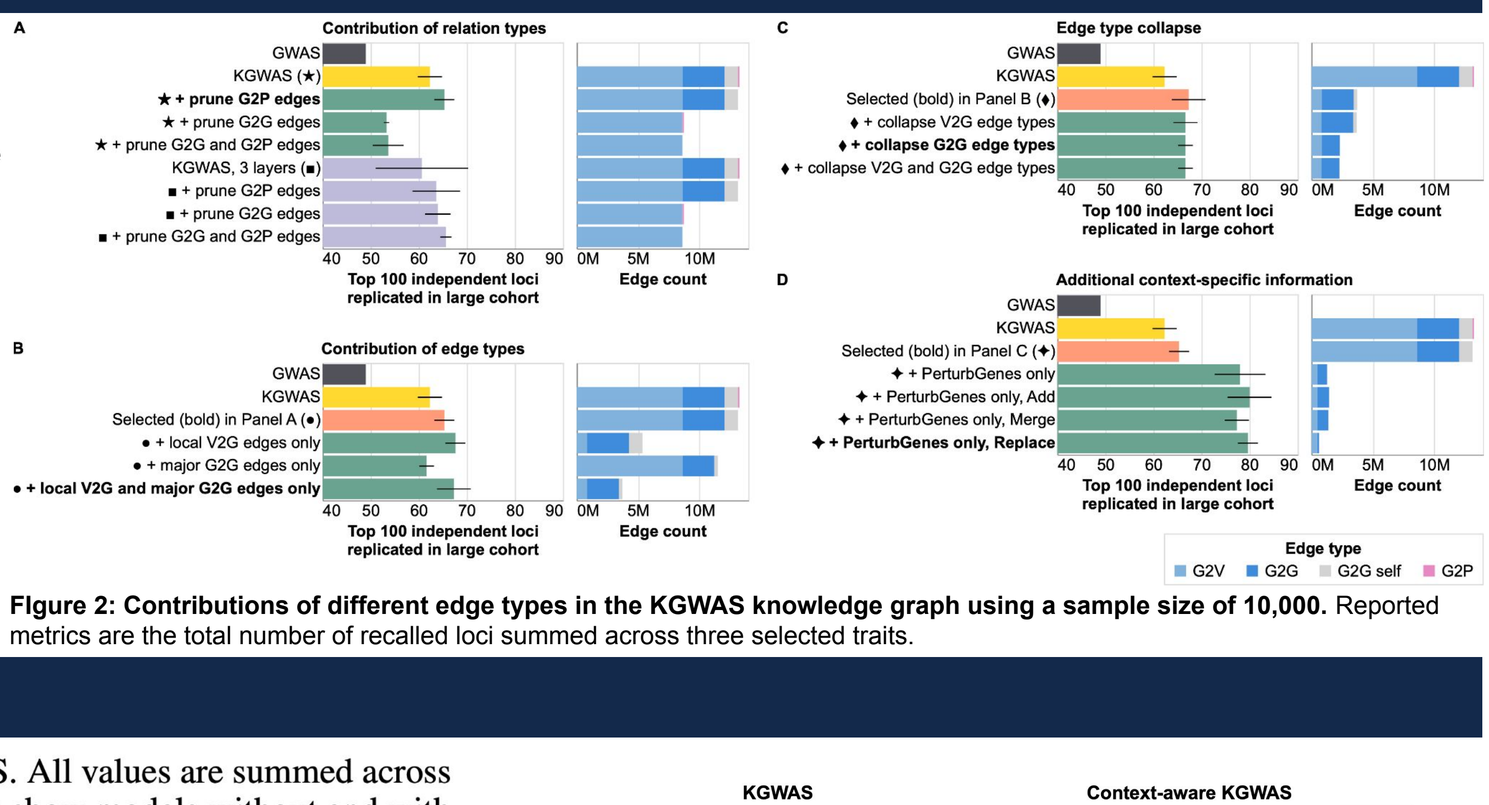

FIgure 2: Contributions of different edge types in the KGWAS knowledge graph using a sample size of 10,000. Reported metrics are the total number of recalled loci summed across three selected traits.

Results and Discussions

Table 1: Top 100 independent loci replicated in large cohort GWAS. All values are summed across three selected traits: MCH, IRF, RDW. The top and bottom groups show models without and with contextual information, respectively.

	# Edges	Small cohort sample size 1,000	2,500	5,000	7,500	10,000	50,000
GWAS	0	14	15	29	36	49	201
KGWAS	12.1M	27.00 (1.00)	24.67 (1.53)	37.67 (2.08)	55.00 (1.73)	62.33 (2.52)	221.00 (4.36)
+ Remove gene programs	12.1M	26.33 (1.53)	24.00 (2.65)	36.33 (1.53)	53.33 (2.31)	65.33 (2.08)	220.00 (4.36)
+ Prune edge types	3.4M	25.67 (3.06)	26.33 (2.52)	46.33 (2.52)	**58.33 (2.52)**	67.33 (3.51)	221.00 (2.00)
+ Collapse GG edges	2.3M	25.67 (3.06)	26.00 (2.00)	46.33 (0.58)	55.00 (2.65)	66.67 (1.53)	218.67 (2.52)
+ Prioritize genes	1.3M	**32.67 (3.79)**	**36.00 (2.65)**	52.33 (1.53)	55.00 (1.00)	78.00 (5.29)	226.33 (0.58)
+ Replace with perturb graph	625K	30.67 (2.31)	34.67 (4.04)	**53.00 (2.00)**	**58.33 (2.08)**	**79.67 (2.08)**	**229.00 (3.61)**

Table 2: Ablation stuides on edge contributions to context-aware KGWAS. Top 100 independent loci replicated in large cohort GWAS over MCH, IRF, RDW traits are reported.

	# Edges	Small cohort sample size 1,000	2,500	5,000	7,500	10,000	50,000
Context-aware KGWAS (⋆)	625K	30.67 (2.31)	34.67 (4.04)	53.00 (2.00)	58.33 (2.08)	79.67 (2.08)	229.00 (3.61)
⋆ with randomized G2G	863K	24.67 (2.31)	27.33 (1.53)	47.33 (2.08)	54.33 (1.15)	67.33 (4.73)	221.33 (3.06)
⋆ with dropped G2G	747K	27.67 (2.52)	29.67 (0.58)	38.33 (2.31)	51.00 (1.73)	72.00 (2.00)	224.33 (4.73)
⋆ with randomized V2G and G2G	863K	19.67 (1.15)	26.33 (4.93)	43.00 (12.74)	50.67 (1.52)	61.67 (5.69)	211.00 (3.00)

We construct the final context-aware KG by combining multiple strategies: (1) removing all gene–program nodes, (2) removing a subset of V2G and G2G edge types, (3) collapsing remaining G2G edges to a single type, (4) removing non-essential and non-expressed genes, and (5) restricting G2G edges to those inferred from Perturb-seq. As shown in Table 1, **these strategies consistently reduce the number of edges while improving recall of loci identified in full-cohort GWAS across a wide range of cohort sizes.** By incorporating context-specific information, such as gene expression and transcriptional-state similarity following genetic perturbation, **we obtain a highly efficient KG with a 19-fold reduction in edge count and significantly improved predictive performance for relevant traits.**

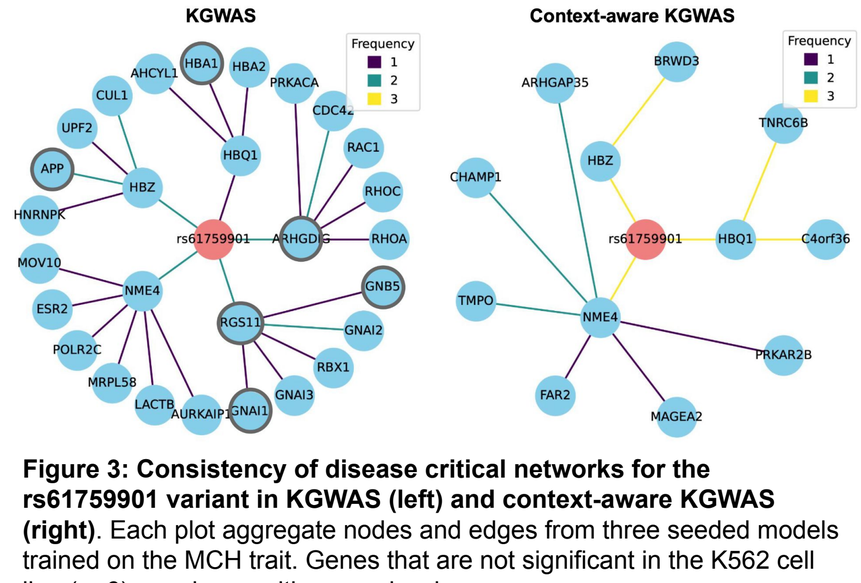

Figure 3: Consistency of disease critical networks for the rs61759901 variant in KGWAS (left) and context-aware KGWAS (right). Each plot aggregate nodes and edges from three seeded models trained on the MCH trait. Genes that are not significant in the K562 cell line (n=6) are shown with a gray border.

We hypothesized that the large size of the KG in KGWAS may undermine the reproducibility of inferred disease-critical networks, and that **restricting the KG to contextually relevant information should improve interpretability.** Figure 3 shows examples of disease-critical networks for the rs61759901 variant derived from both KGWAS and context-aware KGWAS. In the original KGWAS graph, limited consistency between V2G and G2G edges leads to many edges appearing only once and includes connections to genes that are not significant in the K562 cell line (shown with a gray border). In contrast, the sparser context-aware KGWAS network shows much more consistent attention patterns.