

LLM4GRN: DISCOVERING CAUSAL GENE REGULATORY NETWORKS WITH LLMs – EVALUATION THROUGH SYNTHETIC DATA GENERATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Gene regulatory networks (GRNs) represent the causal relationships between transcription factors (TFs) and target genes in single-cell RNA sequencing (scRNA-seq) data. Understanding these networks is crucial for uncovering disease mechanisms and identifying therapeutic targets. In this work, we investigate the potential of large language models (LLMs) for GRN discovery, leveraging their learned biological knowledge alone or in combination with traditional statistical methods. We employ a task-based evaluation strategy to address the challenge of unavailable ground truth causal graphs. Specifically, we use the GRNs suggested by LLMs to guide causal synthetic data generation and compare the resulting data against the original dataset. Our statistical and biological assessments show that LLMs can support statistical modeling and data synthesis for biological research.

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is a cutting-edge technology that enables the collection of gene expression data from individual cells. This approach opens up new avenues for a wide range of scientific and clinical applications. One crucial application of scRNA-seq data is the reconstruction and analysis of gene regulatory networks (GRNs), which represent the interactions between genes. GRN analysis can deepen our understanding of disease mechanisms, identify key regulatory pathways, and provide a foundation for the development of interventional gene therapies and targeted drug discovery.

Statistical causal discovery algorithms (Mercatelli et al., 2020; Scheines et al., 1998; Zheng et al., 2018; Brouillard et al., 2020; Lippe et al., 2021; Roohani et al., 2024; Yu & Welch, 2022) can reveal potential causal links between TFs and their target gene. However, they often lack robustness and are prone to detecting spurious correlations, especially in high-dimensional, noisy single-cell data. Furthermore, many of these approaches rely heavily on prior knowledge from curated databases (e.g., TRANSFAC (Wingender et al., 1996), RegNetwork (Liu et al., 2015), ENCODE (de Souza, 2012), BioGRID (de Souza, 2012), and AnimalTFDB (Hu et al., 2019)), which frequently lack essential contextual information such as specific cell types or conditions, leading to inaccuracies in the inferred regulatory relationships (Zinati et al., 2024).

The recent advancements in and success of large language models (LLMs) have opened up new possibilities for their use in scientific discovery (Sheth et al., 2024; Lu et al., 2024; AI4Science & Quantum, 2023), including causal discovery (Kıçıman et al., 2023; Abdulaal et al., 2023; Kasetty et al., 2024; Vashishtha et al., 2023; AI4Science & Quantum, 2023; Abdulaal et al., 2023; Khatibi et al., 2024). Most of the above methods involve the refinement of the statistically inferred causal graph by LLM. However, LLMs excel at synthesizing vast amounts of heterogeneous knowledge, making them well-suited to tasks that require the integration of diverse datasets, such as constructing full causal graphs based on scientific literature (Sheth et al., 2024). In addition, LLMs have already been shown to perform various genomic data analysis tasks (Märtens et al.; Jin et al., 2024; Tang & Koo, 2023; Fang et al., 2024; Toufiq et al., 2023; Elsborg & Salvatore, 2023), including foundation models pre-trained for genomic tasks (Wang et al., 2024a; Cui et al., 2024).

Inspired by recent advancements, we harness LLMs for inferring gene regulatory networks (GRNs) from scRNA-seq data. Specifically, we utilize LLMs either to generate complete GRNs (causal

054 graphs) directly or to provide prior knowledge in the form of a list of potential transcription factors
055 (TFs), which can then be integrated into traditional statistical causal discovery algorithms.

056 We use causal synthetic data generation as a downstream task to evaluate GRNs or priors suggested by
057 LLMs in the absence of reliable ground-truth graphs. This method embeds gene regulatory networks
058 into the scRNA-seq data generation process and has been shown to preserve biological plausibility
059 better (Zinati et al., 2024). By comparing synthetic data to an oracle dataset, we assess the practical
060 utility of the inferred causal graph. Our results highlight the potential of general-purpose LLMs for
061 GRN inference, as evidenced by both statistical and biological metrics. The best performance is
062 achieved by combining LLMs with statistical GRN inference, pointing to a promising direction for
063 scRNA-seq data analysis.

064 **Our contributions:**

- 065 1. We show the effectiveness of using out of the box Large Language Models (LLMs) for
066 inferring gene regulatory networks (GRNs), showcasing their ability to capture complex
067 biological interactions.
- 068 2. Comprehensive performance evaluation based on synthetic data generation. Our methodol-
069 ogy effectively addresses the challenge posed by the absence of ground-truth causal graphs,
070 facilitating a more rigorous assessment of inference methods.
- 071 3. Extensive biological insight on the best-performing LLM for PBMC-All dataset.

072 2 RELATED WORKS

073 **LLMs and Causality.** Gene regulatory network inference from scRNA-seq data traditionally relies
074 on statistical causal discovery methods (Pratapa et al., 2020; Huynh-Thu et al., 2010; Moerman
075 et al., 2019). However, causal discovery often require external knowledge in the form of inter-
076 ventions (Brouillard et al., 2020), expert input (Kleinegesse et al., 2022), or priors from curated
077 databases (Zinati et al., 2024). Recent advances in large language models (LLMs) offer a promising
078 solution, as LLMs excel at integrating diverse knowledge and providing contextual information (Wan
079 et al., 2024; Ma, 2024). Many of the recent works leverage LLMs for causal discovery by utilizing
080 metadata, such as variable names, to infer causal relationships (Kıçıman et al., 2023; Ban et al.,
081 2023b; Vashishtha et al., 2023; Ban et al., 2023a). Further optimizations, including more advanced
082 prompting strategies beyond pairwise variable comparisons, have been developed to enhance causal
083 discovery (Vashishtha et al., 2023; Jiralerspong et al., 2024). Further (Sheth et al., 2024) explored
084 the effectiveness of completing a partial causal graph across diverse domains. This work, however,
085 considers LLM as an oracle to discover causal graphs for GRN inference.

086 **LLMs and Biology.** Models based on transformer architecture and trained on DNA or RNA
087 genomic sequences are effective at prediction and generation tasks (Zhang et al., 2024). However,
088 the availability of generative Gene-LLMs is limited, they lack rich contextual information and are
089 focused on specific tasks (Zhang et al., 2024). To overcome these limitations, foundation models are
090 tailored for the genomic tasks (Wang et al., 2024a; Cui et al., 2024). However, models trained on the
091 genetic data do not achieve the wide context of LLMs like GPT4 that inject information from the
092 scientific literature and freely available genomic databases. LLMs have been used for the tasks such
093 as enhancing gene perturbation generators (Märtens et al.), protein interaction prediction (Jin et al.,
094 2024), gene selection (Toufiq et al., 2023), analyzing biomarkers (Elsborg & Salvatore, 2023) and
095 cell annotation (Fang et al., 2024). To the best of our knowledge general purpose LLMs have not
096 been used for gene regulatory network inference.

097 **Causal Synthetic Data Generation.** Causally synthetic data generation is an approach that focuses
098 on embedding true causal relationships within the generated data. Several approaches have been
099 proposed in which the generator is guided by the causal acyclic graph (Cinquini et al., 2021; van
100 Breugel et al., 2021; Wen et al., 2021; Wang et al., 2024b; Bandyopadhyay & Sarkar, 2023). We use
101 a recently proposed causal GAN designed for scRNA-seq data generation, shown to produce more
102 biologically plausible results (Zinati et al., 2024).

3 PRELIMINARIES

3.1 CAUSAL GRAPH

A directed acyclic graph (DAG) $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ is composed of a set of variables/vertices \mathbf{V} and a set of (directed) edges \mathcal{E} between them such that no cycle is formed. Let P be the probability distribution over the same set of variables \mathbf{V} . \mathcal{G} and P satisfy the Markov condition if every variable is conditionally independent of its non-descendants given its parents. Assuming the Markov condition, the joint distribution of variables $V_1, V_2, \dots \in \mathbf{V}$ can be factorized as:

$$P[V_1, V_2, \dots, V_d] = \prod_i P[V_i | Pa(V_i)] . \quad (1)$$

where $Pa(V_i)$ denotes the set of parents of V_i .

A bipartite acyclic graph $\mathcal{G} = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ is an instance of a DAG whose vertices are divided into two subsets \mathbf{U} and \mathbf{V} . Each edge \mathcal{E}_i connects only two vertices from \mathbf{U} and \mathbf{V} .

3.1.1 GENE REGULATORY NETWORK (GRN)

A gene regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to control the gene expression levels of mRNA and proteins. GRNs describe the relationships between genes, transcription factors (TFs), RNA molecules, and other regulatory elements within a biological system, illustrating how genes are turned on or off and how their expression is modulated over time and in different conditions. GRN can be expressed as a directed acyclic graph (DAG) 3.1. In this work we are considering a bipartite DAG consisting of TFs and target genes.

3.2 GENERATIVE ADVERSARIAL NETWORK (GAN)

A Generative Adversarial Network (GAN) consists of two neural networks, the generator G and the discriminator D , which are trained simultaneously in a zero-sum game. The generator G maps a random noise vector \mathbf{z} sampled from a prior distribution $p_{\mathbf{z}}(\mathbf{z})$ (typically Gaussian) to the data space, producing a synthetic sample $G(\mathbf{z})$. The discriminator D maps an input \mathbf{x} to a scalar value $D(\mathbf{x}) \in [0, 1]$, representing the probability that \mathbf{x} is a real sample.

The training objective for GANs is formulated as a minimax game defined by the following function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))]$$

where $p_{\text{data}}(\mathbf{x})$ is the distribution of real data, and $p_g(\mathbf{x})$ is the distribution induced by G . The discriminator tries to maximize the probability of distinguishing real data from G 's samples, while G aims to minimize $\log(1 - D(G(\mathbf{z})))$ to generate realistic data. At equilibrium, the optimal generator replicates the true data distribution, making $p_g(\mathbf{x}) = p_{\text{data}}(\mathbf{x})$. The Wasserstein GAN (WGAN) introduces the Wasserstein distance as an alternative to the Kullback-Leibler or Jensen-Shannon divergences traditionally used in GANs. This distance metric measures the similarity between the distributions of real data and the generated data, leading to more stable training and mitigating issues like mode collapse (Arjovsky et al., 2017). WGAN with Gradient Penalty (WGAN-GP) (Gulrajani et al., 2017) further enhances this stability by replacing the original weight clipping technique of WGAN with a gradient penalty, addressing problems such as vanishing or exploding gradients and underutilization of model capacity.

3.2.1 GROUNdGAN

Causal GAN (CGAN) proposed by Kocaoglu et al. (2018) is a variant of GAN that incorporate a causal directed acyclic graph (DAG) into the data generation process. GRouNdGAN (Zinati et al., 2024) builds on CGAN, integrating components such as a causal controller, target generators, a critic, a labeler, and an anti-labeler. The training process for GRouNdGAN occurs in two stages. First, the causal controller, which generates transcription factor (TF) expression values, is pre-trained as the generator in a WGAN-GP framework. In the second stage, the pre-trained causal controller's TF

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

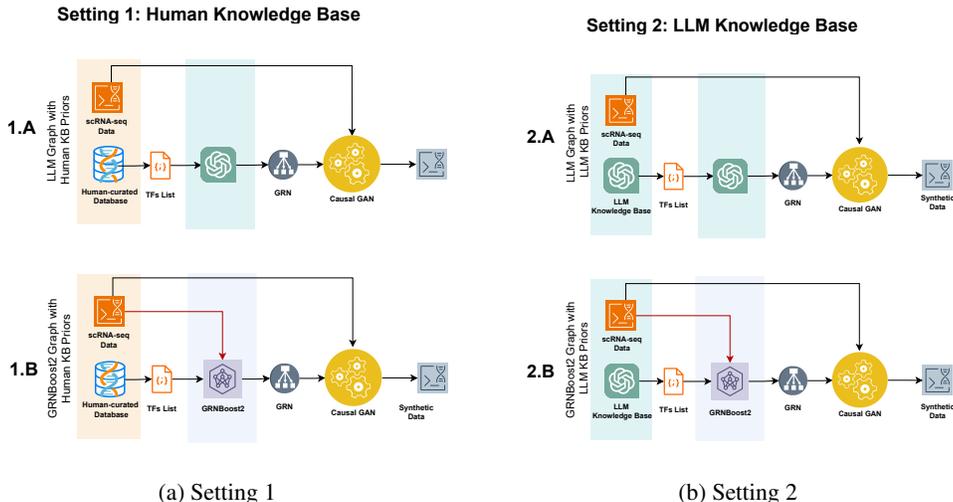


Figure 1: **Overview of LLM4GRN.** Setting **1.A** combines Human Knowledge Base (KB) with LLM. Setting **1.B** is the baseline setting that combines Human KB with GRNBoost2. Setting **2.A** is full LLM pipeline that combines LLM KB and LLM Inference. Setting **2.B** combines LLM KB with GRNBoost2 Inference.

expression values, along with randomly generated noise, are provided as input to target generators. These target generators, each an instance of WGAN-GP, produce the expressions of target genes while respecting the TF-gene relationships dictated by the input gene regulatory network (GRN). For a detailed explanation of GRouNdGAN’s architecture and training, please refer to the original paper.

4 LLM4GRN

In this work, we propose a novel approach integrating Large Language Models (LLMs) for GRN inference. The overall goal is to leverage the potential of LLMs in capturing complex biological interactions and to assess their utility. Given the lack of ground-truth data, we consider causal synthetic data generation as a downstream task to perform biological and statistical evaluations. Our methodology involves two distinct experimental settings that leverage different knowledge about the potential transcription factors (TFs) to guide gene regulatory network (GRN)-informed data generation. In each of the settings, we introduce LLMs into the pipeline motivated by their ability to incorporate extensive contextual information.

- In the first setting (Figure 1a, Setting **1.A**), we use LLM to infer the GRN graph by providing the LLM with a potential list of TF candidates sourced from a human-curated database.
- In the second setting (Figure 1b, Setting **2.A**), we utilize the LLM as the knowledge base, incorporating it earlier in the pipeline to infer potential TFs and to deduce the GRN graph.

We compare with Setting **1.B** and **2.B** where the Human knowledge base and LLM knowledge base is used by a statistical causal inference approach, GRNBoost2, respectively.

4.1 GRN GRAPH

To model the regulatory relationships between transcription factors and target genes. In this framework, we maintain a knowledge base (KB) that contains comprehensive information regarding the regulatory interactions, specifically indicating which genes are target genes of specific transcription factors. KB takes as input a list of genes and outputs the corresponding target genes and their regulating transcription factors:

$$KB : \text{List of Genes} \rightarrow \{(T_i, R_j) \mid T_i \in \mathbf{T}, R_j \in \mathbf{R}\}$$

We represent the gene regulatory network as a *bipartite graph* $\mathcal{G} = (\mathbf{T}, \mathbf{R}, \mathcal{E})$, where \mathbf{T} represents transcription factors (TFs), \mathbf{R} contains target genes regulated by these TFs. The set \mathcal{E} includes directed edges, where an edge $(\mathbf{T}_i, \mathbf{R}_j) \in \mathcal{E}$ indicates that transcription factor \mathbf{T}_i regulates target gene \mathbf{R}_j .

4.2 SETTING 1: HUMAN KNOWLEDGE BASE

Given the ability of large language models (LLMs) to generate causal graphs from metadata (such as variable names) (Abdulaal et al., 2023), we establish the foundational components of gene regulatory networks (GRNs) using a human knowledge base, denoted as KB^H . Let \mathbf{T}^H represent the set of transcription factors identified through the human-curated database, and let \mathbf{R}^H denote the set of target genes regulated by these factors. Total set of genes are defined by G . The directed edges in the graph are represented as \mathcal{E} , where an edge $(\mathbf{T}_i^H, \mathbf{R}_j^H) \in \mathcal{E}$ signifies that transcription factor \mathbf{T}_i^H regulates target gene \mathbf{R}_j^H .

In Setting **1A**, (Figure 1a), the LLM is employed to establish causal relationships between TFs and target genes, utilizing a list of TFs transcription factor candidates sourced from a human-curated database. We model the LLM as a function \mathcal{F}_{LLM} that, given a set of metadata \mathcal{M} , produces a bipartite graph $\mathcal{G} = (\mathbf{T}, \mathbf{R}, \mathcal{E})$, expressed as:

$$\mathcal{F}_{\text{LLM}}(\mathcal{M}, \mathbf{T}^H, \mathbf{R}^H) = \mathcal{G} \quad (2)$$

The input metadata \mathcal{M} encompasses gene names, transcription factors (TFs), single-cell RNA sequencing (scRNA-seq) data, and relevant biological context (such as species or experimental conditions) sourced from relevant literatures about the dataset.

It is important to highlight that, unlike traditional inference methods like GRNBoost2, the LLM-based GRN inference approach in this work does not rely on observational data, ensuring that individuals’ privacy in the dataset remains uncompromised. By integrating metadata from scRNA-seq datasets, the LLM can construct GRNs that are tailored specifically for the biological context of the data, potentially capturing nuances that statistical methods cannot. Detailed descriptions of the various prompting strategies employed can be found in the Appendix A.2.

4.3 SETTING 2: LLM KNOWLEDGE BASE

In Setting **2A**, (Figure 1b), we utilize LLM knowledge base, denoted as KB^{LLM} , to establish partition between transcription factors and target genes. In this context, the LLM is tasked with extracting relevant information directly from its knowledge base, which includes extensive biological data and relationships derived from various sources.

$$\mathcal{F}_{\text{LLM}}(\mathcal{M}) = (\mathbf{T}^{\text{LLM}}, \mathbf{R}^{\text{LLM}}) \quad (3)$$

Here, \mathbf{T}^{LLM} denotes the set of transcription factors identified from the LLM knowledge base, and \mathbf{R}^{LLM} represents the corresponding target genes. The input metadata \mathcal{M} includes gene names, transcription factors (TFs), and relevant biological context, leveraging the extensive knowledge embedded in the LLM.

4.4 GRN-INDUCED CAUSAL SYNTHETIC DATA GENERATION

We use the GRNs obtained by either of the settings, to perform synthetic causal data generation. These GRNs are fed into a causal GAN algorithm, specifically GRouNdGAN (Zinati et al., 2024), which uses them to generate synthetic datasets that correspond to each GRN structure in a two-stage approach.

5 RESULTS

We evaluate and compare the resulting synthetic datasets based on a range of statistical and biological metrics, allowing us to assess the quality and biological relevance of the data produced in each setting. Additionally, we analyze the TFs list generated by the LLM to gain insights, comparing them to priors from human curated database.

270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

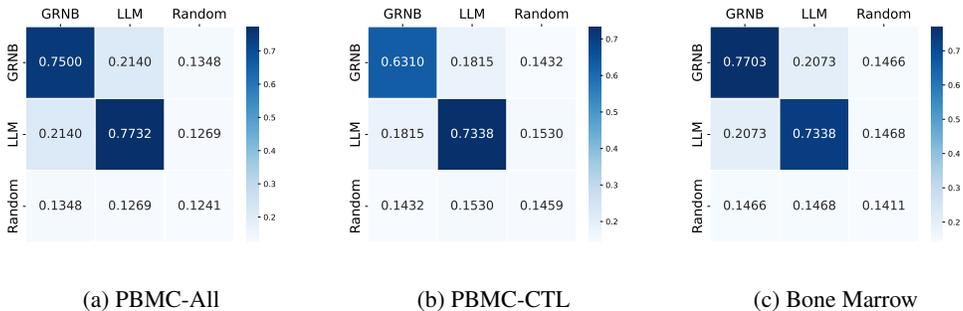


Figure 2: Overlap between GRNs proposed by different methods.

Experimental Setup. We generated synthetic data using datasets and protocols from (Zinati et al., 2024). Our preprocessing focused on creating train, test, and validation sets while maintaining 1,000 genes across all datasets (see Appendix A.1 for details). For each setting, we construct three different GRNs: one derived directly from the LLM, one generated by a causal discovery algorithm that incorporates the prior information (the dataset), and a third, randomly generated graph (based on TF list extracted from Human or LLM KB). For the GRN inference of LLM, we prompted with contextual knowledge. We used the state-of-art pretrained GPT-4 model (Achiam et al., 2023). Given GPT-4 strong performance for causal discovery (Abdulaal et al., 2023) across different domains including genomics, we prompted (see Appendix A.2) in zero-shot fashion. We also compare open-source model, Llama-3.1-70B in Setting 2.

Evaluation. We employed four statistical metrics: Cosine and Euclidean distances to measure the differences between centroids, maximum mean discrepancy (MMD) to assess the proximity of high-dimensional distributions without centroid creation, and Random Forest Area Under the Receiver Operating Characteristic (RF-AUROC) to determine the distinguishability of real and synthetic cells. We evaluate biological plausibility of the datasets by performing gene expression analysis and cell type annotation task. For more details, refer to Appendices A and B.

Datasets. For the direct comparison of the different settings, we focus on the dataset and genomic database information by (Zinati et al., 2024). Specifically, PBMC-All, PBMC-CTL and BoneMarrow data sets (details in Appendix A.1.1). Unlike Zinati et al. (2024), whose goal is to generate causal synthetic scRNA datasets, our objective is to evaluate various GRN graphs. We employ causal synthetic data generation as a means to address the lack of reliable ground truth.

5.1 COMPARISON AGAINST BASELINE

Evaluation of LLM graphs. For Setting 1, in the absence of ground truth for the GRN, we compare the overlap between the graph proposed by GRNBoost2, a statistical method, and the graph hypothesized by the LLM. While GRNBoost2 is not the definitive ground truth, it serves as a useful reference point. Analyzing this overlap allows us to assess how the hypothesized connections align with established statistical methods and examine how these differences impact downstream synthetic data generation metrics. As a baseline, we also compared against a randomly generated causal graph. Additionally, we test the consistency of the LLM’s performance across different random seeds by measuring the overlap between graphs produced from multiple seeds. This helps us evaluate the robustness and stability of the LLM-generated hypotheses. We plot the overlaps for all of the datasets in Figure 2. Our two main observations are that (1) the LLM-derived GRN demonstrates greater robustness compared to the GRNBoost2 (GRNB) GRN, particularly in terms of higher certainty; and (2) the overlap between the LLM-generated GRN and the random GRN is smaller than between LLM GRN and GRNBoost2 GRN, suggesting, that the LLM could be generating meaningful graphs.

Evaluation of LLM KB. In Setting 2, we introduced the LLM to filter TF and target genes from a list of genes. Similar to calculating overlaps in GRN evaluation, we also compute the overlap of TF produced by both approaches. From Table 5 (Appendix C), we observe there exists around just about half an overlap between the two knowledge bases. Interestingly we observe less than 50% overlap

between KB^H and KB^{LLM} for PBMC-CTL while a much higher overlap for the PBMC-ALL dataset. We also observed that LLM proposed a higher number of TFs, over 2 times in the case of the Bone Marrow dataset. As stated in the Table 6 (Appendix C), would change the density of the graphs potentially affecting the downstream tasks.

5.2 STATISTICAL EVALUATION OF GRN INFERENCE METHODS ON SYNTHETIC DATA

In Table 1, we present the statistical metrics results for the three datasets. Metrics are computed between a synthetic dataset of 1000 cells and a held-out test set of 1000 real cells for PBMC-ALL and PBMC-CTL. For BoneMarrow, 500 samples were used for synthetic and held-out test set. In GRouNdGAN’s imposed GRN, each gene is regulated by 10 transcription factors (TFs). Lower values (\downarrow) indicate better performance for all metrics, with the first two metrics representing the distance between the centroids of the real and synthetic cells. The “control” metrics are based on the real training dataset. The best performance values (excluding control and Stage1 which is a non-causal baseline) are highlighted in bold. Evaluations were carried out on 4 synthetic datasets, with experiments repeated using 2 cross-validation seeds.

Setting 1 Comparison. The Setting 1 (Table 1) result shows the LLM (GPT-4)-inferred GRN achieves the best performance across all metrics. For example, for PBMC-ALL, the LLM model shows the lowest Cosine distance of 0.00024 and Euclidean distance of 89, indicating that the synthetic data generated is most similar to the real data in terms of overall structure. Its lower MMD of 0.0072 compared to the GRNBoost2 model suggests it effectively captures subtle gene expression distributions. Additionally, the LLM model performs well on the Random Forest metric of 0.63, which measures binary classification accuracy in distinguishing real from synthetic data. Unsurprisingly, the Random Graph method performs the worst across all metrics, particularly with a high MMD of 0.0166 and the largest Euclidean distance of 121, highlighting the importance of informed GRN inference methods. In contrast, for the PBMC-CTL and BoneMarrow, the GRNBoost2 graph outperform the LLM (GPT-4) in this setting.

Setting 2 Comparison. In Setting 2 (Table 1), where we incorporated the LLM knowledge base, the GRNBoost2 method—combining LLM-proposed transcription factor (TF) lists with GRN inference—outperforms the fully LLM-based GRN approach (i.e., where the LLM proposes both the TF list and the connections between TFs and genes) across all datasets. GRNBoost2 also achieves the best overall performance for all datasets. For instance, in the PBMC-ALL dataset, it achieves a Euclidean distance of 83, improving from 121 in Setting 1, and an RF AUROC of 0.59, compared to 0.73 in Setting 1. Overall, GRNBoost2 shows improvement over its performance in Setting 1, delivering better results across all metrics for both PBMC-ALL and PBMC-CTL. Notably, it surpasses the top results from Setting 1, where the LLM achieved a Cosine distance of 0.00024 and a Euclidean distance of 89 for PBMC-ALL, while GRNBoost2 reached a Cosine distance of 0.00025 and a Euclidean distance of 65 for PBMC-CTL. These findings suggest that integrating LLM-derived priors with GRNBoost2’s statistical inference yields a more accurate and robust representation of gene regulatory networks (GRNs). In contrast, BoneMarrow performed best in Setting 1 but remains competitive in Setting 2 when the LLM knowledge base is combined with GRNBoost2.

Setting 2 Comparison using Llama-3.1-70B as Knowledge Base. Table 2 presents the results of Setting 2 for Llama-3.1-70B Knowledge Base. The findings indicate that Llama, when queried for knowledge, achieves the best overall performance in combination with GRNBoost2, outperforming GPT-4’s knowledge base (Table 1) across all metrics. These results are promising, suggesting that integrating LLM knowledge bases with causal inference approaches like GRNBoost2 offers a highly promising direction for scRNA causal synthetic data generation.

5.3 BIOLOGICAL PLAUSIBILITY OF CAUSAL SYNTHETIC DATA

We perform gene expression analysis and cell-type annotation tasks on the best statistically best-performing datasets. Full analysis can be found in Appendix D(subsection C.3).

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

		Cosine distance ↓	Euclidean distance ↓	MMD ↓	RF AUROC ↓
PBMC-ALL	<i>Baseline</i>				
	Control	0.00029±0.00008	100±16	0.0051±0.001	0.49±0.017
	Stage 1	0.00036±0.00009	107±15	0.0057±0.005	0.55±0.021
	<i>Setting 1 Human-KB</i>				
	GPT-4	<u>0.00024±0.00004</u>	<u>89±7</u>	<u>0.0072±0.001</u>	<u>0.63±0.043</u>
	GRNBoost2	0.00047±0.00021	121±25	0.0139±0.006	0.73±0.050
	Random	0.00045±0.00013	121±22	0.0166±0.003	0.85±0.019
	<i>Setting 2 GPT-4-KB</i>				
	GPT-4	0.00026±0.00009	90±13	0.0206±0.001	0.86±0.018
	GRNBoost2	0.00023±0.00008	83±17	0.0069±0.001	0.59±0.028
	Random	0.00026±0.00011	92±14	0.0226±0.002	0.87±0.018
	PBMC-CTL	<i>Baseline</i>			
Control		0.00020±0.00005	57±7	0.0045±0.000	0.54±0.007
Stage 1		0.00023±0.00003	63±4	0.0049±0.000	0.57±0.030
<i>Setting 1 Human-KB</i>					
GPT-4		0.00265±0.00253	176±115	0.0238±0.020	0.79±0.206
GRNBoost2		<u>0.00025±0.00004</u>	<u>65±6</u>	<u>0.0053±0.000</u>	0.59±0.028
Random		0.00027±0.00001	66±7	0.0085±0.001	0.75±0.021
<i>Setting 2 GPT-4-KB</i>					
GPT-4		0.00028±0.00009	67±11	0.0067±0.001	0.70±0.025
GRNBoost2		0.00020±0.00004	57±6	0.0049±0.000	0.59±0.019
Random		0.00022±0.00005	59±7	0.0080±0.000	0.76±0.023
Bone Marrow		<i>Baseline</i>			
	Control	0.00205±0.00018	80±4	0.0109±0.001	0.60±0.037
	Stage 1	0.00320±0.00115	101±16	0.0197±0.007	0.66±0.037
	<i>Setting 1 Human-KB</i>				
	GPT-4	0.00238±0.00052	86±9	0.0124±0.001	0.70±0.080
	GRNBoost2	0.00190±0.00023	77±5	0.0118±0.001	0.64±0.023
	Random	0.00219±0.00030	82±6	0.0150±0.003	0.75±0.041
	<i>Setting 2 GPT-4-KB</i>				
	GPT-4	0.00217±0.00050	82±10	0.0137±0.003	0.72±0.044
	GRNBoost2	<u>0.00193±0.00023</u>	<u>78±4</u>	<u>0.0119±0.001</u>	0.64±0.033
	Random	0.00297±0.00080	96±13	0.0172±0.004	0.79±0.046

Table 1: Performance of Different GRN Inference Methods in Simulating Realistic scRNA-seq Data across all 3 datasets. The best value for each dataset is presented in boldface and the best value in each setting is underline. The lower (↓) the better for all metrics.

	Cosine distance ↓	Euclidean distance ↓	MMD ↓	RF AUROC ↓
<i>Setting 2 Llama-KB</i>				
Llama-3.1	0.00032±0.00005	100±8	0.0105±0.0002	0.79±0.02
GRNBoost2	0.00016±0.00001	74±2	0.0064±0.0006	0.59±0.02
Random	0.00026±0.00001	89±1	0.0143±0.0006	0.89±0.01

Table 2: Comparison using LLM (Llama-3.1-70B) as Knowledge Base - performance GRN Inference Methods in Simulating Realistic scRNA-seq Data for PBMC-ALL dataset. The best value is presented in boldface. The lower (↓) the better for all metrics.

5.3.1 GENE EXPRESSION ANALYSIS

General Performance of Llama-KB GRNBoost2 in Cell-Specific Expression Profiles Although Llama-KB GRNBoost2 synthetic data Figure 3b introduces some noise, it generally improves cell-

486 in the databases (Members & Partners, 2024) making the information on mice genes scarcer in the
487 LLMs training data.

488 Biological plausibility analysis reveals that the Setting 2, Llama-KB dataset significantly improves
489 cell type differentiation and reduces noise compared to other generated datasets, particularly for
490 CD4+/CD45RA+/CD25- Naïve T cells and dendritic cells, which indicates better performance of the
491 model due to its ability to handle large contexts (Xiong et al., 2023).

492 In contrast, while the Setting 2, GPT4-KB model exhibited higher noise and less specificity, the
493 Setting 1, GRNBoost2 model also showed noisy patterns, especially in Naïve T and cytotoxic T
494 cells, emphasizing the need for models that clearly delineate cell type-specific expression. Notably,
495 discrepancies in cell type proportions across generated datasets, particularly the inflated presence
496 of CD8+/CD45RA+ Naïve Cytotoxic T cells, raise concerns about the biological relevance of these
497 models. As reported previously, large-scale single-cell studies tend to be more noisy which can
498 lead to sub-optimal biological inferences (Kavran & Clauset, 2021). Therefore, further refinements
499 are essential across all approaches to enhance specificity and reliability in representing true cellular
500 compositions, which is crucial for accurate downstream biological analyses and interpretations.
501 Finally, the Human-KB LLM model exhibited highly noisy or non-specific expression profiles,
502 particularly for CD4+/CD45A+/CD25- Naïve T cells, CD8+/CD45RA+ Naïve cytotoxic T cells,
503 CD4+ T helper cells, CD4+/CD25+ T regulatory cells, and CD4+/CD45RO+ memory cells. For
504 other cell types, the model either failed to express most top markers or showed high cell fractions
505 with reduced mean expression across different cell types.

506 One of the limitations of the study is the unifying constraints that are imposed on the GRN discovery
507 due to the parametric requirements of GRouNdGan (Zinati et al., 2024). Namely, all GRN graphs
508 are set to be bipartite graphs with the same number of TFs and same number of target genes for
509 each TF. These constraints restrict the diversity of the generated graphs, resulting in fairly similar
510 performance metrics among different GRN discovery approaches. In addition, multi-layer graphs with
511 transcription factors, cofactors, and target genes are more realistic and can better reflect biological
512 complexity (Karlebach & Shamir, 2008).

513 Despite the promising results, we note, that LLMs should be used cautiously in any high-stakes
514 decision-making. The models are prone to reporting false information with confidence (Ji et al.,
515 2023; Farquhar et al., 2024). In addition, they are susceptible to biases in the training data. One
516 of the most common sources of bias in machine learning is the under-representation of minority
517 populations in the training data. Transcription factor databases and literature often lack balanced
518 representation across ethnicity, ancestry, gender, and age groups. The vast majority of genetic data
519 and studies are based on individuals of European ancestry (Sirugo et al., 2019; Bentley et al., 2017).
520 The underrepresentation of non-European ancestry populations in genomic databases can obscure
521 gene-disease associations that are rare in European groups, leading to treatments that may not be
522 effective for these communities (Landry et al., 2018; Tawfik et al., 2023; Barral-Arca et al., 2019).
523 Age groups, specifically children and elderly, females, low-income populations and individuals from
524 remote rural areas are also underrepresented in clinical trials and data, making them less likely to
525 benefit from the achievements of precision medicine (Mosenifar, 2007; Davis et al., 2019; Steinberg
526 et al., 2021). LLMs are shown to exacerbate some existing health disparities (Pfohl et al., 2024), and
527 are likely to be influenced by the under-representation bias in the genomic data.

528 6 CONCLUSION

529
530 In conclusion, our analysis highlights the potential of using Large Language Models (LLMs) for gene
531 regulatory network (GRN) inference, as supported by both statistical and biological performance
532 metrics. Notably, the best performing approach combines LLM-derived TF priors with GRNBoost2
533 for statistical inference. In addition, we see good potential for using open-source models such as
534 Llama for GRNs, and we aim to explore their utility further, either directly or with fine-tuning. Our
535 biological results suggest that additional refinement is needed. For this we foresee developing GRNs
536 tailored to individual cell types, as evidence indicates that GRNs are often cell-type-specific.

REFERENCES

- 540
541
542 Ahmed Abdulaal, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C
543 Castro, Daniel C Alexander, et al. Causal modelling agents: Causal graph discovery through
544 synergising metadata-and data-driven reasoning. In *The Twelfth International Conference on*
545 *Learning Representations*, 2023.
- 546
547 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
548 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.
549 *arXiv preprint arXiv:2303.08774*, 2023.
- 550
551 Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models
552 on scientific discovery: a preliminary study using gpt-4. *arXiv*, 2023.
- 553
554 Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In
555 Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Ma-*
556 *chine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–
11 Aug 2017. URL <https://proceedings.mlr.press/v70/arjovsky17a.html>.
- 557
558 Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning
559 supervised by large language model. *arXiv*, 2023a.
- 560
561 Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects:
562 Harnessing large language models for advanced causal discovery from data. *arXiv*, 2023b.
- 563
564 Soma Bandyopadhyay and Sudeshna Sarkar. Exploring causality aware data synthesis. In *Proceedings*
565 *of the Third International Conference on AI-ML Systems*, pp. 1–9, 2023.
- 566
567 Ruth Barral-Arca, Jacobo Pardo-Seco, Xabi Bello, Federico Martinon-Torres, and Antonio Salas.
568 Ancestry patterns inferred from massive rna-seq data. *RNA*, 25(7):857–868, 2019.
- 569
570 Amy R Bentley, Shawneequa Callier, and Charles N Rotimi. Diversity and inclusion in genomic
571 research: why the uneven progress? *Journal of community genetics*, 8:255–266, 2017.
- 572
573 Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre
574 Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information*
575 *Processing Systems*, 33:21865–21877, 2020.
- 576
577 Martina Cinquini, Fosca Giannotti, and Riccardo Guidotti. Boosting synthetic data generation with
578 effective nonlinear causal discovery. In *2021 IEEE Third International Conference on Cognitive*
579 *Machine Intelligence (CogMI)*, pp. 54–63. IEEE, 2021.
- 580
581 Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang.
582 scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature*
583 *Methods*, pp. 1–11, 2024.
- 584
585 Terry C Davis, Connie L Arnold, Glenn Mills, and Lucio Miele. A qualitative study exploring
586 barriers and facilitators of enrolling underrepresented populations in clinical trials and biobanking.
587 *Frontiers in Cell and Developmental Biology*, 7:74, 2019.
- 588
589 Natalie de Souza. The encode project. *Nature methods*, 9(11):1046–1046, 2012.
- 590
591 Jonas Elsborg and Marco Salvatore. Using llms and explainable ml to analyze biomarkers at
592 single-cell level for improved understanding of diseases. *Biomolecules*, 13(10):1516, 2023.
- 593
Chen Fang, Yidong Wang, Yunze Song, Qingqing Long, Wang Lu, Linghui Chen, Pengfei Wang,
Guihai Feng, Yuanchun Zhou, and Xin Li. How do large language models understand genes and
cells. *bioRxiv*, pp. 2024–03, 2024.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large
language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

- 594 Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville.
595 Improved training of wasserstein gans. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach,
596 R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing*
597 *Systems*, volume 30. Curran Associates, Inc., 2017.
- 598 Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. Animalfdb 3.0:
599 a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic*
600 *acids research*, 47(D1):D33–D38, 2019.
- 602 Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory
603 networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.
- 604 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,
605 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM*
606 *Computing Surveys*, 55(12):1–38, 2023.
- 608 Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan
609 Du, and Yongfeng Zhang. Prollm: Protein chain-of-thoughts enhanced llm for protein-protein
610 interaction prediction. *bioRxiv*, pp. 2024–04, 2024.
- 612 Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal
613 graph discovery using large language models. *arXiv*, 2024.
- 614 Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature reviews*
615 *Molecular cell biology*, 9(10):770–780, 2008.
- 617 Tejas Kasetty, Divyat Mahajan, Gintare Karolina Dziugaite, Alexandre Drouin, and Dhanya Srid-
618 har. Evaluating interventional reasoning capabilities of large language models. *arXiv preprint*
619 *arXiv:2404.05545*, 2024.
- 620 A. J. Kavran and A. Clauset. Denoising large-scale biological data using network filters. *BMC*
621 *Bioinformatics*, 22:157, 2021. doi: 10.1186/s12859-021-04075-x.
- 623 Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. Alcm:
624 Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.
- 625 Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language
626 models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.
- 628 Steven Kleinegesse, Andrew R Lawrence, and Hana Chockler. Domain knowledge in a*-based causal
629 discovery. *arXiv preprint arXiv:2208.08247*, 2022.
- 630 Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath. CausalGAN:
631 Learning causal implicit generative models with adversarial training. In *International Conference*
632 *on Learning Representations*, 2018.
- 634 Latrice G Landry, Nadya Ali, David R Williams, Heidi L Rehm, and Vence L Bonham. Lack of
635 diversity in genomic databases is a barrier to translating precision medicine research into practice.
636 *Health Affairs*, 37(5):780–785, 2018.
- 637 Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity
638 constraints. *arXiv preprint arXiv:2107.10483*, 2021.
- 640 Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. Regnetwork: an integrated database of
641 transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015:
642 bav095, 2015.
- 644 Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist:
645 Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.
- 646 Jing Ma. Causal inference with large language model: A survey. *arXiv preprint arXiv:2409.09822*,
647 2024.

- 648 Kaspar Märtens, Rory Donovan-Maiye, and Jesper Ferkinghoff-Borg. Enhancing generative perturba-
649 tion models with llm-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning*
650 *for Genomics Explorations*.
- 651 National Genomics Data Center Members and Partners. Database commons: A catalog of databases in
652 life sciences. <https://ngdc.cncb.ac.cn/databasecommons/stat>, 2024. Accessed:
653 2024-09-27.
- 654
655 Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene
656 regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta*
657 *(BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020.
- 658
659 Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan
660 Aerts, and Stein Aerts. Grnboost2 and arboreto: efficient and scalable inference of gene regulatory
661 networks. *Bioinformatics*, 35(12):2159–2161, 2019.
- 662
663 Zab Mosenifar. Population issues in clinical trials. *Proceedings of the American Thoracic Society*, 4
664 (2):185–188, 2007.
- 665
666 Franziska Paul, Ya’ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-
667 Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, et al. Transcriptional
668 heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015.
- 669
670 Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad
671 Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for
672 surfacing health equity harms and biases in large language models. *Nature Medicine*, pp. 1–11,
673 2024.
- 674
675 Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking
676 algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature*
677 *methods*, 17(2):147–154, 2020.
- 678
679 Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel
680 multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.
- 681
682 Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The
683 tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral*
684 *Research*, 33(1):65–117, 1998.
- 685
686 Ivaxi Sheth, Sahar Abdelnabi, and Mario Fritz. Hypothesizing missing causal variables with llms.
687 *arXiv preprint arXiv:2409.02604*, 2024.
- 688
689 Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The missing diversity in human genetic
690 studies. *Cell*, 177(1):26–31, 2019.
- 691
692 Jecca R Steinberg, Brandon E Turner, Brannon T Weeks, Christopher J Magnani, Bonnie O Wong,
693 Fatima Rodriguez, Lynn M Yee, and Mark R Cullen. Analysis of female enrollment and participant
694 sex by burden of disease in us clinical trials between 2000 and 2020. *JAMA Network Open*, 4(6):
695 e2113749–e2113749, 2021.
- 696
697 Ziqi Tang and Peter K Koo. Building foundation models for regulatory genomics requires rethinking
698 large language models. In *Proceedings of the ICML Workshop on Computational Biology*, 2023.
- 699
700 Sherouk M Tawfik, Aliaa A Elhosseiny, Aya A Galal, Martina B William, Esraa Qansuwa, Rana M
701 Elbaz, and Mohamed Salama. Health inequity in genomic personalized medicine in underrepre-
702 sented populations: a look at the current evidence. *Functional & Integrative Genomics*, 23(1):54,
703 2023.
- 704
705 Mohammed Toufiq, Darawan Rinchai, Eleonore Bettacchioli, Basirudeen Syed Ahamed Kabeer,
706 Taushif Khan, Bishesh Subba, Olivia White, Marina Yurieva, Joshy George, Noemie Jourde-Chiche,
707 et al. Harnessing large language models (llms) for candidate gene prioritization and selection.
708 *Journal of Translational Medicine*, 21(1):728, 2023.

- 702 Pedro Valero-Lara, Alexis Huante, Mustafa Al Lail, William F Godoy, Keita Teranishi, Prasanna
703 Balaprakash, and Jeffrey S Vetter. Comparing llama-2 and gpt-3 llms for hpc kernels generation.
704 *arXiv preprint arXiv:2309.07103*, 2023.
- 705
706 Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating
707 fair synthetic data using causally-aware generative networks. *Advances in Neural Information*
708 *Processing Systems*, 34:22221–22233, 2021.
- 709 Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Bal-
710 asubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint*
711 *arXiv:2310.15117*, 2023.
- 712
713 Guangya Wan, Yuqi Wu, Mengxuan Hu, Zhixuan Chu, and Sheng Li. Bridging causal discovery and
714 large language models: A comprehensive survey of integrative approaches and future directions.
715 *arXiv preprint arXiv:2402.11068*, 2024.
- 716 Yuchen Wang, Xingjian Chen, Zetian Zheng, Lei Huang, Weidun Xie, Fuzhou Wang, Zhaolei Zhang,
717 and Ka-Chun Wong. scgreat: Transformer-based deep-language model for gene regulatory network
718 inference from single-cell transcriptomics. *Iscience*, 27(4), 2024a.
- 719 Zhehao Wang, Arran Zeyu Wang, David Borland, and David Gotz. Causalsynth: An interactive web
720 application for synthetic dataset generation and visualization with user-defined causal relationships.
721 *IEEE VIS Posters*, 6, 2024b.
- 722
723 Bingyang Wen, Luis Oliveros Colon, KP Subbalakshmi, and Rajarathnam Chandramouli. Causal-
724 tgan: Generating tabular data using causal generative adversarial networks. *arXiv preprint*
725 *arXiv:2104.10680*, 2021.
- 726 Edgar Wingender, Peter Dietze, Holger Karas, and Rainer Knüppel. Transfac: a database on
727 transcription factors and their dna binding sites. *Nucleic acids research*, 24(1):238–241, 1996.
- 728
729 Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin,
730 Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of
731 foundation models. *arXiv preprint arXiv:2309.16039*, 2023.
- 732 Hengshi Yu and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and
733 genetic perturbations. *BioRxiv*, pp. 2022–07, 2022.
- 734
735 Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao
736 Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological
737 & chemical domains. *arXiv preprint arXiv:2401.14656*, 2024.
- 738 Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson,
739 Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel
740 digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.
- 741
742 Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous
743 optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.
- 744 Yazdan Zinati, Abdulrahman Takiddeen, and Amin Emad. Groundgan: Grn-guided simulation of
745 single-cell rna-seq data using causal generative adversarial networks. *Nature Communications*, 15
746 (1):4055, 2024.
- 747
748
749
750
751
752
753
754
755

A EXPERIMENTAL SETUP

A.1 DATA PREPARATION AND MODIFICATIONS

In synthetic data generation, we use the same data sets and follow the protocol described by Zinati et al. Zinati et al. (2024).

We downloaded the three datasets using the link provided at <https://emad-combine-lab.github.io/GRouNdGAN/tutorial.html#demo-datasets>.

To ensure consistent results across different runs, we modified the code in two key ways: (1) we seeded the randomness in the ‘main.py’ file. While the GRNBoost2 algorithm is already seeded by default, ensuring consistent output, the rest of the code was not. Given that we are testing different GRN graphs, it was critical to control for randomness to prevent it from influencing our results. (2) We addressed additional bugs to ensure smooth execution, such as handling sparse data loading and ensuring that the code was properly assigned to the correct device for training. All modifications have been incorporated into our forked version of the repository, available upon publication. Some of the bugs we identified were also reported and fixed by the original authors.

After downloading the raw data, we preprocessed it to create separate train, test, and validation files. Following the paper’s recommendations, we used a test set of 1,000 samples for the PBMC and CTL datasets, and 500 samples for the BoneMarrow dataset. For validation, we set aside 1,000 samples each for PBMC and CTL, and 500 for BoneMarrow. The subsets were disjoint. We refer to these preprocessed datasets as “Real Data” throughout the paper. The preprocessed data will be made available upon publication.

During preprocessing, we observed that the BoneMarrow dataset had two genes absent in the training data, as they were not expressed in any cells. This issue did not occur in the full dataset, where genes expressed in fewer than three cells and cells expressing fewer than 10 genes had been filtered out. However, after splitting the data into train, test, and validation sets, some genes in specific splits were not expressed at all.

To address this, we modified the code to filter out genes expressed in fewer than one cell for each data split. This reduced the number of genes in the BoneMarrow dataset from 1,000 to 910. In the PBMC dataset, we found that 493 genes were not expressed in the test set, requiring extensive filtering.

To resolve this, we revised our approach. Instead of applying the original threshold, which discarded genes expressed in fewer than three cells, we increased this threshold to 680 cells for the PBMC dataset, equivalent to 0.01% of the full dataset. This change allowed us to retain the full 1,000 highly variable genes. However, the resulting set of 1,000 genes may differ slightly from those identified with the initial criteria, given that the full PBMC dataset contains over 32,738 genes, while BoneMarrow contains 3,451 genes.

This revised strategy allowed us to retain the full 1,000 genes for the PBMC and CTL datasets but not for BoneMarrow. To address this, we further adjusted the strategy by filtering out genes expressed in fewer than the number of cells in the test set for each dataset (1,000 for PBMC and CTL, and 500 for BoneMarrow). This approach successfully retained 1,000 genes across all datasets.

After preprocessing, we followed the rest of the process as detailed in the tutorial provided by the authors.

A.1.1 DATA SETS

- Human peripheral blood mononuclear cell (PBMC-A1). 68579 samples corresponding to 11 cell types
- Human peripheral blood mononuclear CD8+ Cytotoxic T-cells (PBMC-CTL). 20773 samples from the most common cell type in PBMC-All
- Mouse bone marrow Hematopoietic stem cells lineage differentiation (BoneMarrow). 2730 cells.

Dataset	# TFs	# Targets	# Genes	# Possible Edges	# Imposed Edges	GRN density Edges
PBMC	75	925	1000	69375	9250	0.1333
CTL	65	935	1000	60775	9350	0.1538
BoneMarrow	68	932	1000	63376	9320	0.1471

Table 3: The density of the causal graphs.

A.2 LLM PROMPTING STRATEGIES

In this study, we utilize prompting techniques to guide the behavior of a Large Language Model (LLM) for gene regulatory network (GRN) inference and other downstream analyses. The model employed is based on a pretrained large language model, specifically GPT-4, which has not been fine-tuned for this task. Instead, we rely on advanced prompting strategies, including Chain of Thought (CoT) reasoning and context provision, to enhance the performance of the LLM in generating biologically plausible results.

A.2.1 CHAIN OF THOUGHT (COT) PROMPTING

Chain of Thought (CoT) prompting is a method that encourages the LLM to reason through intermediate steps, producing a more transparent and logical progression towards its final answer. By guiding the model to provide step-by-step explanations before arriving at a conclusion, we aim to improve the interpretability and accuracy of its predictions. CoT prompting is particularly valuable for tasks that require complex reasoning, such as GRN inference, where multiple factors, such as transcription factor interactions and gene expression patterns, must be considered.

A.2.2 CONTEXT PROVISION

To further enhance the performance of the LLM, we utilize context provision by supplying the model with relevant background information prior to prompting. This technique is especially important when the task requires domain-specific knowledge, such as GRN inference from single-cell RNA sequencing (scRNA-seq) data. By embedding relevant context in the prompt, we can better align the model’s responses with the biological characteristics of the data being analyzed. Specifically, we provide excerpts from the original articles where the analyzed data sets were introduced Paul et al. (2015); Zheng et al. (2017).

We need to find transcriptomic factors related to CONTEXT. What are the transcriptomic factors genes related to GENE-X out of LIST-OF-TFs. Which of these 10 TFs have a causal relationship with GENE-X gene? Do not include any genes beyond these. Give potential candidates. Think step by step and return the answer in the format <Answer> [first suggestion, second suggestion, third suggestion, fourth suggestion and so on....] </Answer>. You have to return the 10 potential TFs that are related to the give gene only, otherwise your answer will be disqualified.

A.2.3 EXTRACTING POTENTIAL TFs FROM THE GENE LIST

we provide the LLM with a curated list of genes, accompanied by relevant contextual information such as biological function, tissue type, and experimental conditions derived from scRNA-seq datasets. The prompt explicitly requests the LLM to identify and propose TFs that are known to regulate the expression of each gene within the list. In our approach, we start with a total of 1000 genes and sequentially query the Large Language Model (LLM) about subsets of 20 genes at a time to extract potential transcription factors (TFs). For each initial query, the LLM identifies and proposes TFs that may regulate the selected genes, leveraging its extensive contextual knowledge of gene regulatory mechanisms. In subsequent prompts, we maintain a 50% overlap with the previous set of genes, ensuring that 10 of the genes are revisited while introducing 10 new genes. This iterative process allows us to refine the extraction of TFs, incorporating insights from the previous queries while continuously expanding our understanding of the regulatory landscape. By doing so, we aim

864 to capture a comprehensive set of potential TFs that interact with the entire gene pool, facilitating a
865 more robust inference of the Gene Regulatory Network (GRN).
866

867

868 We need to find transcriptomic factors related to CONTEXT out of given genes. What are
869 the transcriptomic factors genes out of LIST-OF-TFs. Which of these can be transcriptomic
870 factors? Do not include any TFs beyond these. Give potential candidates. Return the
871 answer in the format <Answer> [first suggestion, second suggestion, third suggestion, fourth
872 suggestion and so on....] </Answer>. Describe the reasoning first and then return answer in
873 requested format with the potential TFs.
874

875 B METRICS

876 B.1 STATISTICAL EVALUATION METRICS

877
878 **Euclidean Distance:** To compute the Euclidean distance between the centroids of the real and
879 simulated cells, we first calculate the centroid by finding the mean along the gene axis across all
880 simulated and real cells. The Euclidean distance $d(r, s)$ is then given by:
881

$$882 d(r, s) = \|\mu(R) - \mu(S)\|_2$$

883
884 Where R and S are matrix of real and simulated cells, with elements $R_{i,j}$ and $S_{i,j}$ where i indexes
885 the cells and j indexes the genes. $\mu(R)$ and $\mu(S)$ is the mean vector of the real and synthetic cells
886 along the gene axis (columns) respectively.
887

$$888 \mu(R) = \left(\frac{1}{n} \sum_{i=1}^n R_{i,j} \right)_{j=1, \dots, m} \quad \text{and} \quad \mu(S) = \left(\frac{1}{n} \sum_{i=1}^n S_{i,j} \right)_{j=1, \dots, m}$$

889
890 **Cosine Distance:** This computes the cosine distance between the centroids of the real and simulated
891 cells. The centroid is obtained by calculating the mean along the gene axis across all simulated and
892 real cells.
893

$$894 d_{\cos}(r, s) = 1 - \frac{\mu(R) \cdot \mu(S)}{\|\mu(R)\|_2 \|\mu(S)\|_2}$$

895
896 Cosine distance measures the difference in orientation of the two centroids while the Euclidean
897 distance measures the absolute difference between the two centroids. Since cosine focus on angle
898 between the vectors, it is useful in comparing the shape of data distributions rather than their scale.
899

900 **Maximum Mean Discrepancy (MMD):** Maximum Mean Discrepancy (MMD) serves as a non-
901 parametric two-sample test to determine if samples are drawn from the same distribution. MMD
902 metric is identified as a particularly convenient method for assessing the similarity of real data. We
903 followed the description in Zinat et al. Zinati et al. (2024).
904

905
906 **Discriminative Metric (RF AUROC):** This metric evaluates a model's ability to distinguish
907 between real and synthetic datasets. A random forest (RF) classifier is employed, and the area under
908 the receiver operating characteristic (AUROC) curve is used to assess whether real and simulated
909 cells can be effectively differentiated.
910
911

912
913
914
915
916
917

C ADDITIONAL RESULTS

C.1 STATISTICAL METRICS

Comparison of synthetically generated dataset with known causal graph approach. We include results for a synthetically generated dataset, LinearUniform (Table 4), to evaluate the impact of causality. Since the LinearUniform dataset was generated with a known causal graph, we applied GRNBoost2 as the causal inference method. For comparison, the non-causal model is represented in Stage 1. The results indicate that the causal model (GRNBoost2) outperforms across all three metrics.

	Cosine distance	Euclidean distance	MMD	RF AUROC
<i>LinearUniform</i>				
Control	0.00082 \pm 0.00012	108 \pm 8	0.0090 \pm 0.000	0.57 \pm 0.016
Stage 1	0.00097 \pm 0.00010	119 \pm 6	0.0139 \pm 0.001	0.64 \pm 0.020
GRNBoost2	0.00061 \pm 0.00011	93 \pm 9	0.0152 \pm 0.001	0.63 \pm 0.028

Table 4: Evaluation of synthetically generated dataset.

C.1.1 OVERLAPS BETWEEN HUMAN KB AND GPT-4 KB

Dataset	# TFs	% Overlap
PBMC	95	64.00
CTL	135	49.23
BM	157	55.82

Table 5: The overlaps of TFs between of Human KB and GPT-4 KB.

C.1.2 DENSITY OF GRAPHS

Dataset	# TFs	# Targets	# Genes	# Possible Edges	# Imposed Edges	GRN density Edges
<i>KB^H</i>						
PBMC	75	925	1000	69375	9250	0.1333
CTL	65	935	1000	60775	9350	0.1538
BoneMarrow	68	932	1000	63376	9320	0.1471
<i>KB^{GPT}</i>						
PBMC	95	905	1000	85975	9050	0.1052
CTL	135	865	1000	116775	8650	0.0740
BoneMarrow	157	843	1000	132351	8430	0.0639
<i>KB^{Llama}</i>						
PBMC	266	843	1000	224238	8430	0.0375

Table 6: The density of different causal graphs wrt different KB.

C.1.3 QUALITATIVE EVALUATION OF LLM PRIORS

The TFs that were missing from LLM list as compared to the human curated TFs list for the PBMC data set: "PLEK", "SNAI3", "ZNF264", "ZNF708", "ZNF277", "SMARCC1", "ZBTB38", "BAZ2A", "ARID5A", "TGIF1", "LYAR", "DNAJC1", "HOPX". These transcription factors (TFs) play diverse roles in regulating immune function, cell differentiation, and responses to stress or inflammation. Their expression in PBMCs highlights their importance in maintaining immune homeostasis, responding to infections, and their potential involvement in disease states such as cancer, autoimmune diseases, and inflammatory disorders.

Extra genes proposed as TFs by LLM: "TNFRSF18", "IL2RB", "NCOA4", "CHD8", "CDK9", "LMO4", "HEXIM1", "TAF10", "CCNK", "POLR2A", "BRD4", "ANKRD11". Notably the extra transcription

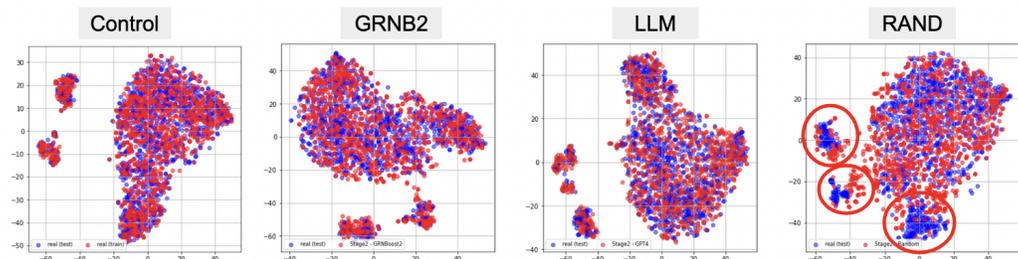


Figure 4: **TSNE projections of synthetic vs. real data for different GRN graphs (Setting 1).** “Control” corresponds to the projection of training and testing data, “GRNB2” -the synthetic data based on GRNBoost2 graph, “LLM” - synthetic data based on LLM graph, and “RAND” - the data based on the random graph. Red dots correspond to the real data points and blue ones - to the synthetic data points. “Hallucinated” extra blue clusters in the RAND graph are marked with red circles.

factors (TFs) proposed by LLM play key roles in regulating immune cell function, influencing various aspects of immune responses, disease states (like cancer and autoimmunity), and cell differentiation pathways.

The list with the missing TFs has more genes that are not as well-defined in the literature. Specifically, genes like SNAI3, ZNF264, ZNF708, and ZNF277 have limited characterization and fewer studies addressing their specific roles in cellular processes, especially in the context of immune cells and PBMCs. In contrast, many genes in the extra proposed TFs list (e.g., TNFRSF18, CDK9, BRD4) are more extensively studied and have clearer functions, particularly in transcriptional regulation and immune processes.

C.2 QUALITATIVE EVALUATION OF SYNTHETIC DATA SETS

In addition we perform a qualitative evaluation of the synthetic data sets by visualizing them using TSNE projections (Figure 4). We observe that using random GRN induces “hallucinated” extra clusters in the data, whereas both GRNB2 and LLM proposed graphs stay faithful to the original distribution.

C.3 BIOLOGICAL EVALUATION

Llama-KB GRNBoost2 Dataset and Improved Cell Type Segregation Compared to the original dataset, it was observed that the Llama-KB GRNBoost2 dataset segregates cell types more effectively. In the original dataset, four of the top five marker genes for CD4+/CD45RA+/CD25- Naïve T cells exhibited similar expression across multiple cell types, introducing noise that could complicate biological analysis. In contrast, the Llama-KB dataset demonstrated more distinct expression for CD4+/CD45RA+/CD25- Naïve T cells, with the exception of a single marker, LTB, which was expressed in other cell types as well.

Reduction of Noise in Other Cell Types in Llama-KB GRNBoost2Dataset Additionally, dendritic cells in the Llama-KB GRNBoost2 model exhibited a less noisy expression pattern compared to the original dataset, a trend also observed in CD8+ cytotoxic T cells and CD4+/CD25+ T regulatory cells. A notable difference was found in CD8+/CD45RA+ cytotoxic cells, where the original dataset showed little to no expression of top markers, while the Llama-KB dataset had a more widespread, noisy expression.

General Performance of Llama-KB GRNBoost2 in Cell-Specific Expression Profiles Although Llama-KB GRNBoost2 synthetic data Figure 3b introduces some noise, it generally improves cell-specific expression profiles, which could be further optimized for even cleaner cell type differentiation.

1026 A key observation was that when mean expression was higher in specific cell types, it tended to be
1027 lower in others, with fewer than 30% of cells displaying similar expression fractions. In contrast,
1028 when expression was noisy across multiple cell types, both mean expression and cell fraction tended
1029 to be similar across these groups.

1030
1031 **Comparison with GPT4-KB GRNBoost2 and Random Datasets** In comparison, the GPT4-KB
1032 GRNBoost2 dataset Figure 3a exhibited more noise than the Llama-KB dataset Figure 3b, with
1033 multiple markers expressed across various cell types at comparable mean expression levels and with
1034 higher cell fractions. The random GRN dataset resembled the Stage 1 non-causal dataset, with mean
1035 expression profiles being more clearly defined for each specific cell type. However, cell fractions for
1036 the same markers in other cell types were still significant, often exceeding 80%, indicating suboptimal
1037 gene expression specificity.

1038
1039 **Stage 1 Dataset and Noise Reduction in Markers** Stage 1 data demonstrated a less noisy expres-
1040 sion pattern, where markers were primarily confined to their specific cell types, showing lower mean
1041 expression in others. However, even though mean expression in non-specific cell types were low, cell
1042 fractions remained similar to those of the main cell type, which is not ideal for clear differentiation.

1043
1044 **Human-KB GRNBoost2 Dataset Performance and Noise Patterns** For the Human-KB GRN-
1045 Boost2d dataset, noise was primarily observed in a few cell types, including CD4+/CD45A+/CD25-
1046 Naïve T cells and CD8+/CD45RA+ Naïve cytotoxic T cells. Some markers from other cell types
1047 also showed similar expression across multiple cell types. These noisy patterns were not observed in
1048 the Stage 1 or random datasets and were less pronounced in the Llama-KB and GPT4-KB models,
1049 though Llama-KB GRNBoost2 outperformed GPT4-KBT in reducing noise for these specific cell
1050 types. Beyond these examples, most cell types in the Human-KB GRNBoost2 dataset did not exhibit
1051 significant noise.

1052
1053 **Human-KB LLM Model’s Noisy Expression Profiles** Finally, the Human-KB LLM model
1054 exhibited highly noisy or non-specific expression profiles, particularly for CD4+/CD45A+/CD25-
1055 Naïve T cells, CD8+/CD45RA+ Naïve cytotoxic T cells, CD4+ T helper cells, CD4+/CD25+ T
1056 regulatory cells, and CD4+/CD45RO+ memory cells. For other cell types, the model either failed to
1057 express most top markers or showed high cell fractions with reduced mean expression across different
1058 cell types.

1059
1060 **Variations in Cell Type Proportions Across Datasets** A notable observation is that the cell type
1061 proportions in each generated dataset differ significantly from those in the original dataset, indicating
1062 a noisy overall expression profile that can alter the distribution of cell types. In the Llama-KB
1063 GRNBoost2 dataset, CD8+/CD45RA+ Naïve Cytotoxic T cells were the most abundant at 34.1%,
1064 followed closely by CD56+ NK cells, CD8+ Cytotoxic T cells, and CD4+/CD25+ T regulatory cells.
1065 This trend was similarly observed in the Stage 1, Human-KB GRNBoost2, Human-KB LLM, and
1066 GPT4-KB GRNBoost2 datasets, albeit with slightly varying percentages.

1067
1068 **Discrepancies in Cell Type Proportions** The most considerable variation in proportions was
1069 noted for CD8+/CD45RA+ Naïve Cytotoxic T cells, with Llama-KB GRNBoost2 reporting the
1070 highest percentage at 34.1% and GPT-4 the lowest at 28.2%. Importantly, all generated datasets
1071 displayed significant differences when compared to the original dataset, where CD8+/CD45RA+
1072 Naïve Cytotoxic T cells constituted only 0.1% of the population, and CD4+/CD45RA+ Naïve T
1073 cells accounted for 65.1%. These findings suggest that the models manipulate the expression profiles
1074 in such a way that the proportions of CD8+/CD45RA+ Naïve Cytotoxic T cells, CD4+/CD45RA+
1075 Naïve T cells, and other cell types are elevated in the generated datasets.

1076
1077 **Overall Performance and Future Improvements for Llama-KB GRNBoost2** Therefore, it
1078 was observed that CD4+/CD45RA+/CD25- Naïve T Cells and CD8+/CD45RA+ Naïve Cytotoxic
1079 T Cells exhibited noisy expression patterns consistently across all datasets. The original dataset
performed better for CD8+/CD45RA+ Cytotoxic T Cells, showing lower noise in their expression
compared to the Llama-KB GRNBoost2 generated dataset. Among the various models, the LLAMA
model emerged as the best performer, providing clearer segregation of cell types and reduced noise,
particularly for CD4+/CD45RA+/CD25- Naïve T Cells and dendritic cells.

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

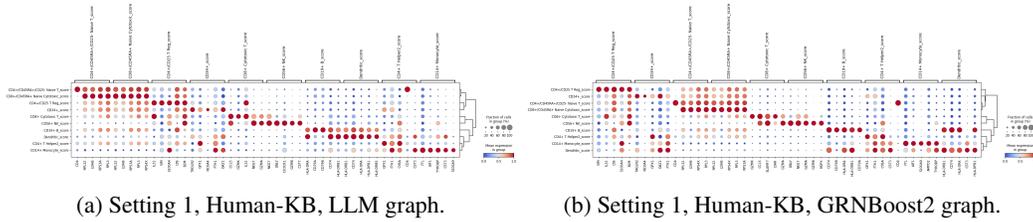


Figure 5: Gene expression analysis. Human-KB LLM and Human-KB GRNBoost2 datasets.

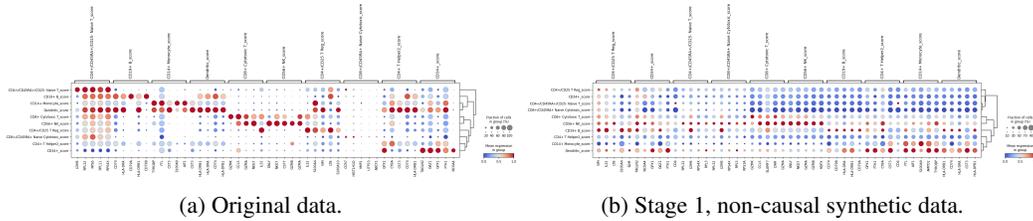


Figure 6: Gene expression analysis. Original and Stage 1 datasets.

Conclusion: Challenges and Future Refinements In contrast, the GPT4-KB GRNBoost2 model exhibited more noise than Llama-KB GRNBoost2, with similar markers expressed across various cell types at comparable mean expression levels. The Stage 1 model displayed less noisy expression patterns, but some markers were still expressed significantly across non-specific cell types, while the random-generated dataset had poorer cell-specific profiling. In conclusion, the discrepancies in cell-type proportions between generated datasets and the original dataset highlight the challenges in achieving accurate cell-type specificity. These variations can significantly impact downstream analyses and interpretations in biological research. Consequently, refining these models to enhance the specificity of marker expression is essential for producing more reliable datasets that reflect the true cellular composition observed in biological samples.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187

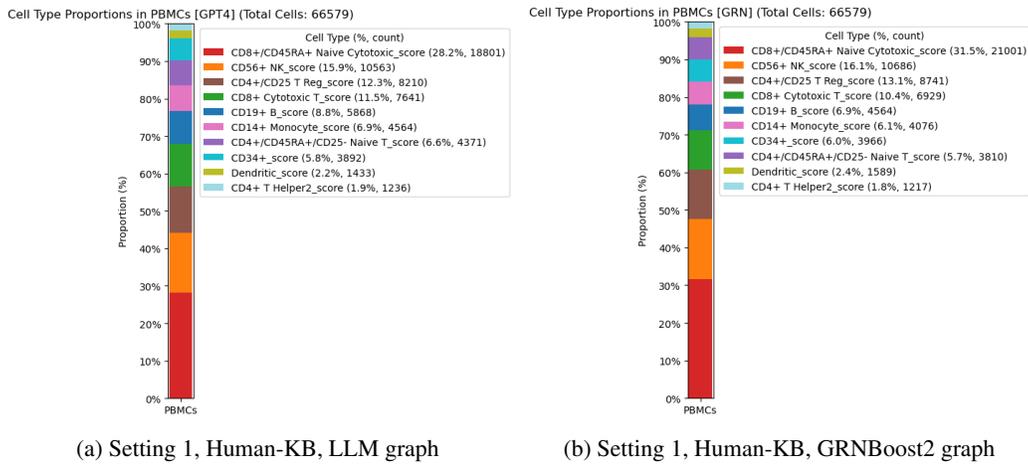


Figure 7: Cell annotation and proportions analysis. Human-KB LLM and Human-KB GRNBoost2 datasets.

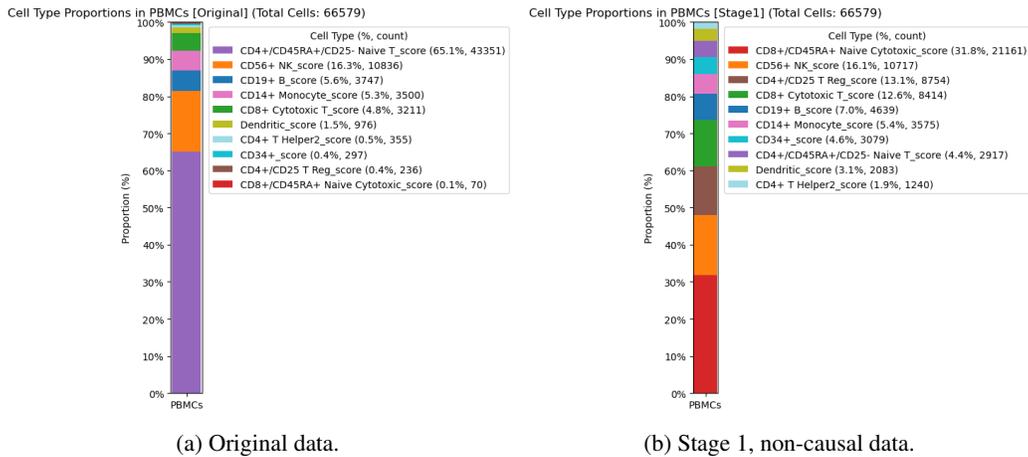


Figure 8: Cell annotation and proportions analysis. Original and Stage 1 datasets.

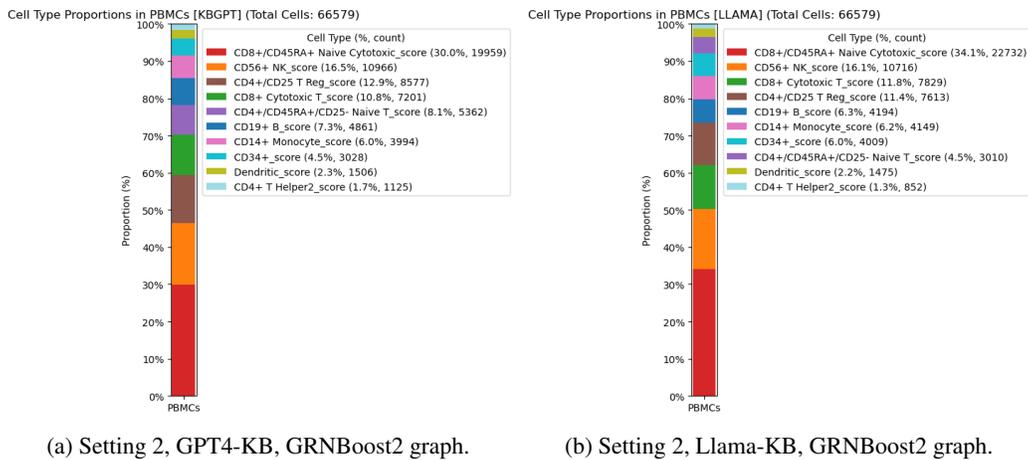


Figure 9: Cell annotation and proportions analysis. GPT4-KB LGRNBoost2 and Llama-KB GRNBoost2 datasets.