INFOSEM: A DEEP GENERATIVE MODEL WITH INFORMATIVE PRIORS FOR GENE REGULATORY NETWORK INFERENCE

Tianyu Cui* **Song-Jun Xu*** **Artem Moskalev*** **Shuwei Li***

Tommaso Mansi* **Mangal Prakash*†** **Rui Liao*†**

ABSTRACT

Inferring Gene Regulatory Networks (GRNs) from gene expression data is crucial for understanding biological processes. While supervised models are reported to achieve high performance for this task, they rely on costly ground truth (GT) labels and risk learning gene-specific biases—such as class imbalances of GT interactions—rather than true regulatory mechanisms. To address these issues, we introduce InfoSEM, an unsupervised generative model that leverages textual gene representations as informative priors, improving GRN inference without GT labels. InfoSEM can also integrate GT labels as an additional prior when available, avoiding biases and further enhancing performance. Additionally, we propose a biologically motivated benchmarking framework that better reflects real-world applications such as biomarker discovery and reveals learned biases of existing supervised methods. InfoSEM outperforms existing models by 38.5% across four datasets using textual embeddings prior and further boosts performance by 11.1% when integrating labeled data as priors.

1 INTRODUCTION

Gene Regulatory Networks (GRNs) govern cellular processes by capturing regulatory relationships which are essential for gene expression, differentiation, and identity (Karlebach & Shamir, 2008) with diverse applications, especially in drug discovery (Basso et al., 2005; Dehmer et al., 2013; Ghosh & Basu, 2012). Single-cell RNA sequencing (scRNA-seq) has revolutionized GRN inference by enabling high-resolution profiling of cell-specific regulatory interactions. However, scRNA-seq data are noisy, sparse, and high-dimensional (Pratapa et al., 2020; Wagner et al., 2016; Dai et al., 2024), requiring advanced computational approaches. Methods have evolved from co-expression frameworks (Chan et al., 2017; Kim, 2015) to cutting-edge machine learning (ML) and deep learning (DL) models (Yuan & Bar-Joseph, 2019; Wang et al., 2024; Anonymous, 2024; Shu et al., 2021), with further improvements leveraging complementary but hard to obtain perturbation experiments, RNA velocity, and chromatin accessibility inputs (Chevalley et al., 2022; Atanackovic et al., 2024; Yuan & Duren, 2024).

Machine learning based GRN inference methods can be broadly classified as supervised (SL) or unsupervised (USL). Supervised models use experimentally derived ground truth (GT) labels, such as ChIP-seq data, where known TF-TG interactions guide learning. While they achieve high performance (AUPRC $\sim$0.85) (Chen & Liu, 2022; Wang et al., 2023), their reliance on costly GT labels limits applicability. Unsupervised methods which infer regulatory relationships directly from gene expression data without using any known interactions, present an attractive alternative but typically lag in performance compared to supervised models (Chen & Liu, 2022; Wang et al., 2023; Mao et al., 2023). Bridging this gap requires advancing unsupervised GRN inference methods.

To address this, we introduce InfoSEM, an unsupervised generative framework trained with variational Bayes that leverages textual gene embeddings as informative priors. By integrating prior

*Johnson and Johnson Innovative Medicine. Correspondence to: Tianyu Cui: tcui8@its.jnj.com
†Equal contribution

biological knowledge, InfoSEM substantially improves GRN inference over existing models. Moreover, InfoSEM can incorporate GT labels as an additional prior when available, rather than using them for direct supervision, avoiding dataset (gene-specific) biases and improving performance further.

Beyond model development, reliably evaluating GRN inference methods is equally important. Existing SL based GRN inference benchmarks typically assume all genes are represented in both the training and test sets, and only the regulatory links between them differ (*unseen interactions between seen genes*, Section 4). While suitable for predicting interactions among well-characterized genes, this setup does not reflect many real-world applications such as biomarker discovery and rare cell-type studies (Lotfi Shahreza et al., 2018; Ahmed et al., 2020), which often involve genes whose interactions are completely absent from the training data (*interactions between unseen genes* in Section 4), thus requiring models to generalize beyond the set of genes encountered during training.

To bridge this gap, we propose a biologically motivated benchmarking framework evaluating *interactions between unseen genes*, better aligning with many real-world applications. This shift in evaluation perspective also inadvertently highlights that, in prior benchmarks, supervised models may have relied on gene-specific biases of the dataset, e.g., class imbalance in known interactions for each gene. This underscores the need for careful evaluation to distinguish between performance gains driven by genuine biological insights from scRNA-seq and those influenced by dataset biases.

In summary, our work makes the following contributions:

- We present InfoSEM, an unsupervised generative framework that leverages textual gene embeddings as priors and can seamlessly integrate GT labels as an additional prior (when available), avoiding dataset biases and improving GRN inference.
- We reveal limitations in existing supervised learning based GRN inference benchmarks, showing that supervised models may exploit dataset biases, such as class imbalance of each gene, rather than capturing true biological mechanisms from scRNA-seq.
- We propose a new evaluation framework focused on regulatory *interactions between unseen genes*, better aligning with real-world applications such as biomarker discovery.
- Finally, we show that our InfoSEM improves GRN inference by 38.5% over models without priors and achieves state-of-the-art performance, even surpassing supervised models. Integrating GT labels as an additional prior further boosts performance by 11.1% while mitigating dataset biases.

2 GRN INFERENCE PROBLEM

Let $G = (V, Y)$ represent a GRN, where $V = \{v_1, \ldots, v_P\}$ denotes the set of P genes (nodes), including transcription factors (TFs) and target genes (TGs). The adjacency matrix $Y \in \{0, 1\}^{P \times P}$ encodes regulatory interactions, where $y_{ik} = 1$ indicates that TF v_i regulates TG v_k. We assume access to scRNA-seq data and the goal is to infer the adjacency matrix Y. The scRNA-seq data, comprising P genes and N cells, is represented as $X \in \mathbb{R}^{P \times N}$, where $x_{ij} \in \mathbb{R}$ denotes the expression of gene v_i in cell j, and $\boldsymbol{x}_i$ (the i-th row of X) denotes the expression profile of v_i across all cells.

Supervised GRN Inference When partial information of Y is available from sources like ChIP-seq or databases (Yevshin et al., 2019), GRN inference can be framed as a supervised learning (SL) task, where the model is trained on labeled data with known interactions in Y as labels. The model f_{sl} learns to predict the probability of an interaction y_{ik} being true based on observed data by minimizing a cross-entropy loss, i.e., assuming a Bernoulli likelihood, using aggregated representations of genes i and k, denoted as $\boldsymbol{s}_i$ and $\boldsymbol{s}_k$, which could be a function of their expression profile $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$:

$$p_{ik} = f_{\text{sl}}(\text{agg}(\boldsymbol{s}_i, \boldsymbol{s}_k)), y_{ik} \sim \text{Bernoulli}(p_{ik}). \tag{1}$$

Deep learning methods such as CNNC (Yuan & Bar-Joseph, 2019), DeepDRIM (Chen et al., 2021), and GENELink (Chen & Liu, 2022) leverage CNNs, GNNs, or transformers to aggregate pairwise gene expressions or their deep representations for interaction prediction (Wang et al., 2023; Mao et al., 2023; Xu et al., 2023). The SL framework can also flexibly incorporate pre-trained embeddings, such as BioBERT embeddings in scGREAT (Wang et al., 2024) or scBERT embeddings combined with GENELink representations in scTransNet (Kommu et al., 2024).

Unsupervised GRN Inference Unsupervised learning (USL) infer gene relationships without labeled data, relying solely on gene expression data and more closely mirroring real-world

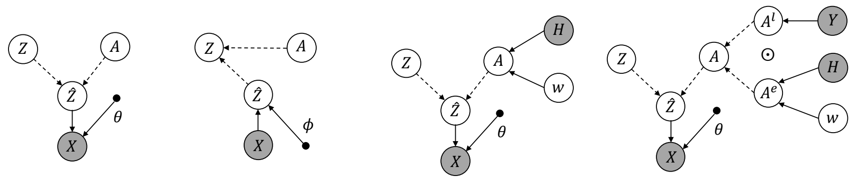


Figure 1: (a) Generative model of **DeepSEM**; (b) Inference model of **DeepSEM**; (c) Generative model of **InfoSEM-B**: InfoSEM with BioBERT gene-embedding priors H on interaction effects; (d) Generative model of **InfoSEM-BC**: InfoSEM with both BioBERT gene-embedding priors on interaction effects A^e and known interactions (e.g.-ChIP-seq) priors on logit of the probability of interactions A^l. Solid and dashed lines represent stochastic and deterministic relations respectively.

scenarios (Pratapa et al., 2020) where labeled data is unavailable. Approaches include information theory-based methods (Margolin et al., 2006; Kim, 2015; Chan et al., 2017) and self-regression models, which predict a gene's expression based on the expressions of all other genes: $\boldsymbol{x}_i = f_i(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_P) + \boldsymbol{z}_i$. The feature importance of $\boldsymbol{x}_k$ in f_i is the interaction effect from gene k to i. TIGRESS (Haury et al., 2012) solves a linear model for each f_i,

$$X = A^T X + Z, \tag{2}$$

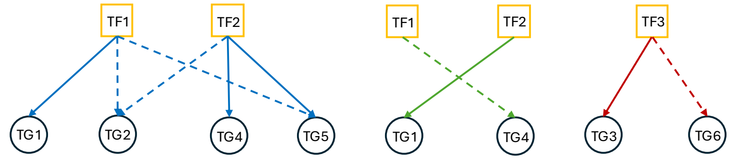
where Z is the noise term and all diagonal elements in the **weighted adjacency matrix** $A \in \mathbb{R}^{P \times P}$ are 0. An L_1 regularization is applied on A. A post-hoc threshold, e.g., only keeping the top 1% interactions, can be further applied on weighted adjacency matrix A to obtain adjacency matrix Y. GENIE3 (Huynh-Thu et al., 2010) and GRNBoost2 (Moerman et al., 2019) use random forests and gradient-boosting trees to enhance model flexibility and feature importance representation. Equation 2 resembles the linear structural equation model (SEM) in Bayesian networks (Zheng et al., 2018), with an added acyclicity constraint on A. However, classical GRNBoost2 still outperforms even the latest Directed Acyclic Graphs (DAG) learning algorithms due to feedback effects (Chevalley et al., 2022). DeepSEM (Shu et al., 2021) extends Eq.2 by modeling gene expression in a latent space without the acyclicity constraint, using $\beta-$VAE to denoise data, and outperforming traditional USL methods (Shu et al., 2021; Zhu & Slonim, 2024; Pratapa et al., 2020). However, it lacks support for incorporating external priors, which have shown to provide accuracy improvements for recent SL methods (Anonymous, 2024; Wang et al., 2024; Kommu et al., 2024).

3 INFOSEM: INFORMATIVE PRIORS FOR DEEPSEM

We begin by reviewing DeepSEM and then detail our proposed InfoSEM, a novel and principled generative model that incorporates textual gene embeddings from pre-trained foundation models and known regulatory interactions (when available), inspired by DeepSEM.

Introduction to DeepSEM The linear SEM model in Eq. 2 can be viewed as a generative model: it generates the dataset X with a GRN structure specified by A: First generating random noise Z; Then solving $X = (\mathrm{I} - A^T)^{-1} Z$ (Yu et al., 2019). Similarly, DeepSEM defines a generative model, shown in Figure 1 (a), as follows: $A \sim p(A), Z \sim p(Z), \hat{Z} = (\mathrm{I} - A^T)^{-1} Z, X \sim p_\theta(X|\hat{Z})$, where the GRN structure, represented by A with a Laplace prior $p(A) = \prod_{i,k} \text{Laplace}(a_{ik}; 0, \sigma_a)$ for encouraging sparsity, is modeled on the latent space $\hat{Z}$ of X, rather than directly on X. Both X and $\hat{Z}$ share the same GRN structure. We use $X \sim p_\theta(X|Z, A)$ to simplify the notation of generating X from Z and A. Similar to the vanilla VAE (Kingma, 2013), an inference network (shown in Figure 1 (b)) that approximates the posterior of Z, $q_\phi(Z|X, A)$, is introduced to learn DeepSEM: $\hat{Z} \sim q_\phi(\hat{Z}|X), Z = (\mathrm{I} - A^T)\hat{Z}$. All model parameters, A, θ, ϕ are optimized by maximizing a lower-bound of the likelihood as shown in Appendix A.

Incorporate pre-trained textual gene embeddings Given the frequent unavailability of costly experimental readouts, we propose to use readily available priors, such as gene embeddings derived from textual gene descriptions. Gene-level information can assist GRN inference: if gene i and gene k are functionally similar, they may also have similar interaction effects with other genes. We use the d dimensional embedding $\boldsymbol{h}_i \in \mathbb{R}^{1 \times d}$ of the textual description of the gene name from BioBERT (Lee et al., 2020), a language model pretrained on extensive biomedical literature, to represent the function of gene i. We design a prior of interaction effect informed by language model



(a) train set (known GRN for training) (b) *seen gene* test set (c) *unseen gene* test set

Figure 2: Strategies of train-test splits of the GT network, with three TFs (TF1-TF3) and six TGs (TG1-TG6). Solid and dashed lines represent positive and negative interactions. (a) Blue links (→) represent known interactions for training. (b) Green links (→) represent the test set $\mathcal{D}^{\text{test}}_{\text{seen}}$: pairs of genes for which the individual genes are in the train data, but not their interaction. (c) Red links (→) represent the test set $\mathcal{D}^{\text{test}}_{\text{unseen}}$: pairs that both the genes and their interactions are not in the train data.

embeddings as shown in Figure 1 (c): $p(A|H,\boldsymbol{w}) = \prod_{i,k} \text{Laplace}(a_{ik}; [\boldsymbol{h}_i, \boldsymbol{h}_k]\boldsymbol{w}^T, \sigma_a)$, where $[\boldsymbol{h}_i, \boldsymbol{h}_k]$ concatenates the gene embedding of genes i and k and $\boldsymbol{w} \in \mathbb{R}^{1\times 2d}$ is the parameter to learn with a prior $p(\boldsymbol{w}) = \mathcal{N}(0, \sigma^2_{\boldsymbol{w}})$. The prior mean can be interpreted as a linear model built on the gene embeddings for predicting their interactions, which has been successful in other gene-level tasks (Chen & Zou, 2024). Moreover, $p(A|H,\boldsymbol{w})$ makes similar genes have similar interaction effects with other genes. Intuitively, if gene 2 and gene 3 are functionally close, i.e., $\boldsymbol{h}_2 \approx \boldsymbol{h}_3$, the prior defines a similar mean for a_{12} and a_{13}: $[\boldsymbol{h}_1, \boldsymbol{h}_2]\boldsymbol{w}^T \approx [\boldsymbol{h}_1, \boldsymbol{h}_3]\boldsymbol{w}^T$. The prior is asymmetric, $p(a_{ik}|\boldsymbol{w}) \neq p(a_{ki}|\boldsymbol{w})$, which reflects the asymmetric nature of GRN. From here on, we refer to this model, which uses BioBERT embeddings as priors, as InfoSEM-B.

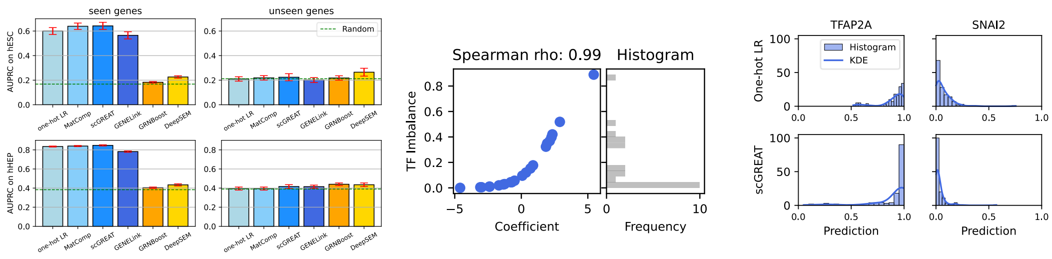
Incorporate both gene embeddings and known gene-gene interactions Although not always readily available, prior knowledge of gene-gene interactions can be obtained in specific scenarios, such as from ChIP-seq experiments (Yuan & Bar-Joseph, 2019; Anonymous, 2024; Chen & Liu, 2022), which can offer valuable biological insights that can guide GRN inference. Here, we show how to incorporate these known interactions in addition to the textual gene embeddings.

However, this comes with a natural challenge in that the partially observed Y is binary which does not inform the continuous weighted adjacency matrix A directly, but it can inform the **probability** of interactions. Therefore, we propose to decompose the weighted adjacency matrix A into $A^e \in \mathbb{R}^{P\times P}$, representing the magnitude of the interaction effect, and $A^l \in \mathbb{R}^{P\times P}$, with each element a^l_{ik} representing the **logit of the probability** that gene i interacts with gene k with $A = A^e \odot \sigma(A^l)$, where $\sigma(\cdot)$ is the sigmoid function and $\odot$ is element-wise product between them, as shown in Figure 1 (d). We work on the logit space of probability to remove the necessary constraints during the model training. The prior of A^e is informed by the gene embeddings from pretrained language models in the same way as before: $p_e(A^e|H,\boldsymbol{w}) = \prod_{i,k} \text{Laplace}(a^e_{ik}; [\boldsymbol{h}_i, \boldsymbol{h}_k]\boldsymbol{w}^T, \sigma_a)$. We define a prior on A^l using the partially observed Y, $p_l(A^l|Y) = \prod_{i,k} p_l(a^l_{ik}|y_{ik})$. We set $p_l(a^l_{ik}|y_{ik}) = \mathcal{N}(\text{logit}(0.95), \sigma^2_l)$ if $y_{ik} = 1$ and $\mathcal{N}(\text{logit}(0.05), \sigma^2_l)$ if $y_{ik} = 0$. This sets the mode of the prior probability that gene i interacts with gene k to 0.95 if we know they interact and 0.05 if they do not. If y_{ik} is not observed, we use a non-informative uniform prior $\boldsymbol{U}[\infty, \infty]$ for the logit, representing that the mode of the prior probability is 0.5. We will refer to this model, which incorporates BioBERT embeddings and known interactions (from e.g., ChIP-seq) when available as priors, as InfoSEM-BC. We train InfoSEM-B and InfoSEM-BC using variational Bayes as shown in Appendix A.

4 A NEW BENCHMARKING FRAMEWORK FOR EVALUATING GRNs

When using parts of the GT network to train a GRN inference model, careful train-test split strategies are essential for reliable evaluation. In most supervised GRN inference studies (Chen & Liu, 2022; Wang et al., 2023; Mao et al., 2023; Xu et al., 2023; Wang et al., 2024; Kommu et al., 2024), datasets are split by **gene pairs**, where edges for each TF are randomly divided into training and test sets. This results in a test set, $\mathcal{D}^{\text{test}}_{\text{seen}}$, containing *unseen interactions between seen genes*. This means that TFs and TGs in test sets are also in training sets, with only the regulatory interactions (edges) being distinct between the training and test sets (see Figure 2 (a) and (b)). This approach has the following major limitations, which are further explored in Section 5:

1. Risk of memorization due to class imbalance: The partially observed GRNs are often heavily imbalanced (Pratapa et al., 2020): most TFs have far more negative interactions (no regulation) than positive ones (Figure 3 (b) right). When all genes are represented in both training and test sets, models can memorize gene-specific patterns (e.g., gene IDs) by exploiting their node degrees from the ground-truth network rather than learning true regulatory mechanisms from scRNA-seq.



(a) Performance of existing methods (b) One-hot LR coefficients (c) One-hot LR and scGREAT predictions

Figure 3: (a) AUPRC of **existing** GRN inference models with **cell-type specific** target on both *seen genes* and *unseen genes* test sets. (b) Relations between class imbalance (proportion of positive interactions, y-axis) with histogram shown on the right and one-hot LR coefficient (x-axis) of each TF. (c) Histograms and kernel density estimations of predicted probabilities from one-hot LR and scGREAT on $\mathcal{D}^{\text{test}}_{\text{seen}}$ for two TFs: TFAP2A and SNAI2, with class imbalance level 0.89 and 0.06.

2. Inadequate representation of real-world scenarios: In real-world cases such as biomarker expansion, the interaction (degree) information of most genes is not represented in the partially observed GRN (Lotfi Shahreza et al., 2018). For instance, TF3, TG3, and TG6 in Figure 2 (c) exhibit this scenario. Evaluating models on *unseen genes* (genes not present in the training set) reflects their ability to generalize to new biological contexts, which is not captured by existing benchmarks.

In this work, we construct another test set, $\mathcal{D}^{\text{test}}_{\text{unseen}}$, containing *interaction between unseen genes*, i.e., all **genes** in $\mathcal{D}^{\text{test}}_{\text{unseen}}$ are not in the training set (e.g., Figure 2 (c)). The model performance and generalization ability on $\mathcal{D}^{\text{test}}_{\text{unseen}}$ would reflect the level of biology that the model has learned from the gene expression data. In practice, we randomly divide all TFs and TGs with known interactions into four sets: seen and unseen TFs, and seen and unseen TGs, with a ratio 3:1. All links between unseen TFs and unseen TGs are in $\mathcal{D}^{\text{test}}_{\text{unseen}}$. We then randomly divide all interactions between seen TFs and seen TGs into training $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}_{\text{seen}}$ with a ratio 3:1 using current strategies.

5 Experiments

We begin by experimentally validating the limitations of existing GRN inference benchmarks to motivate our new benchmarking setup, followed by a comparison of our proposed InfoSEM model with state-of-the-art SL and USL models. For reproducibility, all details are available in Appendix C.

We use scRNA-seq datasets of four cell lines and two different GRNs from BEELINE (Pratapa et al., 2020) (see Appendix B). We compare nine methods, including: 1. a random baseline; 2. two trivial methods that do not use scRNA-seq: one-hot LR (a logistic regression that takes the concatenation of one-hot embeddings of two genes as input) and MatComp (a matrix completion that imputes the partially observed adjacency matrix Y (Troyanskaya et al., 2001)); 3. two recent SL methods: scGREAT and GENELink; 4. two recent USL methods: GRNBoost2 and DeepSEM; 5. our proposed InfoSEM-B and InfoSEM-BC models. Inputs and training objectives for all models are summarized in Appendix Table 2. We consider two evaluation metrics (Pratapa et al., 2020): AUPRC and number of positive interactions among the top 1% (Hit@1%) predicted interactions. We repeat the train-test splits 10 times with different seeds to estimate the error bars.

5.1 Exploring the biases with existing benchmarks

Under the existing benchmarks involving *unseen interactions between seen genes*, we observe that the performance of existing SL and USL methods aligns with prior studies (Wang et al., 2024; Chen & Liu, 2022; Pratapa et al., 2020). However, no previous work has compared these methods to simple baselines, such as one-hot LR and matrix completion, which only use the partially observed Y without any information from scRNA-seq data. Surprisingly, those trivial methods perform similarly to advanced SL models across all cell lines tested (Figure 3 (a) and Appendix E.1). This performance is driven by one-hot LR's ability to exploit the node degree from the partially known GRN, e.g., the probability of a TF regulating a TG. We observe a strong rank correlation between the class imbalance level of each TF and the corresponding coefficient in the one-hot LR (Figure 3 (b)), supporting that one-hot LR predicts interactions based on the class imbalance level rather than biological relationships. Similarly, MatComp works well by learning degree biases with low-rank latent features. The cross-entropy loss (Eq.1) allows models to exploit this class imbalance and predict *unseen interactions between seen genes* based solely on these biases. For illustration, both

	hESC		hHEP		mDC		mESC	
	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%
One-hot LR	0.210 (0.018)	0.205 (0.041)	0.395 (0.016)	0.345 (0.056)	0.247 (0.019)	0.225 (0.104)	0.329 (0.026)	0.397 (0.036)
MatComp	0.219 (0.018)	0.191 (0.048)	0.395 (0.016)	0.356 (0.034)	0.240 (0.008)	0.225 (0.043)	0.342 (0.023)	0.345 (0.038)
scGREAT	0.224 (0.029)	0.222 (0.046)	0.416 (0.020)	0.433 (0.047)	0.245 (0.017)	0.183 (0.035)	**0.393 (0.027)**	0.390 (0.066)
GENELink	0.201 (0.020)	0.144 (0.050)	0.415 (0.016)	0.428 (0.060)	0.249 (0.013)	0.207 (0.054)	0.381 (0.030)	0.454 (0.094)
GRNBoost2	0.218 (0.018)	0.162 (0.037)	0.440 (0.014)	**0.553 (0.058)**	0.226 (0.009)	0.167 (0.053)	0.356 (0.023)	0.348 (0.049)
DeepSEM	0.265 (0.032)	0.419 (0.059)	0.435 (0.019)	0.517 (0.043)	0.277 (0.014)	0.292 (0.095)	0.343 (0.024)	0.369 (0.048)
InfoSEM-B	**0.331 (0.055)**	**0.547 (0.091)**	**0.498 (0.020)**	0.533 (0.048)	**0.298 (0.028)**	**0.472 (0.076)**	0.388 (0.023)	**0.522 (0.047)**
InfoSEM-BC	**0.331 (0.056)**	**0.585 (0.094)**	**0.499 (0.020)**	**0.550 (0.053)**	**0.322 (0.032)**	**0.498 (0.069)**	**0.408 (0.020)**	**0.575 (0.045)**
Random	0.222	0.215	0.390	0.397	0.232	0.231	0.349	0.351

Table 1: The GRN inference performance, measured by AUPRC and Hit@1% on *interactions between unseen genes* test set with **cell-type specific ChIP-seq**.

one-hot LR and scGREAT predict most unseen interactions of a seen TF (e.g., TFAP2A in Figure 3 (c)) positive if the TF has more positive interactions in the training set and vice versa (e.g., SNAI2).

To systematically confirm further whether sophisticated SL models, such as scGREAT and GENELink, also rely on dataset biases, we evaluated them on our new benchmark designed to predict *interactions between unseen genes* using scRNA-seq data. In this scenario, we observe a significant performance drop for SL methods, averaging 42% for cell-specific ChIP-seq and 79% for non-specific ChIP-seq (see Figure 3 (a), Table 1, and Appendix Table 5), even underperforming the random baseline. In contrast, USL methods, especially DeepSEM, outperformed SL methods on three datasets (hESC, hHEP, mDC) for cell-specific ChIP-seq without using known interactions. This further confirms the dependence of SL methods on dataset-specific biases, which holds across both cell-type-specific and non-cell-type-specific benchmarks, and for metrics such as Hit@1% (see Appendix E). Using simple techniques such as downsampling to address class imbalance in SL methods reduces their accuracy inflation but does not improve their accuracy on *unseen genes* (see Appendix D for details).

5.2 Utility of informative priors for InfoSEM

Next, we study the performance of our proposed InfoSEM models on our new *unseen genes* benchmark. Table 1 and Appendix Table 5 demonstrate that InfoSEM-BC and InfoSEM-B emerge as the best-performing models on all datasets that we tested for both cell-specific and non-cell-specific ChIP-seq. InfoSEM-B, which does not use known interactions, always improves upon DeepSEM on all datasets that we tested for both cell-specific and non-cell-specific cases by 25% and 52% on average, respectively. InfoSEM-B, without using any known interaction labels, achieves top-2 performance for both AUPRC and Hit@1% metrics on 3 out of 4 datasets, even when compared to SL baselines that incorporate known interactions (see Table 1 and Appendix Tables 5 and 6). InfoSEM-BC that incorporates information from known interactions achieves the best performance on all datasets with respect to AUPRC and on 3 out of 4 datasets with respect to Hit@1% for both cell-specific and non-cell-specific ChIP-seq target on *unseen genes* test set with average improvements over DeepSEM by 31% and 66% (see Table 1 and Appendix Tables 5 and 6). By integrating prior knowledge from known interactions, InfoSEM-BC reconstructs scRNA-seq data effectively, enabling strong performance on *unseen genes* without the pitfalls of class imbalance exploited by supervised methods relying on their discriminative loss (Eq.1). We also investigate the impact of alternative gene embeddings and the number of cells in Appendix E.3.

6 Conclusion and discussion

We study Gene Regulatory Network (GRN) inference using scRNA-seq data. Existing benchmarks focus on interactions between *seen genes*, i.e., all genes appear in both training and test sets. While suitable for predicting novel interactions within well-characterized gene sets, they fail to address applications where models must generalize to interactions between unseen genes such as biomarker expansion (Lotfi Shahreza et al., 2018). Additionally, we show that the high performance of supervised methods on existing benchmarks may be influenced by dataset (gene-specific) biases rather than their ability to learn true biological mechanisms. We propose a new, biologically motivated benchmarking framework that evaluates a model's ability to infer interactions between unseen genes. We also introduce InfoSEM, an unsupervised generative model that integrates biologically meaningful priors by leveraging textual gene embeddings from BioBERT. InfoSEM achieves 38.5% on average improvement over existing state-of-the-art models without informative priors for GRN inference. Furthermore, we show that InfoSEM can incorporate known interaction labels when available, further enhancing performance by 11.1% on average across datasets while avoiding the pitfalls of training with class imbalance. We believe InfoSEM's ability to generalize to unseen gene interactions, even with no or minimal labeled data, makes it a powerful tool for real-world applications.

References

Khandakar Tanvir Ahmed, Sunho Park, Qibing Jiang, Yunku Yeu, TaeHyun Hwang, and Wei Zhang. Network-based drug sensitivity prediction. *BMC medical genomics*, 13:1–10, 2020.

Anonymous. Deciphering cell lineage gene regulatory network via MTGRN. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Ecb6HBoo1r`. under review.

Lazar Atanackovic, Alexander Tong, Bo Wang, Leo J Lee, Yoshua Bengio, and Jason S Hartford. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. *Advances in Neural Information Processing Systems*, 36, 2024.

Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human b cells. *Nature genetics*, 37(4): 382–390, 2005.

J Gray Camp, Keisuke Sekine, Tobias Gerber, Henry Loeffler-Wirth, Hans Binder, Malgorzata Gac, Sabina Kanton, Jorge Kageyama, Georg Damm, Daniel Seehofer, et al. Multilineage communication regulates human liver bud development from pluripotency. *Nature*, 546(7659): 533–538, 2017.

Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

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.

Jiaxing Chen, ChinWang Cheong, Liang Lan, Xin Zhou, Jiming Liu, Aiping Lyu, William K Cheung, and Lu Zhang. Deepdrim: a deep neural network to reconstruct cell-type-specific gene regulatory network using single-cell rna-seq data. *Briefings in bioinformatics*, 22(6):bbab325, 2021.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, 2024.

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*, 2022.

Haoyue Dai, Ignavier Ng, Gongxu Luo, Peter Spirtes, Petar Stojanov, and Kun Zhang. Gene regulatory network inference in the presence of dropouts: a causal view. *arXiv preprint arXiv:2403.15500*, 2024.

Matthias Dehmer, Laurin AJ Mueller, and Frank Emmert-Streib. Quantitative network measures as biomarkers for classifying prostate cancer disease states: a systems approach to diagnostic biomarkers. *PloS one*, 8(11):e77602, 2013.

Sourish Ghosh and Anirban Basu. Network medicine in drug design: implications for neuroinflammation. *Drug discovery today*, 17(11-12):600–607, 2012.

Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean-Philippe Vert. Tigress: trustful inference of gene regulation using stability selection. *BMC Systems Biology*, 6:1–17, 2012.

Tetsutaro Hayashi, Haruka Ozaki, Yohei Sasagawa, Mana Umeda, Hiroki Danno, and Itoshi Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, 9(1):619, 2018.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, 9(10):770–780, 2008.

Seongho Kim. Ppcor: an r package for a fast calculation to semi-partial correlation coefficients. *Communications for statistical applications and methods*, 22(6):665, 2015.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sindhura Kommu, Yizhi Wang, Yue Wang, and Xuan Wang. Gene regulatory network inference from pre-trained single-cell transcriptomics transformer with joint graph learning. *arXiv preprint arXiv:2407.18181*, 2024.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Jinyu Li, Yutong Lai, Chi Zhang, and Qi Zhang. Tgcna: temporal gene coexpression network analysis using a low-rank plus sparse framework. *Journal of Applied Statistics*, 47(6):1064–1083, 2020.

Maryam Lotfi Shahreza, Nasser Ghadiri, Sayed Rasoul Mousavi, Jaleh Varshosaz, and James R Green. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics*, 19(5):878–892, 2018.

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.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC Bioinformatics*, volume 7, pp. 1–15. Springer, 2006.

Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, 2020.

Alex Rubinsteyn and Sergey Feldman. fancyimpute: An imputation library for python. URL `https://github.com/iskandr/fancyimpute`.

Alex K Shalek, Rahul Satija, Joe Shuga, John J Trombetta, Dave Gennert, Diana Lu, Peilin Chen, Rona S Gertner, Jellert T Gaublomme, Nir Yosef, et al. Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369, 2014.

Hantao Shu, Jingtian Zhou, Qiuyu Lian, Han Li, Dan Zhao, Jianyang Zeng, and Jianzhu Ma. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science*, 1(7): 491–501, 2021.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.

Jiacheng Wang, Yaojia Chen, and Quan Zou. Inferring gene regulatory network from single-cell transcriptomes with graph autoencoder model. *PLoS Genetics*, 19(9):e1010942, 2023.

Yuchen Wang, Xingjian Chen, Zetian Zheng, Lei Huang, Weidun Xie, Fuzhou Wang, Zhaolei Zhang, and Ka-Chun Wong. scgreat: Transformer-based deep-language model for gene regulatory network inference from single-cell transcriptomics. *Iscience*, 27(4), 2024.

Deborah Weighill, Marouen Ben Guebila, Camila Lopes-Ramos, Kimberly Glass, John Quackenbush, John Platig, and Rebekka Burkholz. Gene regulatory network inference as relaxed graph matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10263–10272, 2021.

Jing Xu, Aidi Zhang, Fang Liu, and Xiujun Zhang. Stgrns: an interpretable transformer-based method for inferring gene regulatory networks from single-cell transcriptomic data. *Bioinformatics*, 39(4): btad165, 2023.

Ivan Yevshin, Ruslan Sharipov, Semyon Kolmykov, Yury Kondrakhin, and Fedor Kolpakov. Gtrd: a database on gene transcription regulation—2019 update. *Nucleic acids research*, 47(D1):D100–D105, 2019.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

Qiuyue Yuan and Zhana Duren. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nature Biotechnology*, pp. 1–11, 2024.

Ye Yuan and Ziv Bar-Joseph. Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, 116(52):27151–27158, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Hao Zhu and Donna Slonim. From noise to knowledge: Diffusion probabilistic model-based neural inference of gene regulatory networks. *Journal of Computational Biology*, 31(11):1087–1103, 2024.

A LEARN PROBABILISTIC MODELS WITH VARIATIONAL BAYES.

All deep generative models for GRN inference inference are learned with variational Bayes, which are introduced below:

A.1 LEARN DEEPSEM WITH VARIATIONAL BAYES

In DeepSEM, the whole inference model is shown in Figure 1 (b). Weighted adjacency matrix A is learned using its maximum a posteriori (MAP) estimate $\tilde{A}$, which is equivalent to using a Dirac measure as the approximated posterior distribution, $q_{\tilde{A}}(A) = \delta(A = \tilde{A})$, in the variational Bayes framework. All parameters, i.e., $\tilde{A}, \theta, \phi$. are optimized by maximizing a lower-bound of the likelihood, i.e., evidence lower-bound (ELBO):

$$\begin{aligned} \log p(X) =& \log \int \frac{p_\theta(X|Z,A)P(A)P(Z)}{q_\phi(Z|X,A)} q_\phi(Z|X,A) dZ dA \\ \geq& \mathbb{E}_{q_\phi(Z|X,\tilde{A})}\left[\log p_\theta(X|Z,\tilde{A})\right] + \log p(\tilde{A}) - \mathrm{KL}\left[q_\phi(Z|X,\tilde{A})|p(Z)\right] = \mathcal{L}(\tilde{A},\theta,\phi), \end{aligned} \tag{3}$$

where the first term is the expected reconstruction error of the gene expression matrix X, $\log p(\tilde{A})$ regularizes the MAP estimate of A, and $\mathrm{KL}\left[q_\phi(Z|X,\tilde{A})|p(Z)\right]$ is the Kullback–Leibler divergence that regularizes the approximated posterior of Z and is weighted by β in practice.

In Eq.3, all distributions are fully factorized as shown below:

$$\begin{aligned} &q_\phi(Z|X,\tilde{A}) : q_\phi(\hat{Z}|X) = \prod_{i,j} \mathcal{N}(\hat{z}_{ij}; f_\phi(x_{ij}), g_\phi(x_{ij})),\ Z = (\mathrm{I} - \tilde{A}^T)\hat{Z}; \\ &p_\theta(X|Z,\tilde{A}) : \hat{Z} = (\mathrm{I} - \tilde{A}^T)^{-1} Z,\ p_\theta(X|\hat{Z}) = \prod_{i,j} \mathcal{N}(x_{ij}; f_\theta(\hat{z}_{ij}), g_\theta(\hat{z}_{ij})); \\ &p(\tilde{A}) = \prod_{i,k} \mathrm{Laplace}(a_{ik}; 0, \sigma_a); \\ &p(Z) = \prod_{i,j} \mathcal{N}(z_{ij}; 0, \sigma_z^2). \end{aligned} \tag{4}$$

A.2 LEARN INFOSEM WITH VARIATIONAL BAYES

Similar as DeepSEM, we use MAP estimates to infer all variables related to the weighted adjacency matrix, i.e., $\boldsymbol{w}$ in both models, A in InfoSEM-B, and A^e and A^l in InfoSEM-BC, given the data X. We use a full-rank matrix $\tilde{A}, \tilde{A}^e \in \mathbb{R}^{P \times P}$ for A and A^e, i.e., $q_{\tilde{A}}(A) = \delta(A = \tilde{A})$ and $q_{\tilde{A}^e}(A^e) = \delta(A^e = \tilde{A}^e)$. We use a low-rank MAP estimate with rank $h \ll P$ to infer A^l, motivated by the fact that the maximum rank of the adjacency matrix of a GRN is the number of transcription factors, as it represents the interactions from transcription factors to target genes (Li et al., 2020; Weighill et al., 2021). Therefore, we use $q_{A_a^l, A_b^l}(A^l) = \delta(A^l = A_a^l A_b^l)$, where $A_a^l \in \mathbb{R}^{P \times h}$, $A_b^l \in \mathbb{R}^{h \times P}$, and $A_a^l A_b^l$ has a rank h.

We derive the ELBO of InfoSEM-B with above model to be: $\mathcal{L}_{\text{InfoSEM-B}}(\tilde{A}, \theta, \phi, \boldsymbol{w}) =$

$$\mathbb{E}_{q_\phi(Z|X,\tilde{A})}\left[\log p_\theta(X|Z,\tilde{A})\right] + \log p(\tilde{A}|H,\boldsymbol{w}) + \log p_w(\boldsymbol{w}) - \mathrm{KL}\left[q_\phi(Z|X,\tilde{A})|p(Z)\right], \tag{5}$$

and the ELBO of InfoSEM-BC to be: $\mathcal{L}_{\text{InfoSEM-BC}}(\tilde{A}^e, A_a^l, A_b^l, \theta, \phi, \boldsymbol{w}) =$

$$\begin{aligned} &\mathbb{E}_{q_\phi(Z|X,\tilde{A}^e \odot \sigma(A_a^l A_b^l))}\left[\log p_\theta(X|Z, \tilde{A}^e \odot \sigma(A_a^l A_b^l))\right] + \log p_e(\tilde{A}^e|H,\boldsymbol{w}) + \\ &\log p_l(A_a^l A_b^l|Y) + \log p_w(\boldsymbol{w}) - \mathrm{KL}\left[q_\phi(Z|X,\tilde{A}^e \odot \sigma(A_a^l A_b^l))|p(Z)\right]. \end{aligned} \tag{6}$$

The detailed derivation of above ELBOs are shown below:

$$
\begin{aligned}
&p(X|H) \\
=&\int p_\theta(X|Z,A)p(A|H,\boldsymbol{w})p_w(\boldsymbol{w})p(Z)dZdA \\
=&\log\int \frac{p_\theta(X|Z,A)p(A|H,\boldsymbol{w})p_w(\boldsymbol{w})p(Z)}{q_\phi(Z|X,A)}q_\phi(Z|X,A)dZdA \\
=&\log\int \frac{p_\theta(X|Z,\tilde{A})p(\tilde{A}|H,\boldsymbol{w})p_w(\boldsymbol{w})p(Z)}{q_\phi(Z|X,\tilde{A})}q_\phi(Z|X,\tilde{A})dZ \\
\geq&\mathbb{E}_{q_\phi(Z|X,\tilde{A})}\left[\log p_\theta(X|Z,\tilde{A})\right]+\log p(\tilde{A}|H,\boldsymbol{w})+\log p_w(\boldsymbol{w})-\mathrm{KL}\left[q_\phi(Z|X,\tilde{A})|p(Z)\right] \\
=&\mathcal{L}_{\text{InfoSEM-B}}(\tilde{A},\theta,\phi,\boldsymbol{w}).
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
&p(X|H,Y) \\
=&\int p_\theta(X|Z,A^e\odot\sigma(A^l))p_e(A^e|H,\boldsymbol{w})p_l(A^l|Y)p_w(\boldsymbol{w})p(Z)dZdA^edA^l \\
=&\log\int \frac{p_\theta(X|Z,A^e\odot\sigma(A^l))p_e(A^e|H,\boldsymbol{w})p_l(A^l|Y)p_w(\boldsymbol{w})p(Z)}{q_\phi(Z|X,A^e\odot\sigma(A^l))}q_\phi(Z|X,A^e\odot\sigma(A^l))dZdA^edA^l \\
=&\log\int \frac{p_\theta(X|Z,\tilde{A}^e\odot\sigma(A^l_aA^l_b))p_e(\tilde{A}^e|H,\boldsymbol{w})p_l(A^l_aA^l_b|Y)p_w(\boldsymbol{w})p(Z)}{q_\phi(Z|X,\tilde{A}^e\odot\sigma(A^l_aA^l_b))}q_\phi(Z|X,\tilde{A}^e\odot\sigma(A^l_aA^l_b))dZ \\
\geq&\mathbb{E}_{q_\phi(Z|X,\tilde{A}^e\odot\sigma(A^l_aA^l_b))}\left[p_\theta(X|Z,\tilde{A}^e\odot\sigma(A^l_aA^l_b))\right]+\log p_e(\tilde{A}^e|H,\boldsymbol{w})+\log p_l(A^l_aA^l_b|Y)+\log p_w(\boldsymbol{w}) \\
-&\mathrm{KL}\left[q_\phi(Z|X,\tilde{A}^e\odot\sigma(A^l_aA^l_b))|p(Z)\right] \\
=&\mathcal{L}_{\text{InfoSEM-BC}}(\tilde{A}^e,A^l_a,A^l_b,\theta,\phi,\boldsymbol{w}).
\end{aligned}
\tag{8}
$$

B DETAILS OF EACH DATASET

For all experiments, we use scRNA-seq datasets of four cell lines from the popular BEELINE suite (Pratapa et al., 2020), including human embryonic stem cells (hESC) (Yuan & Bar-Joseph, 2019), human mature hepatocytes (hHEP) (Camp et al., 2017), mouse dendritic cells (mDC) Shalek et al. (2014), and mouse embryonic stem cells (mESC) (Hayashi et al., 2018). We consider two available ground-truth networks, cell-type specific ChIP-seq, collected from databases such as ENCODE and ChIP-Atlas, on the same or similar cell type, and non-cell-type specific transcriptional regulatory network from BEELINE (Pratapa et al., 2020).

B.1 CELL-TYPE SPECIFIC DATASETS

	hESC	hHEP	mDC	mESC
number of genes	844	908	1216	1353
number of cells	758	425	383	421
number of positive links	4404	9684	1129	40083
number of negative links	23448	17556	24407	78981
averaged node degree of TFs	133.5	322.8	53.8	455.5

B.2 NON-CELL-TYPE SPECIFIC DATASETS

	hESC	hHEP	mDC	mESC
number of genes	844	908	1216	1353
number of cells	758	425	383	421
number of positive links	3318	4033	3694	7705
number of negative links	229626	279263	300306	678266
averaged node degree of TFs	12.0	12.9	14.8	15.2

C REPRODUCIBILITY: DETAILS AND HYPERPARAMETERS OF EACH MODEL

Methods	scRNA-seq	Known GRN	External prior	Framework
One-hot LR	✗	✓	✗	SL
Matrix Completion	✗	✓	✗	SL
scGREAT	✓	✓	✓	SL
GENELink	✓	✓	✗	SL
GRNBoost2	✓	✗	✗	USL
DeepSEM	✓	✗	✗	USL
InfoSEM-B (Ours)	✓	✗	✓	USL
InfoSEM-BC (Ours)	✓	✓	✓	USL

Table 2: Properties of a list of benchmarking methods. Existing methods that make use of known GRN interactions treat them as prediction targets in supervised learning (SL) framework while our approaches consider them as a prior for constructing the scRNA-seq in an unsupervised learning (USL) framework.

In general, we cross-validate hyper-parameters using the partially known GRN on the training set to ensure a fair comparison with existing methods that use the same strategy. Code and datasets will be made available on acceptance.

One-hot LR: A logistic regression implemented by scikit-learn (Pedregosa et al., 2011) with the L_2 regularization coefficient cross-validated on the training set.

MatComp: A matrix completion algorithm implemented by fancyimpute (Rubinsteyn & Feldman) with rank 128.

scGREAT: Use hyperparameters and implementation provided by (Wang et al., 2024).

GENELink: Use hyperparameters and implementation provided by (Chen & Liu, 2022).

DeepSEM: Use hyperparameters and implementation provided by (Yuan & Bar-Joseph, 2019).

InfoSEM-B and InfoSEM-BC: We set hyper-parameters, e.g., neural network architectures, learning rates schedule, prior scale of latent variable Z, to be the same as DeepSEM (Yuan & Bar-Joseph, 2019). For unique hyper-parameters of InfoSEM-B and InfoSEM-BC, we cross-validate them on the training set using known GRN (shown as below).

Ground-truth	Methods	hESC	hHEP	mDC	mESC
cell-type specific ChIP-seq	$\sigma_{\boldsymbol{w}}$ (InfoSEM-B, InfoSEM-BC)	0.1	10	10	1
	σ_l (InfoSEM-BC)	0.1	$\sqrt{0.1}$	$\sqrt{0.1}$	0.1
	h (InfoSEM-BC)	128	128	128	128
non-cell-type specific ChIP-seq	$\sigma_{\boldsymbol{w}}$ (InfoSEM-B, InfoSEM-BC)	100	100	100	100
	σ_l (InfoSEM-BC)	$\sqrt{0.1}$	0.1	$\sqrt{0.1}$	$\sqrt{0.1}$
	h (InfoSEM-BC)	128	128	128	128

D Performance of supervised methods with downsampling

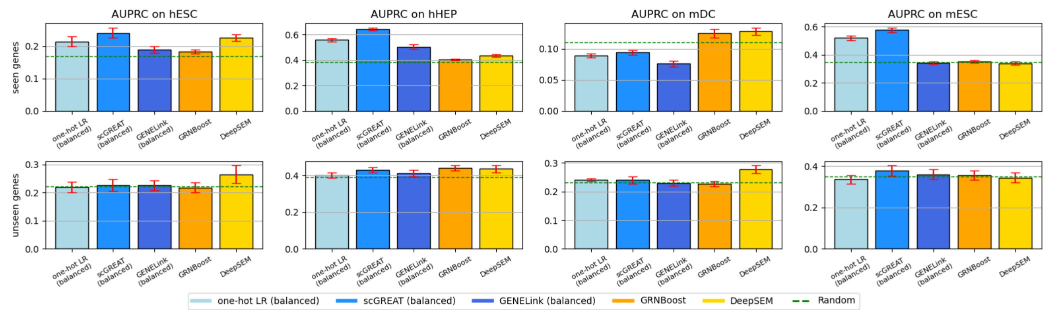


Figure 4: AUPRC with corresponding standard error of the mean of GRN inference models with **cell-type specific** target on *unseen interactions between seen genes* test sets. Although we apply downsampling to remove the class imbalance level associated with each TF, supervised methods still show an inflated accuracy on most of cell lines by using the class imbalance level associated with each TG.

If the class imbalance level associated with each gene causes the inflated accuracy of supervised method, one straightforward solution is to remove the class imbalance with downsampling. Specifically, we randomly select the same number of negative edges as the number of positive edges for each TF when training a supervised model, and the performance is shown in Fig 4.

We observe that supervised methods still have an inflated accuracy by comparing their performance on *unseen interactions between seen genes* and *interactions between unseen genes*. One reason is that edges associated with TGs can still be imbalanced even the edges associated with TFs are balanced, and supervised learning methods can still make use of such information easily.

Moreover, removing the class imbalance of TFs does not improve the performance on *interactions between unseen genes* and unsupervised methods, especially DeepSEM, are still outperform supervised methods on *unseen genes*.

E ADDITIONAL EXPERIMENTAL RESULTS

E.1 PERFORMANCE OF ALL METHODS ON *unseen interaction between seen genes* TEST SET.

	hESC		hHEP		mDC		mESC	
	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%
One-hot LR	**0.600 (0.028)**	0.932 (0.039)	0.835 (0.005)	**1.000 (0.003)**	0.176 (0.006)	0.250 (0.024)	0.844 (0.009)	**0.991 (0.004)**
MatComp	0.638 (0.026)	0.913 (0.037)	**0.840 (0.006)**	**1.000 (0.006)**	**0.351 (0.013)**	**0.419 (0.034)**	**0.860 (0.007)**	0.974 (0.008)
scGREAT	**0.642 (0.029)**	**0.966 (0.039)**	**0.847 (0.006)**	**1.000 (0.000)**	0.249 (0.011)	0.325 (0.047)	**0.858 (0.009)**	**1.000 (0.001)**
GENELink	0.565 (0.029)	**0.964 (0.043)**	0.782 (0.007)	0.942 (0.015)	0.178 (0.006)	0.162 (0.026)	0.690 (0.016)	0.716 (0.101)
GRNBoost2	0.173 (0.006)	0.230 (0.017)	0.384 (0.007)	0.423 (0.015)	0.115 (0.007)	0.141 (0.036)	0.352 (0.011)	0.342 (0.024)
DeepSEM	0.216 (0.010)	0.318 (0.016)	0.424 (0.011)	0.509 (0.028)	0.118 (0.006)	0.165 (0.033)	0.340 (0.013)	0.431 (0.023)
Random	0.168	0.171	0.383	0.384	0.110	0.111	0.346	0.344

Table 3: The GRN inference performance for each method evaluated on *unseen interactions between seen genes* with **cell-type specific ChIP-seq** targets. The top-2 best models are **bold**. We observe that all supervised learning methods, including one-hot LR, achieve much better performance than unsupervised methods.

	hESC		hHEP		mDC		mESC	
	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%
One-hot LR	0.159 (0.001)	0.257 (0.013)	0.186 (0.004)	0.306 (0.008)	0.126 (0.004)	0.228 (0.007)	0.127 (0.004)	0.205 (0.006)
MatComp	**0.207 (0.017)**	**0.334 (0.020)**	**0.247 (0.012)**	**0.375 (0.016)**	**0.396 (0.009)**	**0.563 (0.015)**	**0.260 (0.009)**	**0.375 (0.013)**
scGREAT	**0.173 (0.002)**	0.288 (0.014)	**0.244 (0.006)**	**0.366 (0.004)**	0.183 (0.012)	0.314 (0.004)	0.139 (0.004)	0.235 (0.004)
GENELink	0.059 (0.011)	0.101 (0.022)	0.088 (0.011)	0.157 (0.022)	0.103 (0.007)	0.141 (0.019)	0.025 (0.005)	0.025 (0.013)
GRNBoost2	0.019 (0.001)	0.031 (0.004)	0.018 (0.001)	0.031 (0.003)	0.020 (0.001)	0.031 (0.005)	0.018 (0.001)	0.042 (0.003)
DeepSEM	0.023 (0.001)	0.034 (0.004)	0.021 (0.001)	0.035 (0.004)	0.021 (0.001)	0.033 (0.005)	0.021 (0.001)	0.044 (0.002)
Random	0.019	0.017	0.018	0.017	0.018	0.018	0.014	0.014

Table 4: The GRN inference performance for each method evaluated on *unseen interactions between seen genes* with **non-cell-type specific ChIP-seq** targets. The top-2 best models are **bold**. We observe that InfoSEM-BC achieves top-2 performance on three dataset.

We show the performance of all methods evaluated under current evaluation setup, i.e., *unseen interactions between seen genes*, in Table 3 and Table 4 for cell-type specific and non-cell-type specific GRNs. We observe that trivial baselines (one-hot LR and matrix completion) without using any gene expression data always achieve top-2 performance.

E.2 PERFORMANCE OF METHODS ON *interaction between unseen genes* TEST SET FOR NON-CELL-TYPE SPECIFIC CHIP-SEQ TARGET.

	hESC		hHEP		mDC		mESC	
	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%	AUPRC	Hit@1%
One-hot LR	0.025 (0.002)	0.029 (0.008)	0.022 (0.002)	0.006 (0.006)	0.022 (0.002)	0.012 (0.004)	0.014 (0.001)	0.008 (0.004)
MatComp	0.024 (0.001)	0.017 (0.004)	0.021 (0.001)	0.017 (0.003)	0.023 (0.001)	0.031 (0.005)	0.014 (0.000)	0.014 (0.003)
scGREAT	0.025 (0.003)	0.031 (0.011)	**0.029 (0.003)**	0.027 (0.010)	0.028 (0.002)	0.024 (0.006)	**0.024 (0.001)**	0.037 (0.006)
GENELink	0.023 (0.001)	0.006 (0.006)	0.020 (0.002)	0.014 (0.007)	0.026 (0.002)	0.028 (0.013)	0.015 (0.001)	0.012 (0.003)
GRNBoost2	0.022 (0.001)	0.018 (0.006)	0.022 (0.001)	0.027 (0.009)	0.024 (0.001)	0.020 (0.004)	0.022 (0.003)	**0.041 (0.009)**
DeepSEM	0.028 (0.002)	0.032 (0.005)	0.026 (0.003)	0.028 (0.010)	0.026 (0.001)	0.028 (0.004)	0.022 (0.002)	0.028 (0.007)
InfoSEM-B	**0.036 (0.004)**	**0.038 (0.007)**	**0.029 (0.003)**	**0.049 (0.009)**	**0.050 (0.004)**	**0.077 (0.014)**	**0.023 (0.003)**	0.032 (0.008)
InfoSEM-BC	**0.038 (0.004)**	**0.042 (0.006)**	**0.030 (0.003)**	**0.054 (0.007)**	**0.051 (0.004)**	**0.082 (0.014)**	**0.023 (0.002)**	**0.045 (0.007)**
Random	0.024	0.024	0.021	0.021	0.024	0.022	0.014	0.014

Table 5: The GRN inference performance for each method evaluated on *interactions between unseen genes* with **non-cell-type specific ChIP-seq** targets. The top-2 best models are **bold**. We observe that InfoSEM-BC achieves top-2 performance on all dataset and InfoSEM-B achieves top-2 on three datasets.

We show the performance of all methods evaluated on *interactions between unseen genes* for the non-cell-type specific GRNs in Table 5, where the proposed InfoSEM is among top-2 best models in all cell lines. We observe that both AUPRC and Hit@1% are very small when benchmarked against non-cell-type specific GRNs unlike cell specific GRNs in Table 1. One reason is that negative links in non-cell-type specific GRNs contain both unknown positive links and negative links due to the data collection process (Pratapa et al., 2020), therefore, negative links in non-cell-type specific GRNs are much noisier than cell-type specific GRNs. However, positive links in non-cell-type specific GRNs do not contain such noise. We compute the recall to evaluate the performance of our methods on positive links only (Table 6) where we observe a much higher accuracy compared with Table 5 when both positive links and negative links are evaluated.

	hESC	hHEP	mDC	mESC
DeepSEM	0.10 (0.04)	0.11 (0.04)	0.60 (0.03)	0.03 (0.02)
InfoSEM-B	0.19 (0.09)	0.34 (0.14)	0.64 (0.10)	0.19 (0.12)
InfoSEM-BC	0.22 (0.09)	0.38 (0.14)	0.64 (0.10)	0.21 (0.11)

Table 6: Averaged recall of each method with non-cell-type specific target GRNs using a threshold 0.5.

E.3 SENSITIVITY ANALYSIS

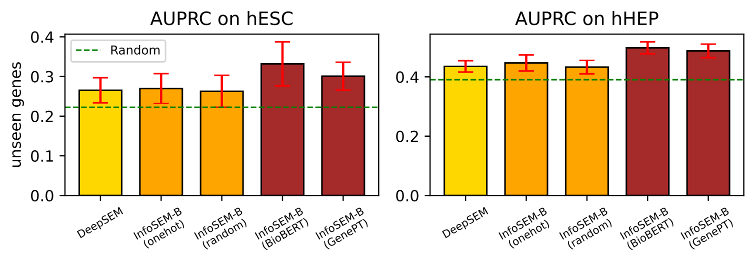


Figure 5: Results of using different prior gene embeddings in InfoSEM-B with **cell-type specific ChIP-seq** targets. InfoSEM-B with informative gene embedding, e.g., from BioBERT and GenePT, are better than models with noninformative embeddings, e.g., one-hot and random.

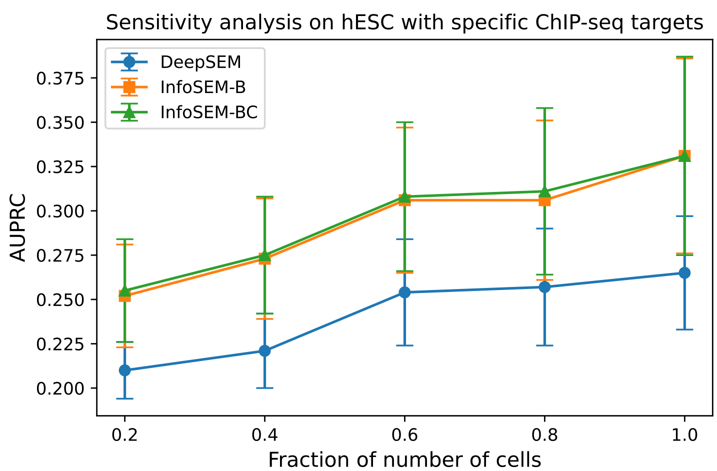


Figure 6: AUPRC of models on the *unseen genes* test set with different numbers of cells in the training data.

Since InfoSEM-B uses embeddings from BioBERT, it is natural to ask: how useful are these embeddings in improving GRN inference? To explore this, we investigate the impact of alternative gene embeddings in Figure 5, where InfoSEM-B (GenePT) replaces BioBERT embeddings with those from another textual gene embedding language model GenePT (Chen & Zou, 2024), still resulting in improved GRN inference compared to DeepSEM. As expected, using non-informative embeddings, such as one-hot or random embeddings, in InfoSEM-B does not enhance performance. Additionally, we evaluate the sensitivity of GRN inference accuracy with respect to the number of cells in the training data, as detailed in Appendix E.3. We illustrate how InfoSEM with different priors and DeepSEM perform on the unseen gene test sets when only a fraction of cells in the hESC dataset are used to train the model. We observe that InfoSEM-BC is only a slightly better than InfoSEM-B. Moreover, InfoSEM-B and InfoSEM-BC trained only on 20% cells can achieve a similar performance as DeepSEM trained on all cells.