LLM4GRN: Discovering Causal Gene Regulatory Networks with LLMs - Evaluation through Synthetic Data Generation

Tejumade Afonja[1*], **Ivaxi Sheth**[1*], **Ruta Binkyte**[1*]& **Mario Fritz**[1]
[1]CISPA Helmholtz Center for Information Security,
Germany
{tejumade.afonja, ivaxi.sheth, ruta.binkyte-sadauskiene}@cispa.de

Waqar Hanif[2,3,4], **Shubhi Ambast**[2], **Charles Mwangi Kaumbutha**[2] & **Matthias Becker**[2]
[2]German Center for Neurodegenerative Diseases (DZNE),
[3]Das Life & Medical Sciences-Institut (LIMES),
[4]University of Bonn,
Germany

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 develop 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 Scheines et al. (1998); Zheng et al. (2018); Mercatelli et al. (2020); Brouillard et al. (2020); Lippe et al. (2021); Yu & Welch (2022); Roohani et al. (2024) 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).

[*]Equal Contribution

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 Kıcıman 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 for 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 in genomic data analysis Märtens et al. (2024); 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 GRNs from scRNA-seq data. Specifically, we utilize LLMs either to generate complete GRNs (causal graphs) directly or to provide prior knowledge in the form of a list of potential transcription factors (TFs), which can then be integrated into traditional statistical causal discovery algorithms.

Gene regulation is highly complex and context-dependent, with relationships between genes varying across cell types, disease states, developmental stages, and other factors. This variability makes it challenging to establish a definitive ground truth for GRNs. To address this, we use causal synthetic data generation as a downstream task to evaluate GRNs or priors suggested by LLMs in the absence of reliable ground-truth graphs. The causal GAN method integrates GRNs into the scRNA-seq data generation process and has been shown to better preserve biological plausibility Zinati et al. (2024). Moreover, causal GAN serves as an effective downstream task for assessing GRN quality, as poor GRNs result in unrealistic or biologically implausible synthetic data.

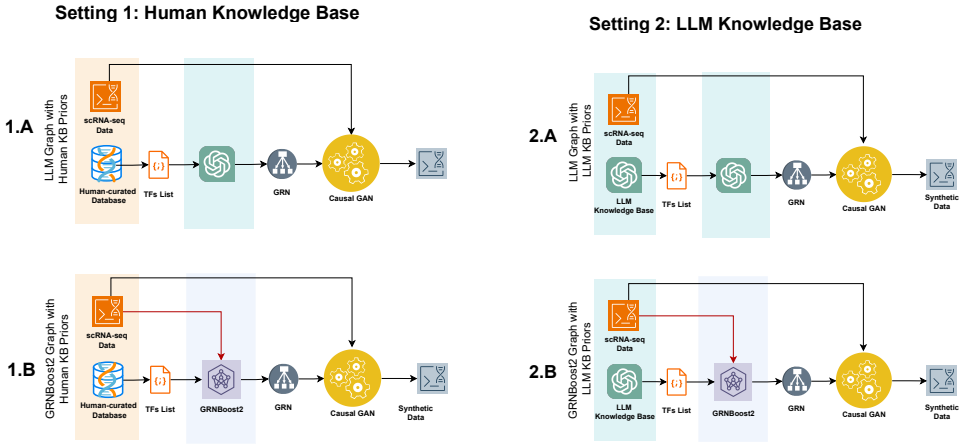
We assess the practical utility of the inferred causal graph by comparing the synthetic data to an oracle dataset, evaluating both its statistical similarity and biological plausibility. This approach allows testing on real-world data instead of relying solely on causal benchmark datasets. Although causal benchmark datasets are paired with ground truth graphs, their widespread use in causal research may lead to their incorporation in the training data, potentially inflating the models' perceived causal performance. Our results highlight the potential of general-purpose LLMs for GRN inference, particularly for the PBMC data, as evidenced by both statistical and biological metrics. The best performance is achieved by combining LLMs with statistical GRN inference, pointing to a promising direction for scRNA-seq data analysis. The related work and preliminaries on Directed acyclic graphs (DAGs), gene regulatory networks (GRNs) and GroundGAN algorithm can be found in the Appendices A and B.

Our contributions:

1. We show the effectiveness of using out of the box Large Language Models (LLMs) for inferring GRNs, showcasing their ability to capture complex biological interactions.
2. Comprehensive performance evaluation via synthetic data generation, enabling rigorous assessment of inference methods despite the lack of ground-truth causal graphs.
3. Extensive biological insight on the best-performing LLM for PBMC-All dataset.

2 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 (causal) 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 setting, we introduce LLMs into the pipeline motivated by their ability to incorporate extensive contextual information.



(a) Setting 1 (b) Setting 2

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.

- 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 are used by a statistically causal GRNBoost2 approach, respectively.

2.1 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:

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

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 $\mathbf{R}_j$.

2.2 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 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 $\mathcal{F}_{\text{LLM}}$ that, given a set of metadata $\mathcal{M}$, produces a bipartite graph $\mathcal{G} = (\mathbf{T}, \mathbf{R}, \mathcal{E})$, expressed as:

$$\mathcal{F}_{\text{LLM}}(\mathcal{M}, \mathbf{T}^H, \mathbf{R}^H) = \mathcal{G} \tag{1}$$

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 C.2.

2.3 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.

$$\mathcal{F}_{\text{LLM}}(\mathcal{M}) = \left(\mathbf{T}^{\text{LLM}}, \mathbf{R}^{\text{LLM}}\right) \tag{2}$$

Here, $\mathbf{T}^{\text{LLM}}$ denotes the set of transcription factors identified from the LLM knowledge base, and $\mathbf{R}^{\text{LLM}}$ 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.

2.4 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.

3 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.

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 AppendixC.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 KB^{H} or KB^{LLM}). 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 C.2) in zero-shot fashion. We also compare open-source model, Llama-3.1-70B Dubey et al. (2024).

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 the biological plausibility of the datasets by conducting single-cell RNA sequencing analysis using the Scanpy Python library. The cell annotations from the original dataset were utilized to annotate the synthetic datasets. Our analysis included log1p normalization and scaling to 10,000, focusing on 1,000 highly variable features. We performed dimensionality reduction using PCA with 50 principal components, followed by Uniform Manifold Approximation and Projection (UMAP). Cell type proportion analysis was conducted with custom code, while gene expression profiling for the top markers of each cell type was visualized using dot plots. For more details, refer to Appendices C and D.

Datasets. For the direct comparison of the different settings, we focus on the dataset and genomic database information by Zinati et al. (2024). Specifically, we used PBMC-All, PBMC-CTL and BoneMarrow data sets, that span (details in Appendix C.1.1). Importantly, the datasets cover a wide range of biological diversity, spanning both human and animal samples as well as single-cell and multicell types. Our objective is to assess the performance of large language models (LLMs) in gene regulatory network (GRN) inference. To evaluate the utility of the inferred GRNs, we employ causal synthetic data generation as a downstream task in scenarios where a reliable ground truth graph is unavailable.

3.1 Comparision against baseline

Evaluation of GPT-4 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 GPT-4 . 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 GPT-4 -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 GPT-4 -generated GRN and the random GRN is smaller than between GPT-4 GRN and GRNBoost2 GRN, suggesting, that the GPT-4 could be generating meaningful graphs.

3.2 Statistical Evaluation of GRN Inference Methods on Synthetic Data

In Table 1 we present the statistical metrics results for the PBMC-ALL data set. Additional results on all three data sets can be found in Appendix E,Table 4. 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 (↓) 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 (using KB^H, Table 1) result shows the GPT-4-inferred GRN achieves the best performance across all metrics 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 GPT-4 in this setting (see Appendix E).

Setting 2 Comparison. Table 1 presents the results for Setting 2, where we incorporate LLM-derived transcription factor (TF) lists into GRN inference. When using GPT-4 as the knowledge base (KB^{GPT4}), GRNBoost2—combining GPT-4-proposed TF lists with statistical GRN inference—outperforms the fully GPT-4-based GRN approach (where the LLM suggests both the TF list and the TF-gene interactions) across all datasets. GRNBoost2 also achieves the best overall performance among KB^{GPT4} approaches, delivering substantial improvements in PBMC-ALL, where it reduces the Euclidean distance from 121 (Setting 1) to 83 and improves RF AUROC from 0.73 to 0.59.

However, Llama-3.1-70B (KB^{Llama}) demonstrates the best overall performance in Setting 2, surpassing GPT-4 across all metrics. When Llama-3.1-70B is used as the knowledge base, GRNBoost2 achieves the most accurate results, further validating the effectiveness of integrating LLM-derived priors with statistical inference. This trend also holds for PBMC-CTL, where GRNBoost2 consistently improves upon its Setting 1 performance. BoneMarrow performed best in Setting 1 but remains competitive in Setting 2 when the KB^{GPT4} is combined with GRNBoost2 (Appendix E, Table 4).

	Cosine distance ↓	**Euclidean distance ↓**	**MMD ↓**	**RF AUROC ↓**
Baseline				
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
KB^{Random}	0.00132±0.00101	187±79	0.0214±0.009	0.86±0.080
Setting 1 KB^{H}				
GPT-4	<u>0.00024</u>±0.00004	<u>89</u>±7	<u>0.0072</u>±0.0012	<u>0.63</u>±0.04
Llama-3.1	0.00114±0.00048	193±52	0.0244±0.0075	0.84±0.04
GRNBoost2	0.00047±0.00021	121±25	0.0139±0.0058	0.73±0.05
Random	0.00045±0.00013	121±22	0.0166±0.0031	0.85±0.02
Setting 2 KB^{GPT4}				
GPT-4	0.00026±0.00009	90±13	0.0206±0.0025	0.86±0.02
GRNBoost2	<u>0.00023</u>±0.00008	**83**±17	<u>0.0069</u>±0.0012	<u>0.59</u>±0.03
Random	0.00026±0.00011	92±14	0.0226±0.0020	0.87±0.02
Setting 2 KB^{Llama}				
Llama-3.1	0.00029±0.00005	97±10	0.0100±0.0007	0.77±0.030
GRNBoost2	**0.00022**±0.00006	**83**±11	**0.0067**±0.0011	**0.58**±0.023
Random	0.00035±0.00009	105±17	0.0152±0.0011	0.86±0.028

Table 1: **Comparison of KB^{Llama} - performance GRN Inference Methods in Simulating Realistic scRNA-seq Data for PBMC-ALL dataset.** The best value is presented in **boldface** and the best value in each setting is <u>underline</u>. The lower (↓) the better for all metrics.

3.3 Biological Plausibility of Causal Synthetic Data

We conduct cell-type annotation based on the original dataset, gene expression analysis to identify top markers for each cell type and cell-type proportion analysis on the most statistically robust datasets. Full analysis can be found in Appendix D (subsection E.4).

3.3.1 Gene Expression Profiling Per Cell Type

General Performance of KB^{Llama} GRNBoost2 in Cell-Specific Expression Profiles. The KB^{Llama} GRNBoost2 model shows that specific cell types, such as CD8+/CD45R+ naive cytotoxic T cells and CD4+ T helper cells, exhibit significantly higher mean expression levels for the same markers across multiple cells while maintaining similar cell fraction of expression. Additionally, the KB^{Llama} GRNBoost2 model demonstrates superior performance in elucidating expression patterns in certain cases, such as with dendritic cells. Although KB^{Llama} GRNBoost2 synthetic data Figure 4a introduces some noise, it generally improves cell-specific expression profiles, which could be further optimized for even cleaner cell type differentiation. A key observation was that when mean expression was higher in specific cell types, it tended to be lower in others, with fewer than 30% of cells displaying similar expression fractions. In contrast, when expression was noisy across multiple cell types, both mean expression and cell fraction tended to be similar across these groups.

KB^{GPT4} GRNBoost2 Dataset Performance and Noise Patterns. The KB^{GPT4} GRNBoost2 model identifies top marker genes, revealing notable differences in expression profiles. While expression levels among some markers are noisy and indistinguishable, this model effectively highlights markers that are more discriminative for certain cell types. The KB^{GPT4} GRNBoost2 dataset Figure 7b exhibited more noise than the KB^{Llama} dataset Figure 4a, with multiple markers expressed across various cell types at comparable mean expression levels and with higher cell fractions. For the KB^{H} GRNBoost2 dataset (Figure 4c), the noise was primarily observed in a few cell types, including CD4+/CD45A+/CD25– Naïve T cells and CD8+/CD45RA+ Naïve cytotoxic T cells. Some markers from other cell types also showed similar expression across multiple cell types. These noisy patterns were not observed in Stage 1 or random datasets and were less pronounced in the KB^{Llama} and KB^{GPT4} models, though KB^{Llama} GRNBoost2 outperformed KB^{GPT4} GRNBoost2 in reducing noise for these specific cell types.

3.3.2 Differential Cell Type Proportion Analysis

Variations in Cell Type Proportions Across Datasets. A notable observation is that the cell type proportions in each generated dataset differ significantly from those in the original dataset, indicating a noisy overall expression profile that alters distribution of cell types. In the KB^{Llama} GRNBoost2 dataset (Figure 8b), CD8+/CD45RA+ Naïve Cytotoxic T cells were the most abundant at 34%, followed closely by CD56+ NK cells, CD8+ Cytotoxic T cells, and CD4+/CD25+ T regulatory cells. This trend was similarly observed in KB^{GPT4} GRNBoost2 (Figure 8a) and other datasets, albeit with slightly varying percentages.

Overall Performance of KB^{Llama} GRNBoost2. CD4+/CD45RA+/CD25– Naïve T Cells and CD8+/CD45RA+ Naïve Cytotoxic T Cells exhibited consistently noisy expression patterns across all datasets. The original dataset performed better for CD8+/CD45RA+ Cytotoxic T Cells, showing lower noise in their expression compared to the KB^{Llama} GRNBoost2 (Figure 8b) dataset. Among the various models, the KB^{Llama} GRNBoost2 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. In contrast, the KB^{GPT4} GRNBoost2 (Figure 8a) model exhibited more noise than KB^{Llama} GRNBoost2, with similar markers expressed across various cell types at comparable mean expression levels.

3.4 Discussion and Limitations

We observe promising results in using LLM for GRN discovery, especially on the PBMC dataset. The hybrid approach incorporating TFs suggested by LLM (KB^{LLM}) and GRNBoost2 causal discovery yield the best overall performance (for PBMC-ALL and PBMC-CTL data sets). Surprisingly, a smaller open-source model Llama has shown the best result in this setting. One possible explanation could be the number of TFs proposed in Setting 2.

The number of TFs in KB^{Llama} is higher than those in KB^{H} or KB^{GPT4}. Our additional experiments (subsection E.2) indicate that the number of transcription factors (TFs) may be significant, as it provides a broader set of variables for the GRNBoost2 algorithm to evaluate during causal discovery. Nevertheless, although further evaluation is needed, new Llama models might perform as good or better than state-of-the-art GPT for specific tasks Valero-Lara et al. (2023). However, Llama performs worse on the more challenging, GRN construction task. The lower performance of LLMs on CTL and BoneMarrow data sets is not surprising. Recent studies suggest that genomic LLMs under-perform on cell-specific data Tang & Koo (2023). In addition, human transcription factors outnumber mice transcriptomic information in the databases Members & Partners (2024) making the information on mice genes scarcer in the LLMs training data.

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

In contrast, while the Setting 2, KB^{GPT4} model exhibited higher noise and less specificity, the Setting 1, KB^{H}, GRNBoost2 model also showed noisy patterns, especially in Naïve T and cytotoxic T cells, emphasizing the need for models that clearly delineate cell type-specific expression. Notably, discrepancies in cell type proportions across generated datasets, particularly the inflated presence of CD8+/CD45RA+ Naïve Cytotoxic T cells, raise concerns about the biological relevance of these models. As reported previously, large-scale single-cell studies tend to be more noisy which can lead to sub-optimal biological inferences Kavran & Clauset (2021). Therefore, further refinements are essential across all approaches to enhance specificity and reliability in representing true cellular compositions, which is crucial for accurate downstream biological analyses and interpretations. Finally, the KB^{H} GPT-4 model exhibited highly noisy or non-specific expression profiles, particularly for CD4+/CD45A+/CD25– Naïve T cells, CD8+/CD45RA+ Naïve cytotoxic T cells, CD4+ T helper cells, CD4+/CD25+ T regulatory cells, and CD4+/CD45RO+ memory cells. For other cell types, the model either failed to express most top markers or showed high cell fractions with reduced mean expression across different cell types.

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

Despite promising results, LLMs should be used cautiously in high-stakes decision-making, as they can confidently generate false information Ji et al. (2023); Farquhar et al. (2024) and inherit biases from training data. A key issue in machine learning is the underrepresentation of minority populations, which also affects transcription factor databases and literature, often lacking diversity in ethnicity, ancestry, gender, and age. Most genetic studies are based on individuals of European ancestry Sirugo et al. (2019); Bentley et al. (2017), potentially overlooking gene-disease associations in other populations and leading to less effective treatments Landry et al. (2018); Tawfik et al. (2023); Barral-Arca et al. (2019). Additionally, children, the elderly, females, low-income populations, and rural communities are underrepresented in clinical trials, limiting their access to precision medicine Mosenifar (2007); Davis et al. (2019); Steinberg et al. (2021). LLMs should be used with caution to avoid exacerbating existing health disparities Pfohl et al. (2024), as they can be influenced by biases present in genomic data.

4 Conclusion

In conclusion, our analysis demonstrates the potential of Large Language Models (LLMs) for gene regulatory network (GRN) inference. The results are most promising for PBMC data, where LLM-based approaches achieve the highest accuracy and biological plausibility. Notably, the best-performing approach combines LLM-derived TF priors with GRNBoost2 for statistical inference. In addition, we see good potential for using open-source models such as Llama for GRNs, and we aim to explore their utility further, either directly or with fine-tuning. Our biological results suggest that additional refinement is needed. For this we foresee developing GRNs tailored to individual cell types, as evidence indicates that GRNs are often cell-type-specific.

Acknowledgements

This work is funded in part by Bundesministeriums fur Bildung und Forschung (PriSyn), grant No. 16KISAO29K. The work is also supported by Medizininformatik-Plattform "Privatsphären-schutzende Analytik in der Medizin" (PrivateAIM), grant No. 01ZZ2316G, and ELSA – European Lighthouse on Secure and Safe AI funded by the European Union under grant agreement No. 101070617. Moreover, the computation resources used in this work are supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Ahmed Abdulaal, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C Castro, Daniel C Alexander, et al. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4. *arXiv*, 2023.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. Causal structure learning supervised by large language model. *arXiv*, 2023a.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv*, 2023b.

Soma Bandyopadhyay and Sudeshna Sarkar. Exploring causality aware data synthesis. In *Proceedings of the Third International Conference on AI-ML Systems*, pp. 1–9, 2023.

Ruth Barral-Arca, Jacobo Pardo-Seco, Xabi Bello, Federico Martinon-Torres, and Antonio Salas. Ancestry patterns inferred from massive rna-seq data. *RNA*, 25(7):857–868, 2019.

Amy R Bentley, Shawneequa Callier, and Charles N Rotimi. Diversity and inclusion in genomic research: why the uneven progress? *Journal of community genetics*, 8:255–266, 2017.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *NeurIPS*, 2020.

Martina Cinquini, Fosca Giannotti, and Riccardo Guidotti. Boosting synthetic data generation with effective nonlinear causal discovery. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pp. 54–63. IEEE, 2021.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

Terry C Davis, Connie L Arnold, Glenn Mills, and Lucio Miele. A qualitative study exploring barriers and facilitators of enrolling underrepresented populations in clinical trials and biobanking. *Frontiers in Cell and Developmental Biology*, 7:74, 2019.

Natalie de Souza. The encode project. *Nature methods*, 9(11):1046–1046, 2012.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jonas Elsborg and Marco Salvatore. Using llms and explainable ml to analyze biomarkers at single-cell level for improved understanding of diseases. *Biomolecules*, 13(10):1516, 2023.

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.

Hui Hu, Ya-Ru Miao, Long-Hao Jia, Qing-Yang Yu, Qiong Zhang, and An-Yuan Guo. Animaltfdb 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research*, 47(D1):D33–D38, 2019.

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. Prollm: Protein chain-of-thoughts enhanced llm for protein-protein interaction prediction. *bioRxiv*, pp. 2024–04, 2024.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *arXiv*, 2024.

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

Tejas Kasetty, Divyat Mahajan, Gintare Karolina Dziugaite, Alexandre Drouin, and Dhanya Sridhar. Evaluating interventional reasoning capabilities of large language models. *arXiv preprint arXiv:2404.05545*, 2024.

A. J. Kavran and A. Clauset. Denoising large-scale biological data using network filters. *BMC Bioinformatics*, 22:157, 2021. doi: 10.1186/s12859-021-04075-x.

Elahe Khatibi, Mahyar Abbasian, Zhongqi Yang, Iman Azimi, and Amir M Rahmani. Alcm: Autonomous llm-augmented causal discovery framework. *arXiv preprint arXiv:2405.01744*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Steven Kleinegesse, Andrew R Lawrence, and Hana Chockler. Domain knowledge in a*-based causal discovery. *arXiv preprint arXiv:2208.08247*, 2022.

Latrice G Landry, Nadya Ali, David R Williams, Heidi L Rehm, and Vence L Bonham. Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs*, 37(5):780–785, 2018.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *arXiv preprint arXiv:2107.10483*, 2021.

Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, 2015:bav095, 2015.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Jing Ma. Causal inference with large language model: A survey. *arXiv preprint arXiv:2409.09822*, 2024.

Kaspar Märtens, Rory Donovan-Maiye, and Jesper Ferkinghoff-Borg. Enhancing generative perturbation models with llm-informed gene embeddings. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

National Genomics Data Center Members and Partners. Database commons: A catalog of databases in life sciences. `https://ngdc.cncb.ac.cn/databasecommons/stat`, 2024. Accessed: 2024-09-27.

Daniele Mercatelli, Laura Scalambra, Luca Triboli, Forest Ray, and Federico M Giorgi. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194430, 2020.

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.

Zab Mosenifar. Population issues in clinical trials. *Proceedings of the American Thoracic Society*, 4(2):185–188, 2007.

Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, pp. 1–11, 2024.

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.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Ivaxi Sheth, Sahar Abdelnabi, and Mario Fritz. Hypothesizing missing causal variables with llms. *arXiv preprint arXiv:2409.02604*, 2024.

Giorgio Sirugo, Scott M Williams, and Sarah A Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019.

Jecca R Steinberg, Brandon E Turner, Brannon T Weeks, Christopher J Magnani, Bonnie O Wong, Fatima Rodriguez, Lynn M Yee, and Mark R Cullen. Analysis of female enrollment and participant sex by burden of disease in us clinical trials between 2000 and 2020. *JAMA Network Open*, 4(6):e2113749–e2113749, 2021.

Ziqi Tang and Peter K Koo. Building foundation models for regulatory genomics requires rethinking large language models. In *Proceedings of the ICML Workshop on Computational Biology*, 2023.

Sherouk M Tawfik, Aliaa A Elhosseiny, Aya A Galal, Martina B William, Esraa Qansuwa, Rana M Elbaz, and Mohamed Salama. Health inequity in genomic personalized medicine in underrepresented populations: a look at the current evidence. *Functional & Integrative Genomics*, 23(1):54, 2023.

Mohammed Toufiq, Darawan Rinchai, Eleonore Bettacchioli, Basirudeen Syed Ahamed Kabeer, Taushif Khan, Bishesh Subba, Olivia White, Marina Yurieva, Joshy George, Noemie Jourde-Chiche, et al. Harnessing large language models (llms) for candidate gene prioritization and selection. *Journal of Translational Medicine*, 21(1):728, 2023.

Pedro Valero-Lara, Alexis Huante, Mustafa Al Lail, William F Godoy, Keita Teranishi, Prasanna Balaprakash, and Jeffrey S Vetter. Comparing llama-2 and gpt-3 llms for hpc kernels generation. *arXiv preprint arXiv:2309.07103*, 2023.

Boris Van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela Van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023.

Guangya Wan, Yuqi Wu, Mengxuan Hu, Zhixuan Chu, and Sheng Li. Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions. *arXiv preprint arXiv:2402.11068*, 2024.

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), 2024a.

Zhehao Wang, Arran Zeyu Wang, David Borland, and David Gotz. Causalsynth: An interactive web application for synthetic dataset generation and visualization with user-defined causal relationships. *IEEE VIS Posters*, 6, 2024b.

Bingyang Wen, Luis Oliveros Colon, KP Subbalakshmi, and Rajarathnam Chandramouli. Causal-tgan: Generating tabular data using causal generative adversarial networks. *arXiv preprint arXiv:2104.10680*, 2021.

Edgar Wingender, Peter Dietze, Holger Karas, and Rainer Knüppel. Transfac: a database on transcription factors and their dna binding sites. *Nucleic acids research*, 24(1):238–241, 1996.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

Hengshi Yu and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *BioRxiv*, pp. 2022–07, 2022.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*, 2024.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.

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.

Yazdan Zinati, Abdulrahman Takiddeen, and Amin Emad. Groundgan: Grn-guided simulation of single-cell rna-seq data using causal generative adversarial networks. *Nature Communications*, 15(1):4055, 2024.

A Related Works

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

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

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

B Preliminaries

B.1 Gene Regulatory Network

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, 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) **??**. In this work we are considering a two layer DAG consisting of transcription factors (TFs) and target genes. TFs are proteins that bind to specific DNA sequences to regulate the transcription and influence the expression levels of target genes. GRNs are crucial for understanding how cells respond to changes in their environment, how different cell types develop, and how malfunctions in these networks can lead to diseases like cancer or Alzheimer's disease.

B.2 GroundGan

GRouNdGAN Zinati et al. (2024) is a deep learning model that generates simulated single-cell RNA-seq data by leveraging causal generative adversarial networks (CausalGAN) and imposed user-defined causal gene regulatory networks (GRNs). The model comprises a causal

controller, target generators, a critic, a labeler, and an anti-labeler, each implemented as separate neural networks. Training involves two steps: pre-training the causal controller using a Wasserstein GAN (WGAN) to generate TF expression values resembling reference data, followed by training the target generators to produce target gene expressions based on these TF values and random noise. The generated expressions undergo library-size normalization (LSN). The critic quantifies the Wasserstein distance between reference and simulated data, while the target generators are adversarially trained to generate realistic data points. The labeler and anti-labeler ensure that target gene expressions encode regulatory information from TFs.

B.3 GRNBoost2

GRNBoost2 Moerman et al. (2019) is a scalable algorithm for inferring gene regulatory networks (GRNs) from single-cell RNA sequencing (scRNA-seq) data. It is part of the arboreto framework and is widely used due to its efficiency and scalability, especially for large datasets. GRNBoost2 employs gradient boosting, a machine learning method, to model the relationships between transcription factors (TFs) and target genes.

C Experimental setup

C.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`.

After downloading the raw data, we followed the pre-processing steps using scanpy as mentioned in the paper. In each dataset, cells with nonzero counts in fewer than ten genes were excluded to eliminate low-quality cells. Genes expressed in less than three cells across the dataset were discarded to reduce noise. We selected 1000 highly variable genes using dispersion based-method. After pre-processing, we split each of the datasets into train, validation and test sets. Following the paper's recommendation. we used a test set of 1000 cells for the PBMC and CTL datasets, and 500 cells for BoneMarrow.

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

C.1.1 Data Sets

- Human peripheral blood mononuclear cell (PBMC-Al). 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
COVID	56	944	1000	52864	9440	0.1786

Table 2: The density of the causal graphs.

C.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.

C.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.

C.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). [Prompt] 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.

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

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

D METRICS

D.1 STATISTICAL EVALUATION METRICS

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

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

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

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

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

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

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

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

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

E Additional Results

E.1 Statistical Metrics

Comparison of synthetically generated dataset with known causal graph approach. We include results for a synthetically generated dataset, LinearUniform (Table 3), 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**
LinearUniform				
Control	0.00082±0.00012	108±8	0.0090±0.000	0.57±0.016
Stage 1	0.00097±0.00010	119±6	**0.0139**±0.001	0.64±0.020
GRNBoost2	**0.00061**±0.00011	**93**±9	0.0152±0.001	**0.63**±0.028

Table 3: **Evaluation of synthetically generated dataset.**

E.2 Llama-3.1 Ablation

We carried out ablation of using Llama-3.1 as knowledge base to study the impact of size of TFs used for GRN creation.The result shown in Table 1 indicate that the number of transcription factors (TFs) may be significant, as it provides a broader set of variables for the GRNBoost2 algorithm to evaluate during causal discovery.

E.2.1 Overlaps between Human KB and GPT-4 KB

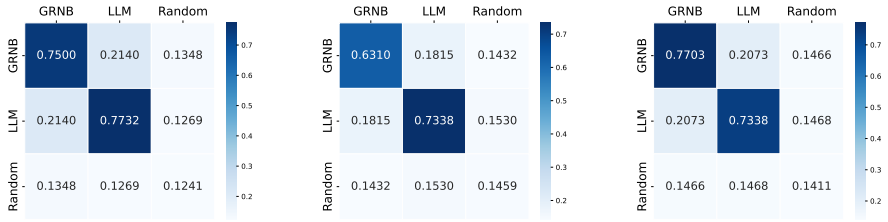
From the Table 6, we observe lower TF overlap between Human KB and GPT-4 KB for CTL than PBMC, suggesting that GPT-4 might be struggling to

E.2.2 Density of graphs

In the Table 7, we observe that LLama suggests more TFs than other methods making the GRNs sparser.

E.3 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 3). 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. The overlap between GRNs proposed by different inference methods can be found in the Table 2.



(a) PBMC-All (b) PBMC-CTL (c) Bone Marrow

Figure 2: Overlap between GRNs proposed by different methods. LLM demonstrates a higher self-overlap compared to GRNBoost2 algorithm.

		Cosine distance ↓	Euclidean distance ↓	MMD ↓	RF AUROC ↓
PBMC-ALL	*Baseline*				
	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
	KB^{Random}	0.00132±0.00101	187±79	0.0214±0.009	0.86±0.080
	Setting 1 KB^{H}				
	GPT-4	<u>0.00024</u>±0.00004	<u>89</u>±7	<u>0.0072</u>±0.001	<u>0.63</u>±0.043
	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
	Setting 2 KB^{GPT4}				
	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	*Baseline*				
	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
	KB^{Random}	0.00016±0.00002	51±3	0.0071±0.000	0.73±0.022
	Setting 1 KB^{H}				
	GPT-4	0.00265±0.00253	176±115	0.0238±0.020	0.79±0.206
	GRNBoost2	<u>0.00025</u>±0.00004	<u>65</u>±6	<u>0.0053</u>±0.000	**0.59**±0.028
	Random	0.00027±0.0001	66±7	0.0085±0.001	0.75±0.021
	Setting 2 KB^{GPT4}				
	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	*Baseline*				
	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
	KB^{Random}	0.00206±0.00034	80±7	0.0156±0.002	0.75±0.048
	Setting 1 KB^{H}				
	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
	Setting 2 KB^{GPT4}				
	GPT-4	0.00217±0.00050	82±10	0.0137±0.003	0.72±0.044
	GRNBoost2	<u>0.00193</u>±0.00023	<u>78</u>±4	<u>0.0119</u>±0.001	**0.64**±0.033
	Random	0.00297±0.00080	96±13	0.0172±0.004	0.79±0.046

Table 4: **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 <u>underline</u>. The lower (↓) the better for all metrics.

	Cosine distance ↓	Euclidean distance ↓	MMD ↓	RF AUROC ↓
Setting 2 KB^{Llama}				
GRNBoost2 (Tf=266)	**0.00022**±0.00006	**83**±11	**0.0067**±0.0011	**0.58**±0.023
GRNBoost2 (Tf=95)	<u>0.00032</u>±0.00008	103±17	<u>0.0083</u>±0.0006	<u>0.65</u>±0.024
GRNBoost2 (Tf=75)	0.00034±0.00023	<u>102</u>±38	0.0097±0.0029	0.67±0.043

Table 5: **Comparison of KB^{Llama} with varying number of TFs - performance of GRNBoost2 Inference Method in Simulating Realistic scRNA-seq Data for PBMC-ALL dataset.** The best value is presented in **boldface**. The lower (↓) the better for all metrics.

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

Table 6: The overlaps of TFs between of KB^{H} and KB^{GPT4}.

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

Table 7: The density of different causal graphs wrt different KB.

Evaluation of KB^{GPT4}. 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 6 (Appendix E), 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^{GPT4} 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 2 (Appendix E), would change the density of the graphs potentially affecting the downstream tasks.

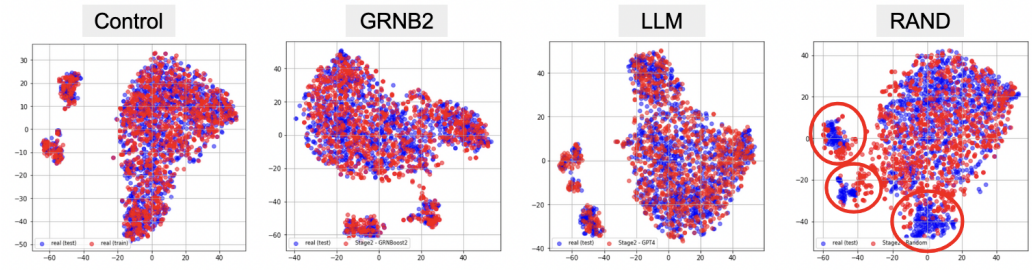


Figure 3: **TSNE projections of synthetic vs. real data for different GRN graphs (Setting 1)**. "Control" corresponds to the projection of training and testing data, "GRNB2" -the synthetic data based on GRNBoost2 graph, "LLM" - synthetic data based on LLM graph, and "RAND" - the data based on the random 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.

E.4 Biological evaluation

KB^{Llama}GRNBoost2 Dataset and Improved Cell Type Segregation Compared to the original dataset, it was observed that the KB^{Llama} 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 KB^{Llama} 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 KB^{Llama} GRNBoost2Dataset Additionally, dendritic cells in the KB^{Llama} 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 KB^{Llama} dataset had a more widespread, noisy expression.

General Performance of KB^{Llama} GRNBoost2 in Cell-Specific Expression Profiles Although KB^{Llama} GRNBoost2 synthetic data Figure 4a introduces some noise, it generally improves cell-specific expression profiles, which could be further optimized for even cleaner cell type differentiation. A key observation was that when mean expression was higher in specific cell types, it tended to be lower in others, with fewer than 30% of cells displaying similar expression fractions. In contrast, when expression was noisy across multiple cell types, both mean expression and cell fraction tended to be similar across these groups.

Comparison with KB^{GPT4} B GRNBoost2 and Random Datasets In comparison, the KB^{GPT4} GRNBoost2 dataset Figure 7b exhibited more noise than the KB^{Llama} dataset Figure 4a, with multiple markers expressed across various cell types at comparable mean expression levels and with higher cell fractions. The random GRN dataset resembled the Stage 1 non-causal dataset, with mean expression profiles being more clearly defined for each specific cell type. However, cell fractions for the same markers in other cell types were still significant, often exceeding 80%, indicating suboptimal gene expression specificity.

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

KB^{H} GRNBoost2 Dataset Performance and Noise Patterns For the KB^{H} GRNBoost2d dataset, noise was primarily observed in a few cell types, including CD4+/CD45A+/CD25– Naïve T cells and CD8+/CD45RA+ Naïve cytotoxic T cells. Some markers from other cell types also showed similar expression across multiple cell types. These noisy patterns were not observed in the Stage 1 or random datasets and were less pronounced in the KB^{Llama} and KB^{GPT4} models, though KB^{Llama} GRNBoost2 outperformed KB^{GPT} in reducing noise for these specific cell types. Beyond these examples, most cell types in the KB^{H} GRNBoost2 dataset did not exhibit significant noise.

KB^{H} LLM Model's Noisy Expression Profiles Finally, the KB^{H} LLM model exhibited highly noisy or non-specific expression profiles, particularly for CD4+/CD45A+/CD25– Naïve T cells, CD8+/CD45RA+ Naïve cytotoxic T cells, CD4+ T helper cells, CD4+/CD25+ T regulatory cells, and CD4+/CD45RO+ memory cells. For other cell types, the model either failed to express most top markers or showed high cell fractions with reduced mean expression across different cell types.

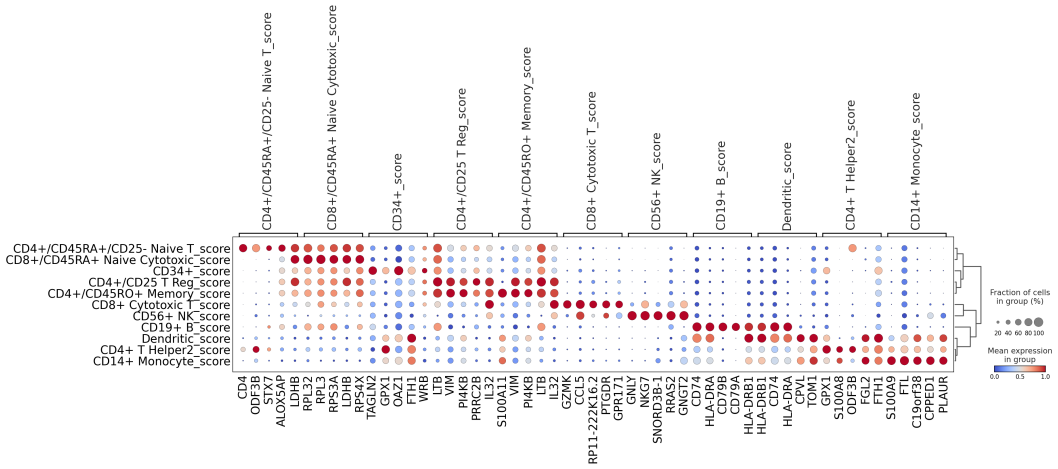
Variations in Cell Type Proportions Across Datasets A notable observation is that the cell type proportions in each generated dataset differ significantly from those in the original dataset, indicating a noisy overall expression profile that can alter the distribution of cell types. In the KB^{Llama} GRNBoost2 dataset, CD8+/CD45RA+ Naïve Cytotoxic T cells were the most abundant at 34.1%, followed closely by CD56+ NK cells, CD8+ Cytotoxic T cells, and CD4+/CD25+ T regulatory cells. This trend was similarly observed in the Stage 1,

KB^{H}GRNBoost2, KB^{H} LLM, and KB^{GPT4} GRNBoost2 datasets, albeit with slightly varying percentages.

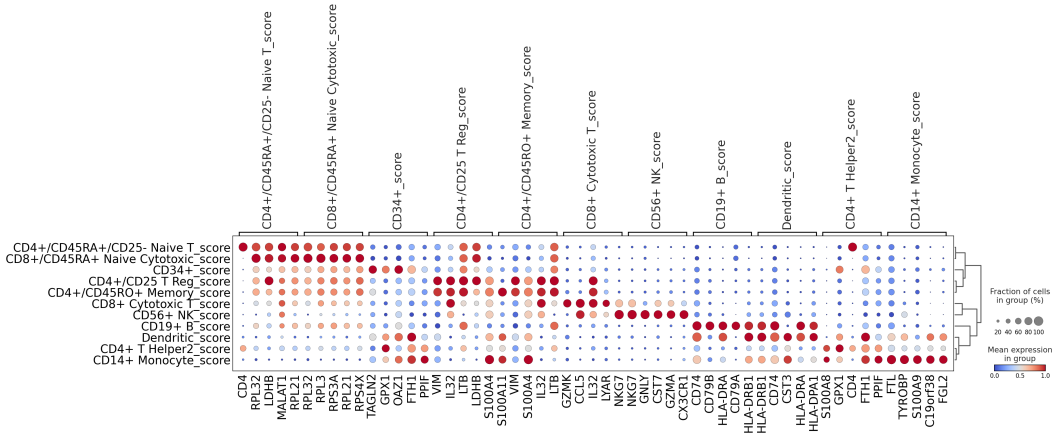
Discrepancies in Cell Type Proportions The most considerable variation in proportions was noted for CD8+/CD45RA+ Naïve Cytotoxic T cells, with KB^{Llama} GRNBoost2 reporting the highest percentage at 34.1% and GPT-4 the lowest at 28.2%. Importantly, all generated datasets displayed significant differences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

Overall Performance and Future Improvements for KB^{Llama} GRNBoost2 Therefore, it was observed that CD4+/CD45RA+/CD25– Naïve T Cells and CD8+/CD45RA+ Naïve Cytotoxic 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 KB^{Llama} 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.

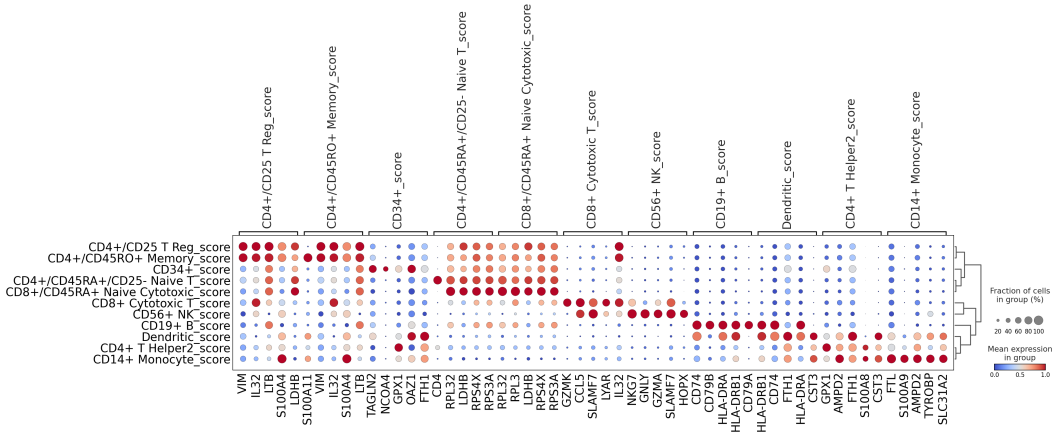
Conclusion: Challenges and Future Refinements In contrast, the KB^{GPT4} GRNBoost2 model exhibited more noise than KB^{Llama} GRNBoost2, with similar markers expressed across various cell types at comparable mean expression levels. The Stage 1 model displayed less noisy expression patterns, but some 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.



(a) Setting 2. KB^{Llama}, GRNBoost2 Top Markers

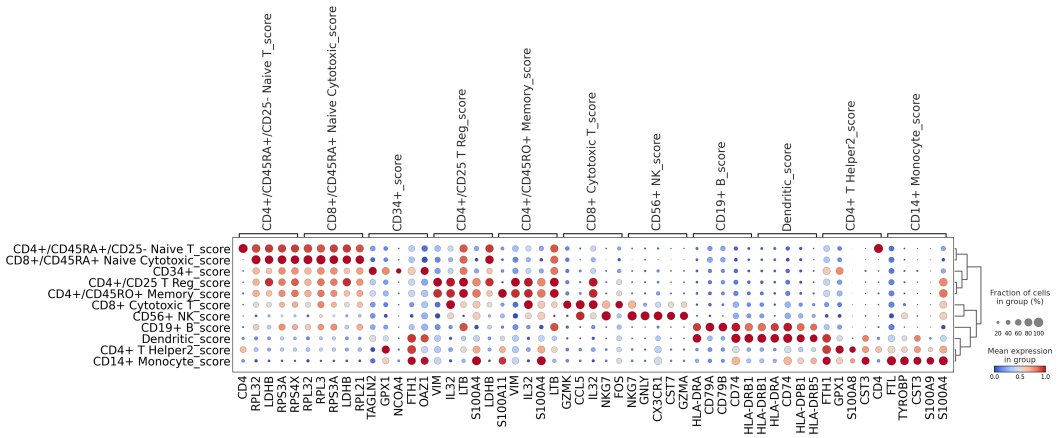


(b) Setting 2. KB^{GPT4}, GRNBoost2 Top Markers

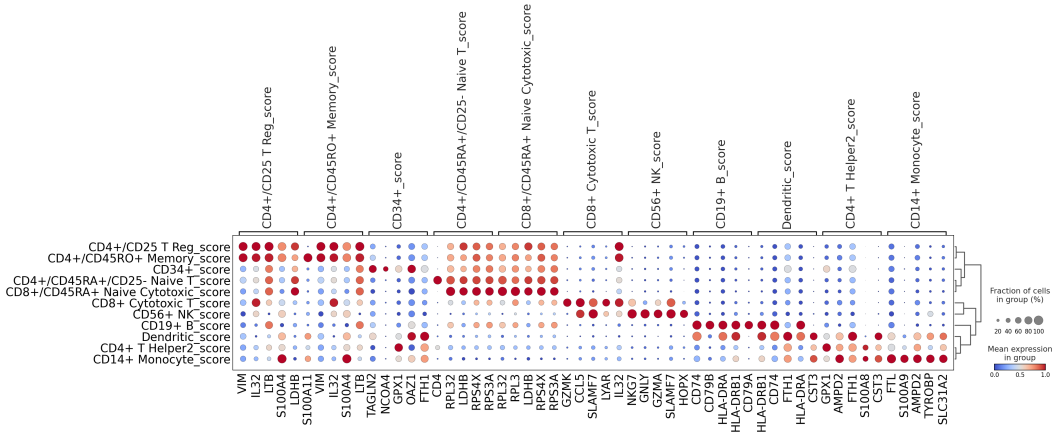


(c) Setting 1, KB^{H}, GRNBoost2 graph.

Figure 4: Dot plots illustrating the gene expression profiles of top marker genes across different cell types. The red color represent overexpression of the marker gene in the cell type while blue color represents the downregulation. Size of the bubble (dot) represents the cell percentage or fraction of the expression.

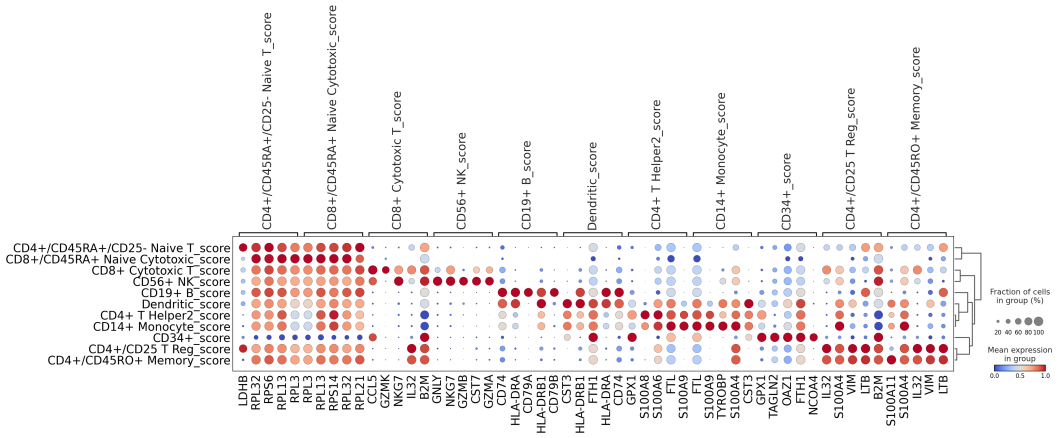


(a) Setting 1, KB^{H}, LLM graph.

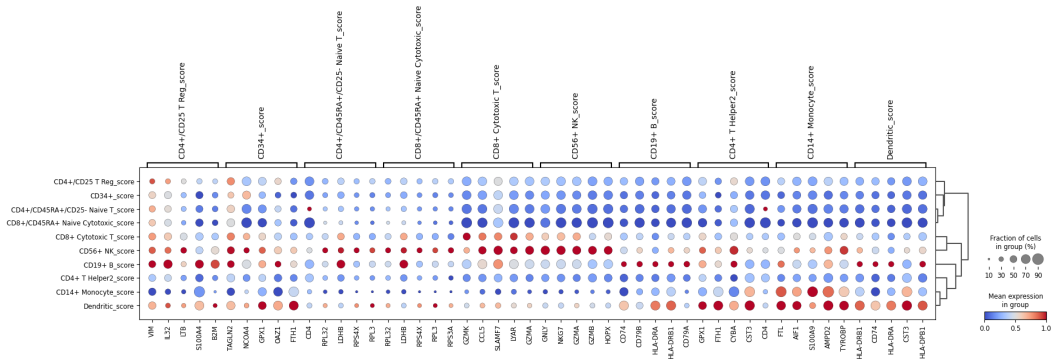


(b) Setting 1, KB^{H}, GRNBoost2 graph.

Figure 5: Dot plots visualization the gene expression profiles of KB^{H} LLM and KB^{H} GRNBoost2 datasets. (a) Expression profiles of the KB^{H} LLM model show significant noise for multiple cell types, including T helper and regulatory cells. (b) Noise patterns observed in the KB^{H} GRNBoost2 dataset, particularly in CD4+/CD45A+/CD25– Naïve T cells and CD8+/CD45RA+ Naïve cytotoxic T cells, with reduced noise in the KB^{Llama} model compared to GKB^{GPT4}.

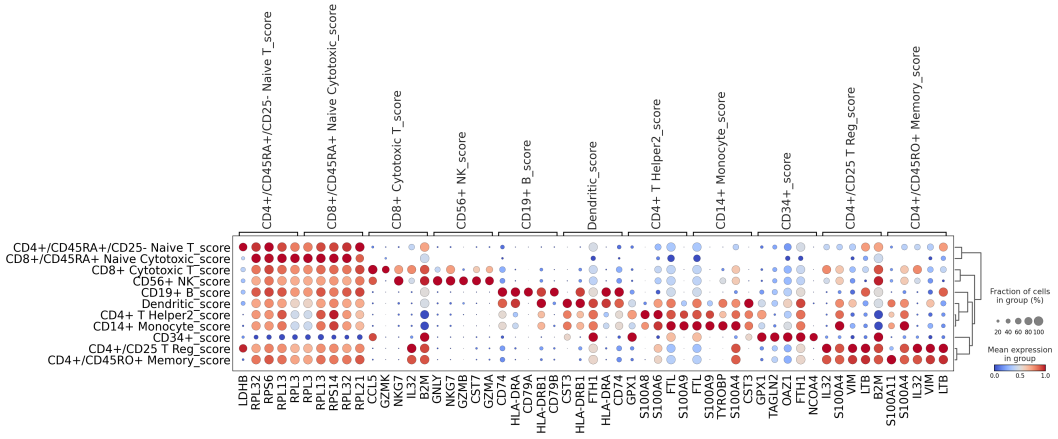


(a) Original data.

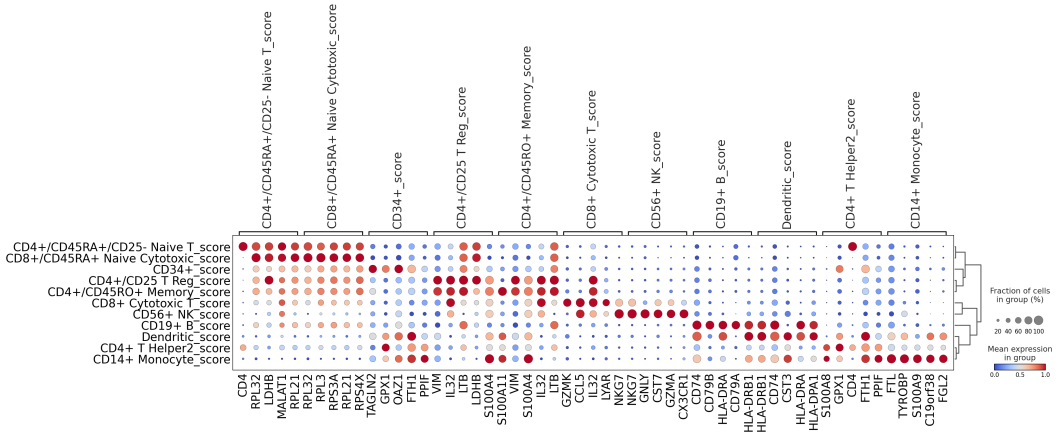


(b) Stage 1, non-causal synthetic data.

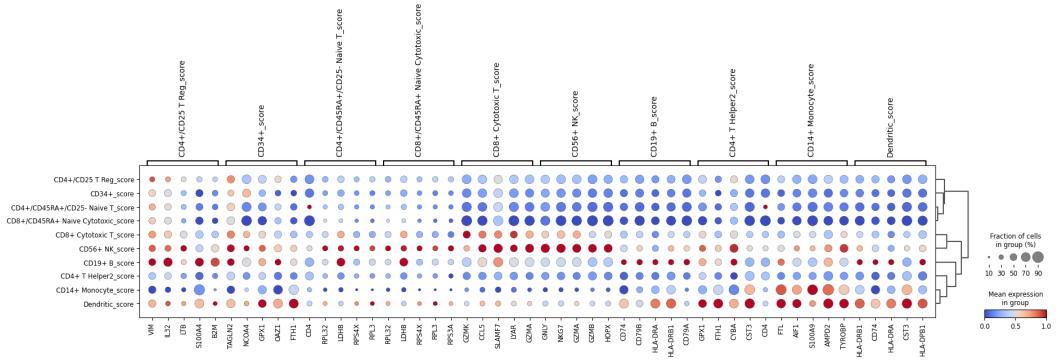
Figure 6: Dot plots illustrating gene expression profiles of key marker genes across various cell types in the Original and Stage 1 datasets. (a) In the original dataset, a dot plot of the top marker genes shows a highly variable expression profile for CD4+/CD45RA Naive T cells. While other cell types also exhibit a noisy expression profile, it is not as pronounced as that of the Naive T cells. (b) The Stage 1 non-causal synthetic dataset, when represented in a dot plot, displays an extremely noisy expression profile for all top markers across all cell types, characterized by lower mean expression but a broader range of expression across a higher cell fraction.



(a) Original data.

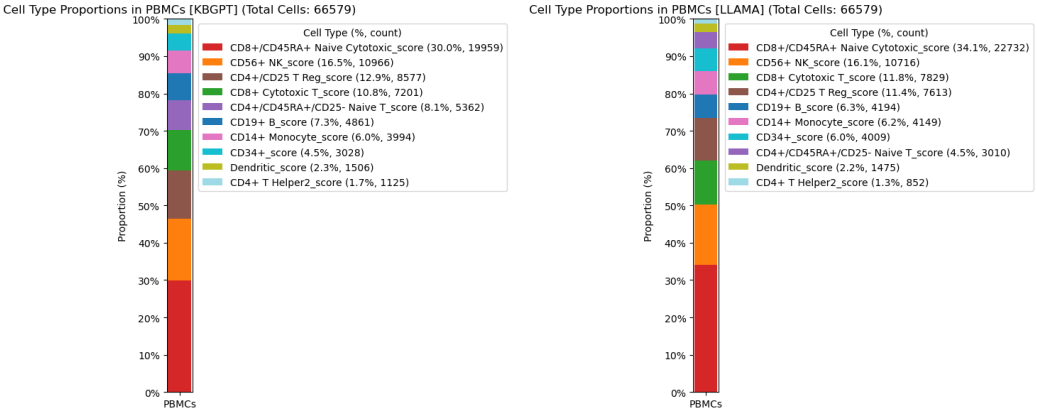


(b) Setting 2. KB$^{\text{GPT4}}$, GRNBoost2 Top Markers



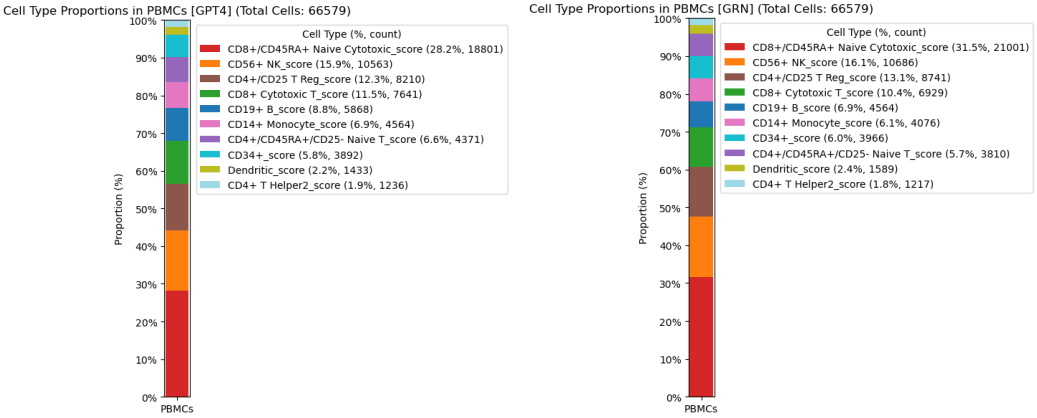
(c) Stage 1, non-causal synthetic data.

Figure 7: Dot plots illustrating gene expression profiles of key marker genes across various cell types in the Original, KB$^{\text{GPT4}}$, GRNBoost2 and Stage 1 datasets. The Stage 1 non-causal synthetic dataset displays an extremely noisy expression profile for all top markers across all cell types, characterized by lower mean expression but a broader range of expression across a higher cell fraction.



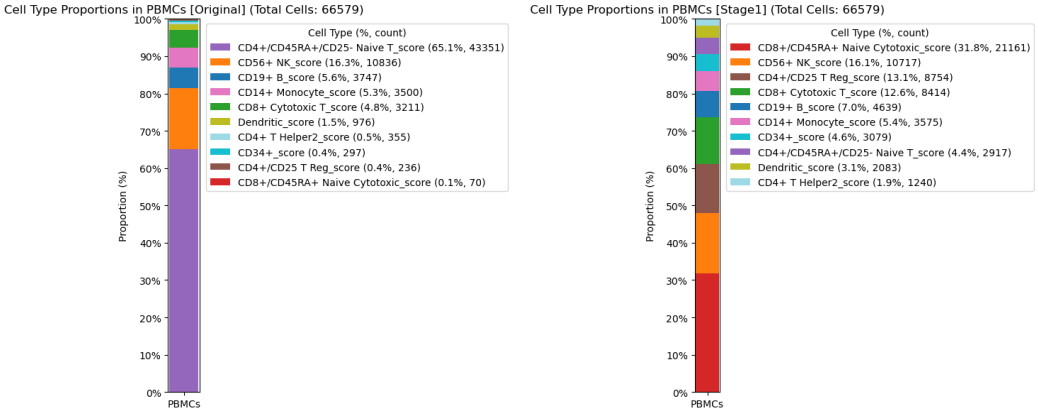
(a) Setting 2, KB^{GPT4}, GRNBoost2 graph. (b) Setting 2, KB^{Llama}, GRNBoost2 graph.

Figure 8: Cell type proportions analysis of GPT4-KB LGRNBoost2 and Llama-KB GRNBoost2 datasets reveals similar cell type proportions where the differences in percentages is between 0.1% to 4% across same cell types between the two datasets.



(a) Setting 1, KB^{H}, LLM graph (b) Setting 1, KB^{H}, GRNBoost2 graph

Figure 9: Cell type proportions analysis of KB^{H} LLM and KB^{H} GRNBoost2 datasets reveals similar cell type proportions where the differences in percentages is between 0.1% to 3% across same cell types between the two datasets.



(a) Original data.

(b) Stage 1, non-causal data.

Figure 10: Comparison of cell annotations and proportions between the Original and Stage 1 datasets reveals significant differences in the prevalence of CD4+/CD45RA+/CD25- Naive T cells. In the Original dataset, these cells constitute 65.1% of the population. In contrast, Stage 1 non-causal and synthetic datasets show a marked reduction, with Naive T cells accounting for only 4.4%.