Bridging Gene Regulatory Networks and Causal Representation Learning in Single-Cell Genomics Data

Vincenzo Lagani [1 2] **Giorgi Sokhadze** [3] **Liliia Nigmetzianova** [3] **Robert Lehmann** [1] **Yiling Ma** [1] **Sumeer Ahmad Khan** [1] **Xabier Martinez-de-Morentin** [1] **Narsis Aftab Kiani** [4 5] **Mikel Hernaez** [6 7 8] **Alexander A. Lukyanov** [3] **Jesper Tegnér** [1 9 10] **David Gomez-Cabrero** [1]

Abstract

Understanding gene regulatory mechanisms is key to advancing our capacity to interpret and manipulate cellular physiology, with significant implications for bioengineering and precision medicine. Two major computational paradigms, namely gene regulatory network (GRN) reconstruction and causal representation learning (CRL), offer distinct perspectives on transcriptional regulation. GRN methods focus on capturing detailed, fine-scale interactions among genes and transcription factors, whereas CRL seeks to identify a small set of latent variables that drive gene expression, providing a coarser but potentially more generalizable representation. In this work, we propose methods that incorporate GRN-derived structures into CRL models, guiding their training and enriching their biological interpretability. Computational experiments on scPerturb-seq datasets demonstrate that GRNs and CRL can work in concert, yielding biologically interpretable latent representations without sacrificing predictive performance.

[1]Division of Biomedical Sciences, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia [2]Institute of Chemical Biology, Ilia State University, Tbilisi, Georgia [3]Quantori, Cambridge, Massachusetts, United States of America [4]Center for Molecular Medicine, Karolinska Institute, Stockholm, Sweden [5]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden [6]CIMA, Universidad de Navarra, IdiSNA, Pamplona, Spain [7]DATAI, Universidad de Navarra, Pamplona, Spain [8]CiberONC, Madrid, Spain [9]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia [10]Science for Life Laboratory, Karolinska Institutet, Solna, Sweden. Correspondence to: Vincenzo Lagani <vincenzo.lagani@kaust.edu.sa>, David Gomez-Cabrero <david.gomezcabrero@kaust.edu.sa>.

Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026).

1. Introduction

Understanding the mechanisms that govern transcriptional regulation is a central challenge in molecular and systems biology. Accurate models of these processes are critical for applications such as dissecting cell-type–specific regulatory programs (Zhang et al., 2023b), predicting cellular responses to perturbations (Roohani et al., 2024), and designing synthetic circuits for bioengineering (De Carluccio et al., 2024).

Among the many computational approaches developed for this task, two complementary paradigms have emerged. Gene Regulatory Network (GRN) inference methods aim to reconstruct networks of regulatory interactions between transcription factors (TFs) and their target genes, typically grounded in statistical dependencies observed in gene expression data (Kim et al., 2023; Badia-I-Mompel et al., 2023). In contrast, Causal Representation Learning (CRL) seeks to identify a small set of latent causal variables that explain observed high-dimensional data (Scholkopf et al., 2021). While GRNs provide fine-grained descriptions of regulatory interactions, CRL offers a more abstract and potentially generalizable view of cellular regulation.

Despite their shared goal of modeling transcriptional regulation, these paradigms operate at different levels of abstraction and remain largely disconnected. GRN methods emphasize biological interpretability and mechanistic grounding, whereas CRL prioritizes disentanglement and predictive generalization. As a result, GRNs often lack a compact representation suitable for generative modeling, while CRL models may yield latent variables that are difficult to interpret biologically.

In this work, we introduce gCRL, a unified framework that bridges GRN inference and CRL by using GRN-derived structure to guide the learning of causal latent representations. Specifically, we leverage transcription factor communities identified in GRNs as biologically grounded priors for CRL models. This integration constrains the latent space toward interpretable regulatory programs while preserving the generative capabilities of CRL.

We demonstrate that incorporating GRN-derived priors enables CRL models to produce latent representations that are both biologically meaningful and predictive of cellular responses to perturbations. Our results suggest that mechanistic network structure and causal representation learning can be combined to obtain models that are simultaneously interpretable and effective.

2. Methods

2.1. Causal Representation Learning in Single Cell Genomics

CRL methods represent a set of measurements X through a set of causal latent factors (CLFs) Z, typically lying on a lower-dimensional manifold. Unlike standard dimensionality reduction techniques, CRL assigns a causal interpretation to Z, aiming to capture the underlying generative factors governing X (Scholkopf et al., 2021). In practice, most CRL methods are implemented as (variational) autoencoders ((V)AEs), where $Z = g(X)$ and $X = f(Z)$, with g and f denoting the encoder and decoder, respectively. Furthermore, the identifiability of Z generally relies on the availability of data collected under distributional shifts, such as external perturbations. In the language of Pearl's do-calculus (Pearl, 1995), interventions induce distributions of the form $P(X \mid do(Z_j = z_j^*))$, where the operator $do(Z_j = z_j^*)$ denotes forcing a latent factor to a specific value.

In single-cell transcriptomics (scRNA-seq), X corresponds to gene expression profiles measured across individual cells, while the latent factors Z can be interpreted as coordinated regulatory programs acting on groups of genes. Although previous work has attempted to associate Z with known biological processes (de la Fuente Cedeño et al., 2024), their interpretation remains challenging (Tejada-Lapuerta et al., 2025). Perturb-seq, which combines scRNA-seq with CRISPR-based perturbations, provides a natural setting for CRL by generating data under controlled genetic interventions (Dixit et al., 2016). In this context, the goal of CRL models is to learn parsimonious representations that capture causal regulatory mechanisms while enabling accurate prediction of unseen perturbations and perturbation combinations (Zhang et al., 2023a).

Importantly, perturb-seq interventions operate at the gene level, inducing distributions of the form $P(X \mid do(X_i = x_i^*))$. A central challenge for CRL models is therefore to reconcile gene-level interventions with latent causal variables, that is, to identify Z_j and z_j^* such that

$$P(X \mid do(Z_j = z_j^*)) \approx P(X \mid do(X_i = x_i^*)).$$

This correspondence is critical for grounding latent variables in biologically meaningful mechanisms.

2.2. Gene Regulatory Networks

A Gene Regulatory Network (GRN) is modeled as a directed graph $G(N, E)$, where nodes N represent genes and edges E denote regulatory interactions. Nodes are partitioned into transcription factors (TFs) and target genes (TGs): TFs regulate both TGs and other TFs, while TGs do not exert regulatory effects. Biologically, TFs encode proteins that bind to promoters or enhancers, modulating gene transcription. Regulation is typically combinatorial, with multiple TFs acting on the same target, and feedback loops are common.

TFs are known to organize into communities, characterized by strong intra-community connectivity and weaker inter-community interactions. This modular structure suggests that TFs within a community tend to reinforce each other's activity, while interactions between communities are more directional (Gyorgy & Vecchio, 2014).

Parametrizing G yields a quantitative GRN. Under a linear assumption, the expression of a target gene is modeled as a weighted combination of its regulators:

$$X_{TG} = \sum_{i \in RE(TG)} w_i \cdot X_i + w_0,$$

where $RE(TG)$ denotes the set of TFs regulating the target gene, w_i are regression coefficients, and w_0 represents baseline expression. Alternative formulations, such as random forest–based models (Bravo González-Blas et al., 2023) or differential equation–based approaches (Bertin et al., 2025), allow for more complex, nonlinear dynamics.

2.3. gCRL: GRN-informed CRL

The central idea of gCRL is to use structures derived from Gene Regulatory Networks (GRNs) as an inductive bias to constrain the latent space of Causal Representation Learning (CRL) models. In particular, gCRL leverages the organization of transcription factors (TFs) into regulatory modules to guide the identification of causal latent factors.

Two biological observations motivate this approach. First, TFs act as primary drivers of transcriptional regulation and explain a large fraction of transcriptomic variability, with a relatively small subset capturing most regulatory signal (Magnusson et al., 2022). Second, TFs are organized into communities characterized by strong intra-group interactions and coordinated activity (Gyorgy & Vecchio, 2014). These communities can be interpreted as low-dimensional regulatory programs, making them natural candidates for causal latent factors.

Building on these principles, gCRL maps GRN structure to latent variables through the following steps (see Appendix for full details):

1. Construct a reference GRN
2. Identify TF communities within the GRN
3. Summarize the activity of each TF community
4. Use TF community activities as priors for CRL models

2.4. Construct a reference GRN

We construct a reference GRN using CellOracle (Kamimoto et al., 2023), a method that integrates epigenomic priors with gene expression data to infer regulatory interactions. CellOracle starts from a base network derived from TF binding evidence and refines it by fitting gene-wise regression models that relate target gene expression to the expression of candidate regulators. This procedure yields a directed network with weighted TF–target interactions and associated statistical significance. In our analysis, we retain regulatory edges with FDR ≤ 0.05.

While we use CellOracle as a representative method, the proposed framework is agnostic to the specific GRN inference approach and can be applied to any method providing weighted TF–target interactions.

2.5. Identify TF communities within the GRN

We identify transcription factor (TF) communities by applying the Leiden algorithm (Traag et al., 2019) to the TF–TF subnetwork derived from the GRN. Edge weights between TF pairs are defined by aggregating the absolute values of their inferred regulatory interactions, capturing the overall strength of association between TFs.

To ensure robustness, community detection is repeated across multiple random initializations and resolution parameters, and the final partition is selected based on stability across runs. The resulting TF communities represent groups of strongly interconnected regulators, which we interpret as candidate regulatory modules.

2.6. Summarize the activity of each TF community

We summarize the activity of each TF community through a one-dimensional representation of the transcriptional profiles of its TFs, denoted as TFA (TF community activity).

In this work, we use eigengenes, defined as the first principal component of the normalized expression values of the TFs within the community (Alter et al., 2000), capturing the dominant mode of variation in TF activity.

In addition to one TFA per TF community, we include a global TFA computed across all transcription factors, capturing broad regulatory trends that may not be fully explained by individual communities.

While eigengenes provide a simple and effective summary, alternative strategies could be used to quantify TF community activity, ranging from simple aggregations (e.g., mean expression across TFs) to more expressive models (e.g., graph-based or neural representations). The proposed framework is therefore agnostic to the specific choice of TFA representation.

These community-level activity profiles provide biologically grounded, low-dimensional summaries of regulatory programs, and serve as proxies for candidate causal latent factors in subsequent CRL models.

2.7. TF community activities as priors for CRL models: gCRL-AE

Ahuja et al. (2023) showed that auto-encoder architectures can recover an affine transformation of the underlying causal factors, $\hat{Z} = AZ$, under suitable assumptions on interventions and when the decoder is a polynomial function. Recovering the true factors Z from $\hat{Z}$ then requires additional structural constraints.

We hypothesize that TF community activities (TFA) provide such structure in the context of perturb-seq data. Specifically, we treat TFA as biologically grounded proxies for latent causal factors, and use them to guide the identification of the transformation A.

To this end, we train an auto-encoder on perturbed and unperturbed transcriptomic profiles, restricting the input to TF expression (X_{TF}) and setting the latent dimensionality to match the number of TF communities. This yields a latent representation $\hat{Z}$, which corresponds to a mixed version of the underlying factors.

We then estimate a linear transformation A that aligns $\hat{Z}$ with TFA by maximizing a modified version of the Test-based Measurement EXclusivity (T-MEX) score (Yao et al., 2025):

$$A^* = \arg\max_A \; T\text{-}MEX(\mathrm{TFA}, \hat{Z}A).$$

Statistical significance of the alignment is assessed via permutation analysis by shuffling TF assignments across communities. We refer to this procedure as gCRL-AE.

2.8. TF community activities as priors for CRL models: gCRL-VAE

The discrepancy-VAE (Zhang et al., 2023a) is a generative CRL model designed for perturb-seq data, capable of (i) inferring latent variables Z, (ii) estimating a directed acyclic graph (DAG) over Z, and (iii) predicting responses to unseen combinations of interventions, when the individual interventions are observed during training.

This is achieved by extending a standard VAE architecture with (a) an additional interventional encoder, which estimates, for each intervention, its target latent factor and effect, and (b) a deep structural causal model (Pawlowski et al., 2020) that infers causal relationships among the components of Z. A key challenge in discrepancy-VAE is the selection and interpretation of the latent variables Z. We address this by introducing gCRL-VAE, which incorporates TF community activities as priors to guide the learning of Z. Our implementation modifies discrepancy-VAE as follows:

1. The encoder input is restricted to TF expression values, and the dimensionality of Z is set to the number of TF communities.
2. A regularization term based on $T\text{-}MEX(TFA, Z[\pi])$ is added to the loss function, encouraging alignment between latent variables and TF community activities. Here, π is a learnable permutation that aligns the two representations.
3. The mapping between genetic interventions and latent variables is derived from π, providing a GRN-informed assignment that complements the effect estimates of the interventional encoder.

In this formulation, GRN-derived structure directly constrains the latent space, promoting representations that are both predictive and biologically interpretable.

2.9. Experimentation protocol

2.9.1. Perturb-seq data preprocessing

Open Reading Frame (ORF) over-expression of transcription factors in human embryonic stem cells. Joung et al. (2023) generated a large-scale atlas comprising over 1.1 million single-cell transcriptomic profiles measured in human embryonic stem cells (hESCs), perturbing 3,266 TFs via ORF over-expression. We randomly subsampled 10,000 unperturbed cells and up to 1,000 cells per perturbed TF. Perturbations with fewer than 500 cells were excluded. The resulting dataset (348,766 cells, 37,528 transcripts, 382 perturbed conditions) was normalized using the standard Scanpy pipeline (Wolf et al., 2018), followed by the selection of 5,000 highly variable genes.

CRISPR activation (CRISPRa) combinatorial experiments in human myelogenous leukemia cells. This dataset was designed to explore how genetic interactions shape high-dimensional transcriptomic landscapes (Norman et al., 2019). We used the preprocessed version provided by the discrepancy-VAE publication, comprising normalized transcriptomic profiles for 8,907 unperturbed cells, 57,831 single-perturbation cells, and 41,759 double-perturbation cells, measured over 5,000 highly variable genes.

2.10. Evaluation of gCRL-VAE predictive performance

We evaluated the predictive performance of gCRL-VAE using the double-perturbation data from Norman et al. (2019). We compared gCRL-VAE against the original discrepancy-VAE (d-VAE) and an informed variant (id-VAE), in which the dimensionality of Z is set to the number of TF communities rather than to the number of training perturbations. In addition, we considered two baselines: (i) *worst-case prediction*, where all cells are predicted as unperturbed, and (ii) *perfect prediction*, where half of the test cells are randomly selected and used as predictions (Mejia et al., 2025), providing an estimate of irreducible biological variability.

For all models (gCRL-VAE, d-VAE, id-VAE), we replaced the Maximum Mean Discrepancy (MMD) term in the loss function with a mean squared error (MSE) computed between the centroids of predicted and observed cell populations. While MMD can capture fine-grained distributional differences, its estimation from small batches can be unstable; we therefore adopt a simpler and more robust alternative. Details on model training are provided in Appendix A.6.

3. Results

3.1. TF communities are detected by interventional CRL methods

We derived a GRN from the unperturbed cells of the Joung et al. (2023) dataset, following the procedure detailed in Appendix A.1. The network coefficients exhibited a bimodal distribution, motivating the retention of the top 50% of edges (Figure S1). The resulting network comprised 25,694 edges connecting 150 transcription factors (TFs) and 1,646 target genes. The TFs partitioned into six communities, displaying significantly higher internal cohesion than expected by chance (Figure S1c).

We performed gene over-representation analysis (ORA) to identify Gene Ontology (GO) (The Gene Ontology Consortium, 2015) Biological Processes (BPs) enriched among the TFs in each community and their respective targets.

We identified 57 background BPs significantly enriched (FDR < 0.05) in at least half of the communities, reflecting overarching biological programs characteristic of the dataset. These include neuronal differentiation and synaptic organization, cell–cell junction assembly, extracellular matrix organization, calcium signaling, and developmental processes such as embryonic morphogenesis and angiogenesis. This pattern is consistent with the broad differentiation potential of hESCs, as reported by Joung et al. (2023), where cells exhibit coordinated activation of several differentiation programs alongside repression of pluripotency.

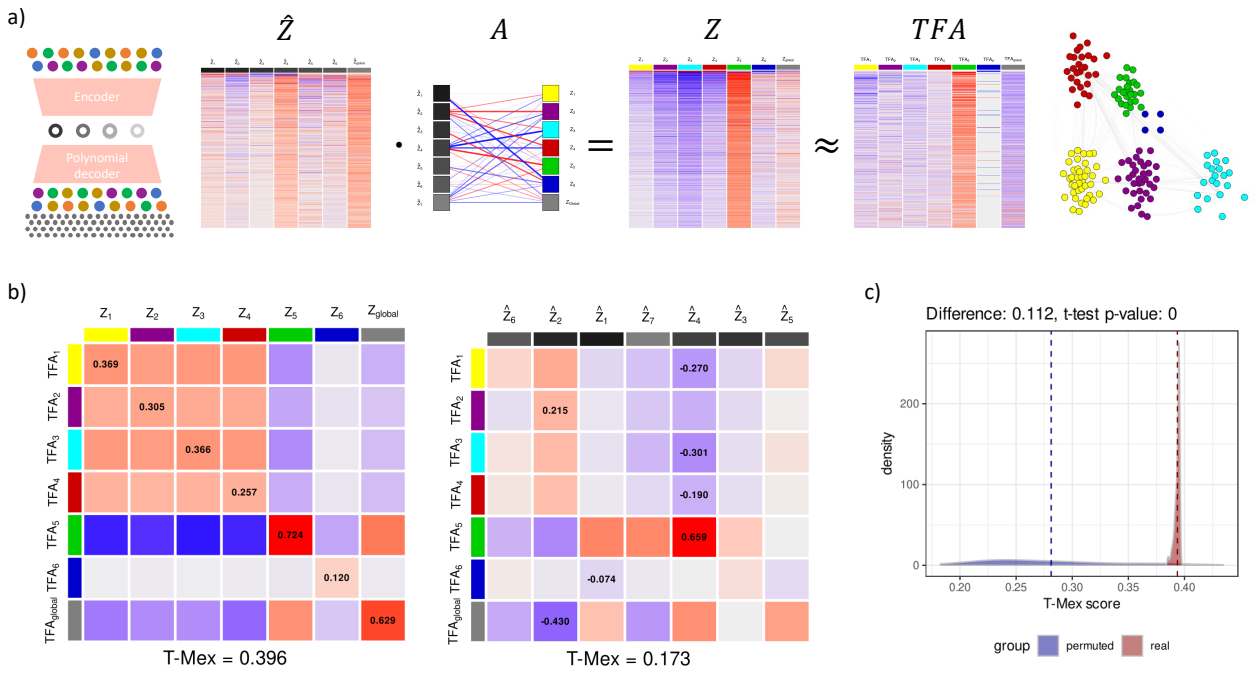

Figure 1. gCRL-AE operation and results on data from Joung et al. (2023). Panel a: The entangled embeddings $\hat{Z}$ are obtained from an autoencoder with a polynomial decoder. A rotation matrix A is then used to align these embeddings with the TF community activities TFA, yielding the disentangled representation Z. Panel b: Correlation matrices between TFA and Z (left), and between TFA and $\hat{Z}$ (right). The disentangled embeddings Z show substantially higher alignment with TF community activities, as further supported by an increased T-MEX score. Panel c: Comparison of technical variability, measured by deriving A across different random seeds (red), with the variability obtained by permuting TFs across communities (blue). Technical variability is an order of magnitude lower, and the T-MEX value corresponding to the original community partition is significantly higher than expected by chance.

Beyond these shared processes, individual TF communities display distinct functional enrichments. For instance, selected communities are enriched for calcium signaling and synaptic transmission, suggesting neuronal-like programs, while others are associated with endothelial proliferation, angiogenesis, and Wnt signaling, indicative of mesodermal and vascular differentiation pathways. Additional communities capture membrane organization, cytokine response, and protein trafficking processes, pointing to signaling and cellular remodeling. Further details on the ORA results are reported in Appendix B.1.

We next applied gCRL-AE to the 1,688 genes contained in the GRN, using both unperturbed cells and single-perturbation conditions targeting genes present in the six TF communities (45,072 cells across 39 interventions).

Figure 1a illustrates the latent representation $\hat{Z}$ learned by the auto-encoder, the rotation matrix A obtained by optimizing the T-MEX criterion, and the resulting aligned representation Z. Panel b reports the correlation matrices between TFA and Z (left) and between TFA and $\hat{Z}$ (right), with the highest correlation per row highlighted. Each Z component is maximally correlated with a single TFA component, and vice versa, indicating a high degree of alignment. In contrast, each $\hat{Z}$ component is associated with multiple TFA components, reflecting a lack of disentanglement prior to alignment. Panel c summarizes the permutation analysis. First, we optimized A 100 times using the original TF communities, varying the random seed (red values). We then permuted TF assignments across communities 100 times and recomputed the optimal A (three runs per permutation, averaged; blue distribution). The variability due to optimization randomness (standard deviation 0.0022) is an order of magnitude smaller than that induced by community permutation (standard deviation 0.0615). More importantly, the average T-MEX score for the true communities (0.3937) is significantly higher than that obtained for permuted communities (0.2812; one-sample, one-tailed t-test $p < 10^{-33}$).

These results suggest that the auto-encoder captures, directly from transcriptomic data, structural patterns that are consistent with those derived from GRN reconstruction and TF community organization.

We performed the same analysis on the Norman et al. (2019) dataset, obtaining qualitatively similar results. The inferred GRN comprised 1,642 genes, including 141 TFs, partitioned into eight communities with higher-than-random internal cohesion (Figure S2). The average T-MEX score for the true communities (0.4771) was significantly higher than for permuted communities (0.4592; $p < 10^{-4}$, see Figure S3), although the gap was smaller than in the Joung dataset. This suggests that the extent to which transcriptomic data alone recapitulates GRN-derived structure depends on the underlying dataset.

3.2. GRN-guided CRL models provide interpretability with competitive predictive performance

Across the eight TF communities inferred from the Norman et al. (2019) GRN, the ORA analysis identified forty-four background biological processes. These include immune and hematopoietic differentiation, TGF-β and ERK/MAPK signaling, apoptosis regulation, angiogenesis, phagocytosis, and cellular stress responses.

These shared processes reflect both the intrinsic regulatory landscape of the K562 leukemia cell line and the experimental context of perturb-seq. K562 cells retain multilineage hematopoietic potential and exhibit active signaling and stress-response pathways. In particular, the prevalence of immune and differentiation-related programs is consistent with their hematopoietic origin, while the enrichment of signaling and apoptosis pathways reflects their transformed, proliferative state. At the same time, recent evidence suggests that perturb-seq experimental procedures can induce stress responses even in nominally unperturbed cells (Viñas Torné et al., 2025), which may contribute to the widespread presence of stress-related programs.

Beyond these global programs, individual TF communities show distinct enrichments corresponding to specialized regulatory modules (Appendix B.2). Some communities are associated with immune activation and cytokine signaling, others with endothelial and vascular processes, and others with erythroid or muscle differentiation. Additional communities specialize further in cell cycle regulation, DNA repair, and stress-response pathways.

We next assessed whether this interpretability comes at the cost of predictive performance. Among the double-perturbation conditions in the Norman et al. (2019) dataset, 25 involved transcription factors present in the GRN reconstructed from the corresponding unperturbed cells. These 25 conditions were assigned to the hold-out test set, totaling 7,330 cells, while 27,718 cells (single-perturbation conditions represented in the GRN along with unperturbed cells) were used to train the gCRL-VAE, id-VAE, and d-VAE models.

Table 1 reports the Average Euclidean Centroid Distance (AECD) between predicted and observed perturbation centroids for held-out double perturbations. gCRL-VAE achieved the lowest AECD among the learned models, outperforming both id-VAE and d-VAE in nominal terms, although the differences were not statistically significant. Importantly, all learned models performed substantially better than the worst-case baseline and remained within the range bounded by the perfect-prediction reference.

These results indicate that incorporating GRN-derived priors improves biological interpretability while preserving predictive performance comparable to discrepancy-VAE models.

Table 1. Average Euclidean Centroid Distance (AECD) between predicted and observed centroids in gene-expression space for held-out double perturbations from the Norman et al. (2019) dataset. Lower values indicate better performance. p-values are paired, two-tailed Wilcoxon signed-rank tests against gCRL-VAE. S.E.: standard error.

MODEL	AECD	S.E.	p-VALUE
PERFECT	1.962	0.107	< 0.01
GCRL-VAE	3.639	0.442	—
ID-VAE	3.945	0.703	0.220
D-VAE	4.324	0.756	0.080
WORST-CASE	5.504	0.322	0.002

4. Discussion

We presented gCRL, a framework that bridges Gene Regulatory Networks (GRNs) and Causal Representation Learning (CRL) by using GRN-derived structure as inductive priors to guide the inference of latent variables. This approach enables the integration of mechanistic regulatory information into data-driven representation learning.

Our results support three main conclusions.

First, TF community structure appears to capture biologically meaningful axes of variation that can be partially recovered from transcriptomic data alone. In gCRL-AE, the polynomial auto-encoder is trained without access to the TFA priors; nevertheless, its embeddings show higher alignment with TFA derived from the original TF communities than with TFA derived from permuted communities. This suggests that the modular structure inferred from the GRN is already reflected, at least in part, in the representation learned by the auto-encoder, although the strength of this signal varies across datasets and depends on GRN reconstruction and community-definition choices.

Second, TF communities inferred from GRNs correspond to coherent regulatory programs, supporting the interpretability of the learned CRL representations.

Third, despite the constrained latent space induced by GRN priors, gCRL-VAE achieves predictive performance comparable to discrepancy-VAE models, indicating that interpretability can be improved without sacrificing predictive accuracy.

Our approach has several limitations. First, neither the latent variables Z nor the TF community activities TFA correspond to ground-truth causal factors. Both are estimated from data and subject to modeling assumptions. Aligning the former to the latter enforces consistency between two approximate representations, rather than establishing that one provides a more faithful description of the underlying biology. In this sense, gCRL should be interpreted as creating a productive tension between two complementary views of the regulatory system: the CRL model is driven by perturbation-induced transcriptomic variation, while the GRN prior pulls the latent space toward regulatory programs inferred from network structure. In gCRL-VAE, this influence is further moderated by the fact that TFA contributes only through one component of the overall loss function, rather than fully determining the latent representation.

More broadly, identifiability guarantees for CRL frameworks hold under idealized assumptions, including sufficiently informative interventions, smooth invertible mappings, and asymptotic data regimes. In realistic biological systems, interventions may affect multiple latent programs simultaneously, regulatory interactions may contain feedback loops, and transcriptomic measurements are noisy and partially observed. Consequently, the learned latent variables should be interpreted as approximate causal abstractions rather than exact recovery of underlying biological mechanisms.

At the same time, the TF community activities used as priors are themselves imperfect estimates derived through a multi-step modeling procedure. Each step of gCRL involves implementation choices: the reference GRN can be reconstructed using different methods, TF communities can be identified using different graph partitioning strategies, community activity can be summarized using eigengenes or alternative representations, and the resulting activities can be incorporated into CRL models through different regularization schemes.

In this work, we instantiate these steps using CellOracle, Leiden clustering, eigengenes, and T-MEX-based alignment, respectively. Other choices, such as graph neural network–based community representations, alternative alignment losses, or different GRN inference methods, could lead to different inductive biases. Therefore, the framework is methodologically agnostic in design, but its robustness across implementation choices remains an important direction for future work.

Finally, the T-MEX score itself relies on conditional association estimation, which can become statistically challenging in high-dimensional settings and may affect the robustness of the alignment procedure.

Despite these limitations, our results indicate that GRN-derived regulatory structure can serve as an effective inductive bias for causal representation learning, enabling latent representations that remain both biologically interpretable and predictively informative.

4.1. Related Work

Gene Regulatory Network inference. A wide range of computational approaches have been proposed for GRN reconstruction. Early methods, such as ARACNe (Margolin et al., 2006), rely on information-theoretic measures to infer regulatory interactions from statistical dependencies in gene expression data. Regression-based approaches, including GENIE3 (Huynh-Thu et al., 2010), frame GRN inference as a predictive task, leveraging ensemble methods to capture nonlinear relationships between transcription factors and target genes. Dynamic models, such as SCODE (Matsumoto et al., 2017), explicitly account for temporal structure in the data by modeling gene regulation through differential equations.

More recent approaches leverage graph neural networks, transformers, perturbation-aware architectures, and foundation models to infer nonlinear regulatory structure directly from transcriptomic and perturbation data (Shu et al., 2021; Roohani et al., 2024; Cui et al., 2024; Qiu et al., 2025).

Within this landscape, a subset of methods combines chromatin accessibility, motif enrichment, and gene expression data to derive mechanistically grounded regulatory networks (Badia-I-Mompel et al., 2023). In this work, we adopt CellOracle (Kamimoto et al., 2023) as a representative of this class of multi-omics, biologically informed GRN inference methods.

Causal abstraction learning (CAL). These methods aim to derive coarser representations of causal systems by grouping sets of variables and appropriately redefining the relationships between them (Massidda et al., 2023; Geiger et al., 2025). In principle, applying CAL to GRNs would enable the identification of higher-level regulatory units, such as TF communities, together with their interactions. However, existing CAL approaches typically assume acyclic causal graphs and are not readily applicable to the large, cyclic, and densely connected structures that characterize gene regulatory networks. This limits their direct use in transcriptomic settings.

CRL modeling in single cell genomics. Several recent works have explored the application of causal representation learning to single cell data.

scConCRL (Dong et al., 2025) leverages CRL frameworks to jointly model confounding variables and regulatory interactions in single-cell data. The SENA discrepancy-VAE method (de la Fuente Cedeño et al., 2024) integrates known biological pathways into the training of the discrepancy-VAE through a masked encoder. GRACE-VAE (Zhang et al., 2025) integrates a GNN-based encoder within a discrepancy-VAE framework, using network information to construct structure-aware embeddings of gene expression prior to causal inference. In this setting, latent variables are inferred from graph-enriched representations rather than directly from gene expression, effectively summarizing transformed features that mix signals across neighboring genes.

While these approaches demonstrate the potential of combining representation learning, graph structure, and causality, they primarily incorporate network information at the level of input features. In contrast, our work takes a complementary perspective, using GRN-derived structure as an inductive bias to directly constrain the latent space, enabling the identification of biologically grounded and interpretable causal factors.

Software and Data

Code and data for replicating the presented results are available at https://github.com/quantori/gCRL

Acknowledgements

This work was supported by the CausalEmbed project, funded by the SDAIA–KAUST Center of Excellence in Data Science and Artificial Intelligence, and by the baseline funds of King Abdullah University of Science and Technology. We thank Zaal Beriashvili-Lianov and Yuriy Gankin for their role in coordinating and supporting the project, and Prof. Takis Benos for valuable discussions and feedback.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. Interventional Causal Representation Learning. In *Proceedings of Machine Learning Research*, volume 202, 2023.

Alter, O., Brown, P. O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, August 2000. doi: 10.1073/pnas.97.18.10101. URL https://www.pnas.org/doi/10.1073/pnas.97.18.10101.

Badia-I-Mompel, P., Wessels, L., Müller-Dott, S., Trimbour, R., Ramirez Flores, R. O., Argelaguet, R., and Saez-Rodriguez, J. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews. Genetics*, 24(11):739–754, November 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00618-5.

Bertin, P., Viviano, J. D., Tejada-Lapuerta, A., Wang, W., Bauer, S., Theis, F. J., and Bengio, Y. A scalable gene network model of regulatory dynamics in single cells, March 2025. URL http://arxiv.org/abs/2503.20027. arXiv:2503.20027 [q-bio].

Bravo González-Blas, C., De Winter, S., Hulselmans, G., Hecker, N., Matetovici, I., Christiaens, V., Poovathingal, S., Wouters, J., Aibar, S., and Aerts, S. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods*, 20(9):1355–1367, September 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01938-4. URL https://www.nature.com/articles/s41592-023-01938-4. Publisher: Nature Publishing Group.

Bruse, N. and Heeringen, S. J. v. GimmeMotifs: an analysis framework for transcription factor motif analysis, November 2018. URL https://www.biorxiv.org/content/10.1101/474403v1. Pages: 474403 Section: New Results.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, August 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

De Carluccio, G., Fusco, V., and di Bernardo, D. Engineering a synthetic gene circuit for high-performance inducible expression in mammalian systems. *Nature Communications*, 15(1):3311, April 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-47592-y. URL https://www.nature.com/articles/s41467-024-47592-y. Publisher: Nature Publishing Group.

de la Fuente Cedeño, J., Lehmann, R., Ruiz-Arenas, C., Voges, J., Marín-Goñi, I., Martínez de Morentin, X., Gomez-Cabrero, D., Ochoa, I., Tegnér, J., Lagani, V., and

Hernaez, M. Interpretable Causal Representation Learning for Biological Data in the Pathway Space. October 2024. URL https://openreview.net/forum?id=3Fgylj4uqL.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., and Regev, A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, December 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2016.11.038. URL https://www.sciencedirect.com/science/article/pii/S0092867416316105.

Dong, J., Li, J., and Wang, F. Conditional Causal Representation Learning for Heterogeneous Single-cell RNA Data Integration and Prediction. volume 9, pp. 7392–7400, September 2025. doi: 10.24963/ijcai.2025/822. URL https://www.ijcai.org/proceedings/2025/822.

Geiger, A., Ibeling, D., Zur, A., Chaudhary, M., Chauhan, S., Huang, J., Arora, A., Wu, Z., Goodman, N., Potts, C., and Icard, T. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. ISSN 1533-7928. URL http://jmlr.org/papers/v26/23-0058.html.

Gyorgy, A. and Vecchio, D. D. Modular Composition of Gene Transcription Networks. *PLOS Computational Biology*, 10(3):e1003486, March 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003486. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003486. Publisher: Public Library of Science.

Hernaez, M., Blatti, C., and Gevaert, O. Comparison of single and module-based methods for modeling gene regulatory networks. *Bioinformatics*, 36(2):558–567, January 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz549. URL https://doi.org/10.1093/bioinformatics/btz549.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985. ISSN 1432-1343. doi: 10.1007/BF01908075. URL https://doi.org/10.1007/BF01908075.

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE*, 5(9):e12776, September 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0012776. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012776.

Joung, J., Ma, S., Tay, T., Geiger-Schuller, K. R., Kirchgatterer, P. C., Verdine, V. K., Guo, B., Arias-Garcia, M. A., Allen, W. E., Singh, A., Kuksenko, O., Abudayyeh, O. O., Gootenberg, J. S., Fu, Z., Macrae, R. K., Buenrostro, J. D., Regev, A., and Zhang, F. A transcription factor atlas of directed differentiation. *Cell*, 186(1):209–229.e26, January 2023. ISSN 0092-8674. doi: 10.1016/j.cell.2022.11.026. URL https://www.sciencedirect.com/science/article/pii/S0092867422014702.

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, February 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05688-9. URL https://www.nature.com/articles/s41586-022-05688-9. Publisher: Nature Publishing Group.

Kim, D., Tran, A., Kim, H. J., Lin, Y., Yang, J. Y. H., and Yang, P. Gene regulatory network reconstruction: harnessing the power of single-cell multi-omic data. *npj Systems Biology and Applications*, 9(1):51, October 2023. ISSN 2056-7189. doi: 10.1038/s41540-023-00312-6. URL https://www.nature.com/articles/s41540-023-00312-6.

Magnusson, R., Tegnér, J., and Gustafsson, M. Deep neural network prediction of genome-wide transcriptome signatures - beyond the Black-box. *NPJ systems biology and applications*, 8(1), February 2022. ISSN 2056-7189. doi: 10.1038/s41540-022-00218-9. URL https://pubmed.ncbi.nlm.nih.gov/35197482/. Publisher: NPJ Syst Biol Appl.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7, March 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-S1-S7. URL https://doi.org/10.1186/1471-2105-7-S1-S7.

Massidda, R., Geiger, A., Icard, T., and Bacciu, D. Causal Abstraction with Soft Interventions. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pp. 68–87. PMLR, August 2023. URL https://proceedings.mlr.press/v213/massidda23a.html.

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., Hayashi, T., and Nikaido, I. SCODE: an efficient regulatory network inference algorithm from

single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15):2314–2321, August 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx194. URL https://doi.org/10.1093/bioinformatics/btx194.

Meilă, M. Comparing Clusterings by the Variation of Information. In Schölkopf, B. and Warmuth, M. K. (eds.), *Learning Theory and Kernel Machines*, pp. 173–187, Berlin, Heidelberg, 2003. Springer. ISBN 978-3-540-45167-9. doi: 10.1007/978-3-540-45167-9_14.

Mejia, G. M., Miller, H. E., Leblanc, F. J. A., Wang, B., Swain, B., and Camillo, L. P. d. L. Diversity by Design: Addressing Mode Collapse Improves scRNA-seq Perturbation Modeling on Well-Calibrated Metrics, June 2025. URL http://arxiv.org/abs/2506.22641. arXiv:2506.22641 [q-bio].

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, August 2019. doi: 10.1126/science.aax4438. URL https://www.science.org/doi/10.1126/science.aax4438.

Pawlowski, N., Castro, D. C., and Glocker, B. Deep Structural Causal Models for Tractable Counterfactual Inference, June 2020. URL https://arxiv.org/abs/2006.06485v2.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, December 1995. ISSN 0006-3444. doi: 10.1093/biomet/82.4.669. URL https://doi.org/10.1093/biomet/82.4.669.

Qiu, M., Hu, X., Zhan, F., Yun, S., Peng, J., Zhang, R., Kailkhura, B., Yang, J., and Chen, T. GRNFormer: A Biologically-Guided Framework for Integrating Gene Regulatory Networks into RNA Foundation Models, March 2025. URL http://arxiv.org/abs/2503.01682. arXiv:2503.01682 [cs].

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, June 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01905-6. URL https://www.nature.com/articles/s41587-023-01905-6. Publisher: Nature Publishing Group.

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Towards Causal Representation Learning. *Proceedings of the IEEE*, 109 (5):612–634, February 2021. ISSN 15582256. doi: 10.48550/arxiv.2102.11107. URL https://arxiv.org/abs/2102.11107v1. Publisher: Institute of Electrical and Electronics Engineers Inc.

Shu, H., Zhou, J., Lian, Q., Li, H., Zhao, D., Zeng, J., and Ma, J. Modeling gene regulatory networks using neural network architectures. *Nature Computational Science 2021 1:7*, 1(7):491–501, July 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00099-8. URL https://www.nature.com/articles/s43588-021-00099-8.

Tejada-Lapuerta, A., Bertin, P., Bauer, S., Aliee, H., Bengio, Y., and Theis, F. J. Causal machine learning for single-cell genomics. *Nature Genetics*, 57(4):797–808, April 2025. ISSN 1546-1718. doi: 10.1038/s41588-025-02124-2. URL https://www.nature.com/articles/s41588-025-02124-2. Publisher: Nature Publishing Group.

The Gene Ontology Consortium, . Gene Ontology Consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gku1179. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1179.

Traag, V. A., Waltman, L., and van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41695-z. URL https://www.nature.com/articles/s41598-019-41695-z. Publisher: Nature Publishing Group.

Viñas Torné, R., Wiatrak, M., Piran, Z., Fan, S., Jiang, L., Teichmann, S. A., Nitzan, M., and Brbić, M. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nature Biotechnology*, pp. 1–10, August 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02777-8. URL https://www.nature.com/articles/s41587-025-02777-8.

Wolf, F. A., Angerer, P., and Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0. URL https://doi.org/10.1186/s13059-017-1382-0.

Yao, D., Huang, S., Cadei, R., Zhang, K., and Locatello, F. The Third Pillar of Causal Analysis? A Measurement Perspective on Causal Representations, May 2025. URL http://arxiv.org/abs/2505.17708. arXiv:2505.17708 [cs].

Zhang, J., Greenewald, K., Squires, C., Srivastava, A., Shanmugam, K., and Uhler, C. Identifiability Guarantees for Causal Disentanglement from Soft Interventions. In *37th Conference on Neural Information Pro-*

cessing Systems (NeurIPS 2023), 2023a. URL https://openreview.net/pdf?id=o16sYKHk3S.

Zhang, J., Li, M. M., and Zheleva, E. Causal representation learning from network data, September 2025. URL https://arxiv.org/abs/2509.01916v1.

Zhang, S., Pyne, S., Pietrzak, S., Halberg, S., McCalla, S. G., Siahpirani, A. F., Sridharan, R., and Roy, S. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature Communications*, 14(1):3064, May 2023b. ISSN 2041-1723. doi: 10.1038/s41467-023-38637-9. URL https://www.nature.com/articles/s41467-023-38637-9. Publisher: Nature Publishing Group.

A. Supplementary Methods

A.1. Construction of the reference GRN (extended description)

We construct the reference Gene Regulatory Network (GRN) using CellOracle (Kamimoto et al., 2023), a multi-omics framework that combines epigenomic information with transcriptomic data to infer directed regulatory interactions between transcription factors (TFs) and target genes.

CellOracle operates in two main stages. First, it defines a base regulatory network by integrating prior knowledge of TF binding. In the standard workflow, this involves combining motif enrichment analysis with chromatin accessibility data (e.g., scATAC-seq) to identify candidate TF binding sites across the genome. Alternatively, pre-computed regulatory networks can be used as input, enabling flexibility depending on data availability.

Second, CellOracle refines this base network using gene expression data. For each target gene, a regression model is fitted in which the expression of the gene is predicted from the expression levels of its candidate regulators. This step prunes spurious regulatory interactions and assigns a quantitative weight $w_{a,b}$ to each TF–target pair (a, b), along with a statistical significance measure derived from the model.

In our implementation, we use a base GRN constructed on the hg38 genome using gimmemotifs v5 (Bruse & Heeringen, 2018). All other parameters are set to their default values as provided by the CellOracle framework. After model fitting, regulatory edges are filtered by controlling the false discovery rate (FDR), and only edges with FDR ≤ 0.05 are retained.

The resulting GRN is a directed, weighted graph capturing putative regulatory relationships between TFs and target genes. Importantly, while CellOracle is used here for concreteness, the downstream steps of the gCRL framework only require a weighted TF–target interaction network, and are therefore compatible with alternative GRN inference methods.

A.2. Identification of TF communities (extended description)

To identify transcription factor (TF) communities, we first derive a TF–TF interaction network from the full GRN by restricting the graph $G(N, E)$ to TF nodes. For each pair of TFs (a, b), we define the weight of their connection by aggregating the absolute values of their bidirectional regulatory interactions:

$$w_{a,b} = \sum_{i \in \{a \to b,\, b \to a\}} |w_i|,$$

where w_i denotes the regression coefficient associated with the corresponding regulatory edge. This construction captures the overall interaction strength between TFs, irrespective of direction or sign.

Community detection is performed using the Leiden algorithm (Traag et al., 2019), which optimizes modularity under a resolution parameter γ. To account for sensitivity to initialization and parameter choice, we perform a stability-based selection procedure. Specifically, we vary γ in the range $[0.9, 2.0]$ with increments of 0.1, and for each value we run the algorithm with 1000 random initializations.

For each γ, stability is assessed by computing pairwise similarity between partitions using the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) and the Variation of Information (VI) (Meilă, 2003). We evaluate a stability score based on the ARI/VI ratio over 200 randomly sampled pairs of partitions. The optimal resolution parameter is selected as the one maximizing this stability criterion, prioritizing solutions with a lower number of communities in case of ties.

Finally, for the selected γ, the consensus TF communities are obtained by majority voting across the 1000 runs. This yields a robust partition of TFs into communities that are consistently recovered across random initializations and parameter settings.

Assessment of community cohesion. To quantify the structural quality of the inferred TF communities, we compute a weighted cohesion metric for each community. Specifically, for a given community, we define:

$$w_{\text{ratio}} = \frac{w_{\text{internal}}}{w_{\text{internal}} + w_{\text{boundary}}},$$

where w_{internal} is the sum of edge weights for interactions between TFs within the community, and w_{boundary} is the sum of edge weights for interactions between TFs inside the community and those outside it. This metric ranges from 0 to 1, with higher values indicating stronger internal connectivity relative to external interactions.

To assess statistical significance, we perform a permutation test by randomly reassigning TFs to communities (while preserving community sizes) and recomputing the mean w_{ratio} across communities. This procedure is repeated 1,000 times to generate a null distribution. The observed mean w_{ratio} is then compared against this null, allowing us to determine whether the inferred communities exhibit greater internal cohesion than expected by chance.

A.3. TF community activity (extended description)

To obtain a low-dimensional representation of transcription factor (TF) activity, we summarize each TF community through an eigengene, defined as the first principal component (PC1) of the normalized expression matrix restricted to the TFs in the community (Alter et al., 2000). Let $X_C \in \mathbb{R}^{n \times |C|}$ denote the expression matrix for the TFs in community C across n cells. The eigengene is obtained as:

$$\mathrm{TFA}_C = X_C v_1,$$

where v_1 is the leading eigenvector of the covariance matrix of X_C. This projection captures the dominant axis of variation in TF expression within the community.

Eigengenes have been widely used in gene expression analysis, particularly in network-based methods, as robust summaries of coordinated gene activity (Hernaez et al., 2020). In the present context, they provide a natural way to map groups of TFs into low-dimensional regulatory signals.

We denote by TFA the collection of all TF community activities. In addition to one eigengene per community, we compute a global eigengene across all transcription factors:

$$\mathrm{TFA}_{\text{global}} = X_{\text{TF}} v_1,$$

where X_{TF} denotes the expression matrix restricted to all TFs. This global component captures broad transcriptional programs that may not be confined to a single community and can account for shared variability across modules.

The resulting set of TF community activities forms a low-dimensional representation of transcriptional regulation, which we interpret as a proxy for underlying regulatory programs. These quantities are subsequently used to guide the learning of causal latent variables in the gCRL framework.

A.4. T-MEX alignment and estimation of A

T-MEX is used to quantify the alignment between a learned latent representation and a reference representation. In its original formulation, T-MEX assumes a known correspondence between latent variables and reference variables, encoded in a binary matrix V, where $V_{ij} = 1$ if latent variable Z_j corresponds to reference variable TFA_i.

The method evaluates conditional associations between pairs of variables, producing a matrix W such that $W_{ij} = 1$ if Z_j remains associated with TFA_i after conditioning on all other reference variables. The T-MEX score is then defined as the discrepancy between V and W.

In our setting, we adapt T-MEX to serve as a differentiable alignment objective. Rather than using binary conditional independence tests, we compute T-MEX as the average partial correlation between corresponding components:

$$T\text{-}MEX(\mathrm{TFA}, Z) = \frac{1}{d} \sum_{i=1}^{d} \rho\big(Z_i, \mathrm{TFA}_i \mid \mathrm{TFA}_{-i}\big),$$

where $\rho(\cdot, \cdot \mid \cdot)$ denotes partial correlation and d is the dimensionality of the representation.

We assume a one-to-one correspondence between components by setting V to the identity matrix. Under this assumption, the optimal linear transformation A is estimated by solving:

$$A^* = \arg\max_A \; T\text{-}MEX(TFA, ZA),$$

which is optimized using gradient-based methods.

Importantly, in this work T-MEX is not used to certify exact recovery of latent causal variables, but rather as a proxy objective to align learned representations with biologically meaningful signals.

A.5. Permutation-based alignment in gCRL-VAE

In gCRL-VAE, the alignment between TF community activities and latent variables is learned through a permutation vector π, derived from a learnable matrix $S \in \mathbb{R}^{d \times d}$, where d is the latent dimensionality.

The matrix S is initialized as the identity and optimized during training. At each forward pass, the permutation is defined as:

$$\pi[k] = \arg\max_j S[k, j],$$

yielding a mapping from TF communities to latent dimensions.

This permutation plays three roles:

- **Routing:** TF community k is associated with latent variable $Z_{\pi[k]}$.
- **Causal ordering:** The DAG over Z is parameterized as an upper triangular matrix, and π determines the ordering of variables, ensuring acyclicity without requiring additional constraints.
- **Alignment:** The TFA matrix is permuted according to π before computing the alignment loss, ensuring correspondence between components.

The use of a learnable permutation allows the model to jointly infer both the latent variables and their alignment with biologically defined regulatory programs. By optimizing over permutations, we avoid fixing an arbitrary ordering of variables, while still enforcing a DAG structure through triangular parameterization.

A.6. AE and VAE training specifications

All models were trained using the Adam optimizer with a learning rate of 10^{-3} over 100 epochs. The latent dimension was inferred automatically and was not tuned as a hyperparameter: $d = n_{\text{communities}} + 1$, for gCRL-(V)AE and id-VAE, and $d = n_{\text{perturbations}}$ for d-VAE.

For the gCRL-AE, the batch size was set to 1024. Input features were standardized to zero mean and unit variance (`zscore`). Ten per cent of training cells were held out as a validation set. The encoder takes as input only TF gene expression, and passes it through a single hidden layer of 64 units (ReLU activation), while the decoder is a quadratic polynomial function that reconstructs the full transcriptome from the latent representation.

For the gCRL-VAE, id-VAE and d-VAE the batch size was set to 32. The KL divergence weight β_{KLD} was set to 2.0. Instead of maximum mean discrepancy (MMD), a centroid alignment loss was used to regularize the latent space, with a maximum weight $\alpha_{\text{centroid}} = 20.0$; the weight was linearly ramped from epoch 5 to the midpoint of training ($\lfloor T/2 \rfloor = 50$). An ℓ_1 sparsity penalty on decoder weights was applied with coefficient $\lambda_{\text{sparse}} = 10^{-3}$. For gCRL-VAE, the MCC alignment loss was included with weight $\lambda_{\text{MCC}} = 1.0$; this term was automatically set to zero for the discrepancy-VAE variants ($\lambda_{\text{MCC}} = 0$), since they do not use GRN priors.

B. Perturb-seq data results

B.1. Community-specific GO biological process for Joung et al. (2023)

Community 1. This community is characterized by processes related to calcium signaling, synaptic regulation, and developmental morphogenesis, including embryonic limb development and miRNA transcription. The combination of neuronal signaling and developmental pathways suggests a regulatory program associated with early differentiation and cell fate specification.

Community 2. Enrichment in actin cytoskeleton organization, filopodium assembly, and receptor clustering indicates a program related to cell morphology and signaling. Additional terms related to muscle differentiation and cardiac function suggest partial activation of mesodermal lineage programs.

Community 3. This community is strongly associated with endothelial biology and tissue development, including endothelial cell proliferation and migration, ossification, and Wnt signaling. These processes are indicative of vascular and mesodermal differentiation programs.

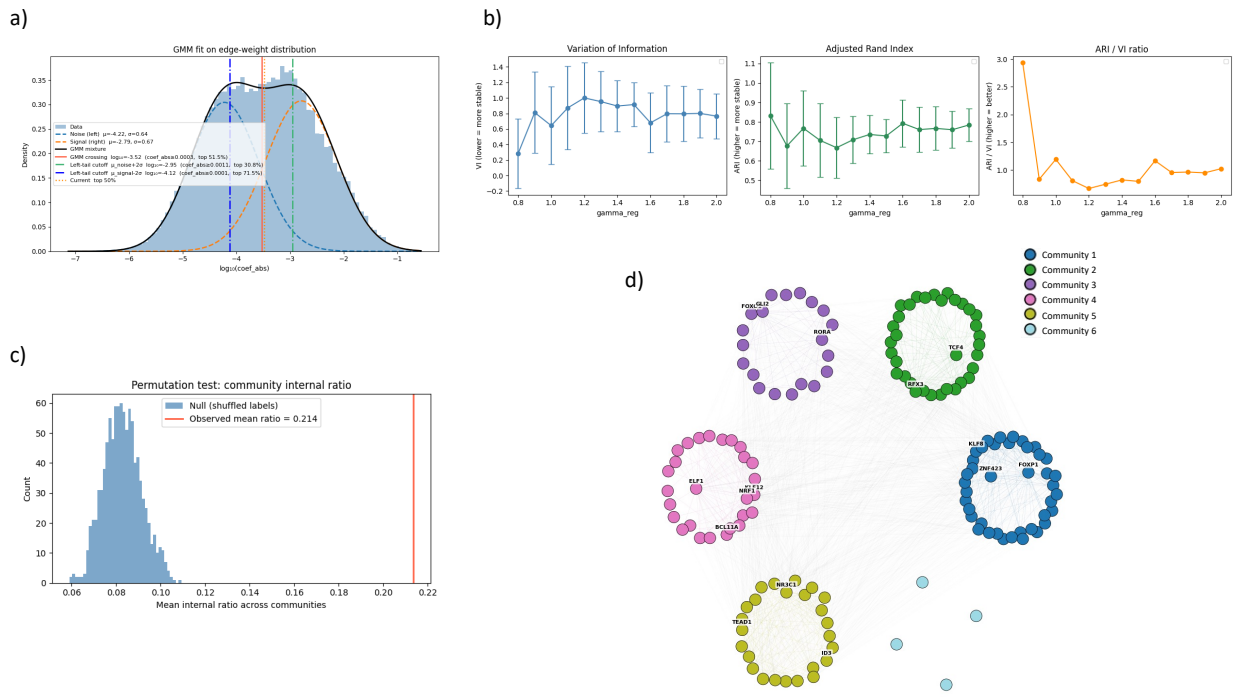

Figure S1. GRN reconstruction on data from Joung et al. (2023). Panel a reports the log-transformed distribution of GRN edges' coefficient (absolute value). A Gaussian Mixture Model (GMM) is used to discriminate between strong connection (right Gaussian) and weak ones (left Gaussian). The two distributions meet at the top 50% (red vertical solid line). Panel b: consistency across 1000 repetitions of the Leiden algorithm, measured over 200 randomly selected repetitions pairs through the Variation of Information (VI) and Adjusted Rand Index (ARI) metrics, as well as their ratio. Panel c: permutation analysis for the weighted cohesion metric. This metric quantifies the extent to which TFs within a community are more connected to each other than with other communities. Red solid line: value computed on the original community partition, contrasted against a null distribution (cyan) obtained by permuting TFs across communities. Panel d: TF-to-TF networks, grouped by community.

Community 4. Similar to Community 1, this cluster is enriched for calcium signaling, synaptic transmission, and developmental processes, but also includes epithelial proliferation and glycosaminoglycan metabolism, suggesting a broader regulatory program combining neuronal and tissue remodeling functions.

Community 5. This community shows strong enrichment for calcium signaling, Wnt signaling (both positive and negative regulation), and cell adhesion processes. The simultaneous presence of signaling and adhesion pathways suggests a regulatory module controlling cell communication and structural organization.

Community 6. This cluster is distinct in being enriched for membrane organization, protein trafficking, cytokine response (e.g., IL-6), autophagy, and VEGF signaling. These processes point to a regulatory program involved in cellular stress response, signaling, and membrane dynamics.

B.2. Community-specific GO biological process for Norman et al. (Norman et al., 2019)

Community 1. This community is enriched for granulocyte chemotaxis, nitric oxide biosynthesis, and cell adhesion processes, suggesting a regulatory program related to immune cell migration and inflammatory signaling. The presence of nervous system development terms indicates potential overlap with broader differentiation programs.

Community 2. This community is characterized by strong enrichment in immune activation and inflammatory signaling, including T cell proliferation, TNF production, and PI3K/AKT signaling, alongside cytoskeletal organization. This suggests a regulatory module associated with immune activation and cell signaling.

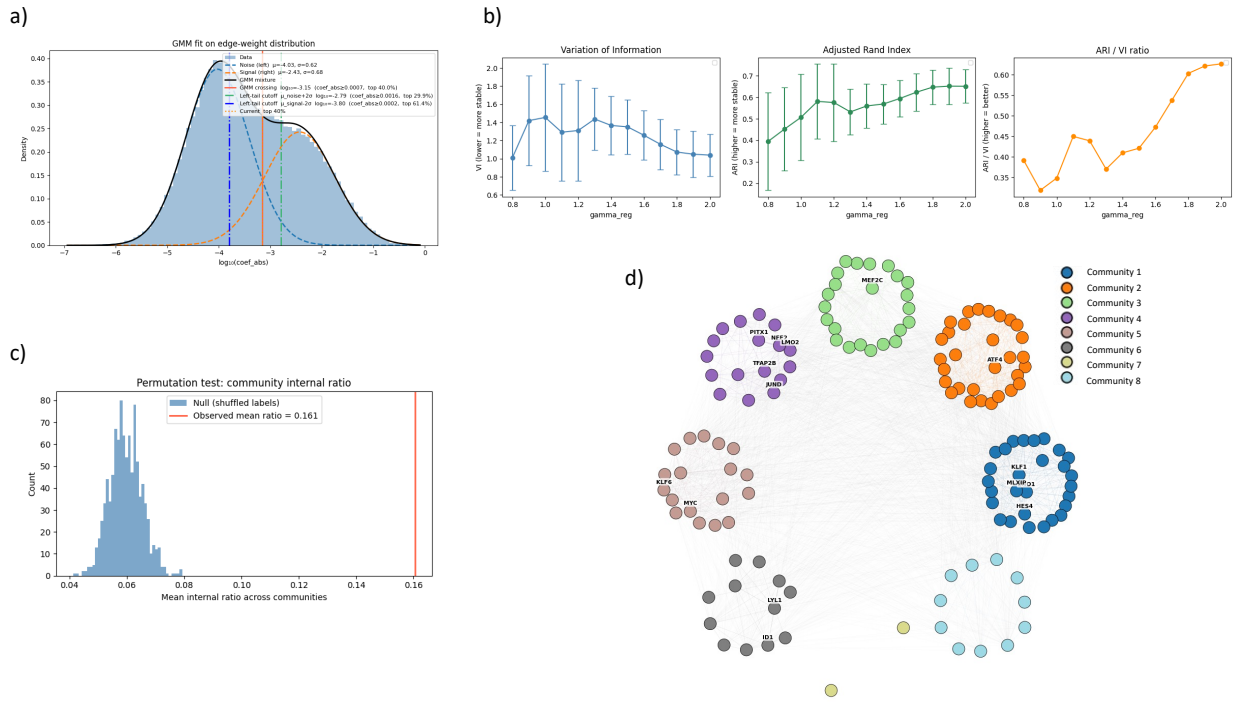

Figure S2. GRN reconstruction on data from Norman et al. (2019). Panel a reports the log-transformed distribution of GRN edges' coefficient (absolute value). A Gaussian Mixture Model (GMM) is used to discriminate between strong connection (right Gaussian) and weak ones (left Gaussian). The two distributions meet at the top 40% (red vertical solid line). Panel b: consistency across 1000 repetitions of the Leiden algorithm, measured over 200 randomly selected repetitions pairs through the Variation of Information (VI) and Adjusted Rand Index (ARI) metrics, as well as their ratio. Panel c: permutation analysis for the weighted cohesion metric. This metric quantifies to which extent TFs within a community are more connected to each other than with other communities. Red solid line: value computed on the original community partition, contrasted against a null distribution (cyan) obtained by permuting TFs across communities. Panel d: TF-to-TF networks, grouped by community.

Community 3. Enrichment in MAPK signaling, cell cycle regulation, endothelial migration, and erythroid differentiation indicates a program combining proliferation, signaling, and lineage commitment. This is consistent with perturbations affecting both growth and differentiation trajectories.

Community 4. This community captures vasculature development, inflammatory signaling, erythroid differentiation, and JAK-STAT signaling. These processes suggest a regulatory module associated with hematopoietic and vascular programs.

Community 5. This cluster shows strong enrichment for cytokine production (especially TNF), inflammatory response, epithelial-to-mesenchymal transition (EMT), and endothelial migration. This profile is indicative of a highly active signaling module driving inflammation and cell state transitions.

Community 6. This community is associated with reactive oxygen species metabolism, collagen biosynthesis, leukocyte migration, and neuronal generation. These processes point to a mixed regulatory program involving tissue remodeling, immune activity, and differentiation.

Community 7. Although enrichment is weaker, this community is associated with cell cycle checkpoints, DNA repair, and apoptosis-related pathways, suggesting a regulatory module linked to genomic stability and stress response.

Community 8. This community combines strong inflammatory signaling (TNF production), muscle development, ER stress response, and apoptosis. This suggests a regulatory program associated with stress-induced differentiation and cellular remodeling.

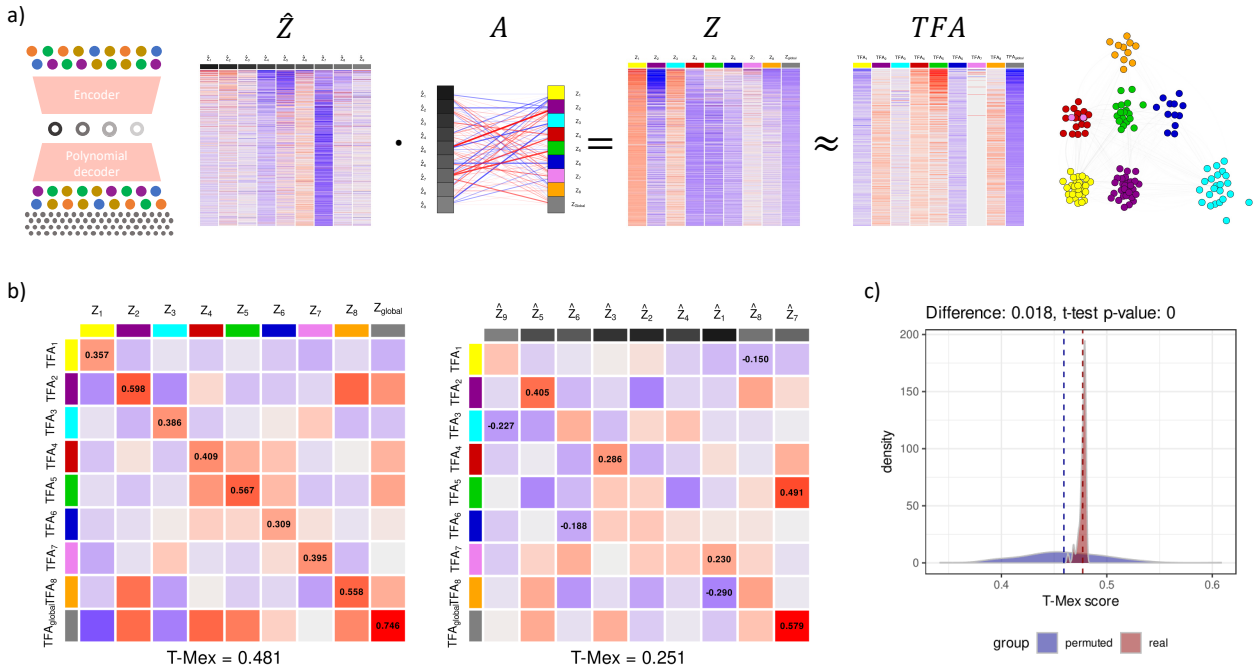

Figure S3. gCRL-AE operation and results on data from Norman et al. (2019). Panel a: The entangled embeddings $\hat{Z}$ are obtained from an autoencoder with a polynomial decoder. A rotation matrix A is then used to align these embeddings with the TF community activities TFA, yielding the disentangled representation Z. Panel b: Correlation matrices between TFA and Z (left), and between TFA and $\hat{Z}$ (right). The disentangled embeddings Z show substantially higher alignment with TF community activities, as further supported by an increased T-MEX score. Panel c: Comparison of technical variability, measured by deriving A across different random seeds (red), with the variability obtained by permuting TFs across communities (blue). Technical variability is an order of magnitude lower, and the T-MEX value corresponding to the original community partition is significantly higher than expected by chance.

Table S1. Representative GO Biological Process enrichments across TF communities in Joung et al. (2023). Background processes correspond to terms enriched in at least 50% of the communities with adjusted p-value (FDR) < 0.05.

CATEGORY	GO BIOLOGICAL PROCESS	ADJ. p-VALUE
BACKGROUND	ADHERENS JUNCTION ORGANIZATION (GO:0034332)	—
BACKGROUND	AXON GUIDANCE (GO:0007411)	—
BACKGROUND	AXONOGENESIS (GO:0007409)	—
BACKGROUND	EXTRACELLULAR MATRIX ORGANIZATION (GO:0030198)	—
BACKGROUND	NEURAL CREST CELL MIGRATION (GO:0001755)	—
BACKGROUND	REGULATION OF SYNAPSE ORGANIZATION (GO:0050807)	—
BACKGROUND	REGULATION OF NEUROTRANSMITTER SECRETION (GO:0046928)	—
BACKGROUND	REGULATION OF CALCIUM ION TRANSMEMBRANE TRANSPORT (GO:1903169)	—
BACKGROUND	REGULATION OF SMALL GTPASE MEDIATED SIGNAL TRANSDUCTION (GO:0051056)	—
BACKGROUND	SYNAPSE ASSEMBLY (GO:0007416)	—
COMMUNITY 1	RELEASE OF SEQUESTERED CALCIUM ION INTO CYTOSOL (GO:0051209)	1.34×10^{-2}
COMMUNITY 1	NEGATIVE REGULATION OF CELL PROJECTION ORGANIZATION (GO:0031345)	1.35×10^{-2}
COMMUNITY 1	REGULATION OF SYNAPTIC TRANSMISSION, GABAERGIC (GO:0032228)	3.12×10^{-2}
COMMUNITY 2	REGULATION OF FILOPODIUM ASSEMBLY (GO:0051489)	6.10×10^{-2}
COMMUNITY 2	HEART CONTRACTION (GO:0060047)	6.83×10^{-2}
COMMUNITY 2	ACTIN FILAMENT ORGANIZATION (GO:0007015)	7.64×10^{-2}
COMMUNITY 3	REGULATION OF ENDOTHELIAL CELL PROLIFERATION (GO:0001936)	7.47×10^{-3}
COMMUNITY 3	POSITIVE REGULATION OF OSSIFICATION (GO:0045778)	1.86×10^{-2}
COMMUNITY 3	REGULATION OF WNT SIGNALING PATHWAY (GO:0030111)	2.45×10^{-2}
COMMUNITY 4	RELEASE OF SEQUESTERED CALCIUM ION INTO CYTOSOL (GO:0051209)	5.92×10^{-3}
COMMUNITY 4	REGULATION OF SYNAPTIC TRANSMISSION, GABAERGIC (GO:0032228)	9.63×10^{-3}
COMMUNITY 4	CALCIUM ION TRANSMEMBRANE TRANSPORT (GO:0070588)	2.50×10^{-2}
COMMUNITY 5	RELEASE OF SEQUESTERED CALCIUM ION INTO CYTOSOL (GO:0051209)	3.44×10^{-3}
COMMUNITY 5	POSITIVE REGULATION OF WNT SIGNALING PATHWAY (GO:0030177)	3.07×10^{-2}
COMMUNITY 5	NEGATIVE REGULATION OF WNT SIGNALING PATHWAY (GO:0030178)	3.43×10^{-2}
COMMUNITY 6	PLASMA MEMBRANE ORGANIZATION (GO:0007009)	5.62×10^{-2}
COMMUNITY 6	POSITIVE REGULATION OF AUTOPHAGY (GO:0010508)	7.18×10^{-2}
COMMUNITY 6	VASCULAR ENDOTHELIAL GROWTH FACTOR RECEPTOR SIGNALING PATHWAY (GO:0048010)	7.18×10^{-2}

Table S2. Representative GO Biological Process enrichments across TF communities in the Norman et al. (2019) dataset. Background processes correspond to terms enriched in at least 50% of the communities with adjusted p-value (FDR) < 0.05.

CATEGORY	GO BIOLOGICAL PROCESS	ADJ. p-VALUE
BACKGROUND	CELLULAR RESPONSE TO TRANSFORMING GROWTH FACTOR BETA STIMULUS (GO:0071560)	—
BACKGROUND	ERK1 AND ERK2 CASCADE (GO:0070371)	—
BACKGROUND	ERYTHROCYTE DIFFERENTIATION (GO:0030218)	—
BACKGROUND	MYELOID LEUKOCYTE DIFFERENTIATION (GO:0002573)	—
BACKGROUND	POSITIVE REGULATION OF ANGIOGENESIS (GO:0045766)	—
BACKGROUND	POSITIVE REGULATION OF CELL CYCLE (GO:0045787)	—
BACKGROUND	POSITIVE REGULATION OF PHAGOCYTOSIS (GO:0050766)	—
BACKGROUND	REGULATION OF TUMOR NECROSIS FACTOR PRODUCTION (GO:0032680)	—
BACKGROUND	T CELL DIFFERENTIATION (GO:0030217)	—
BACKGROUND	REGULATION OF INTRINSIC APOPTOTIC SIGNALING PATHWAY (GO:2001242)	—
COMMUNITY 1	POSITIVE REGULATION OF GRANULOCYTE CHEMOTAXIS (GO:0071624)	1.49×10^{-2}
COMMUNITY 1	POSITIVE REGULATION OF NITRIC OXIDE BIOSYNTHETIC PROCESS (GO:0045429)	2.45×10^{-2}
COMMUNITY 1	POSITIVE REGULATION OF KINASE ACTIVITY (GO:0033674)	2.96×10^{-2}
COMMUNITY 2	POSITIVE REGULATION OF T CELL PROLIFERATION (GO:0042102)	1.04×10^{-2}
COMMUNITY 2	POSITIVE REGULATION OF TUMOR NECROSIS FACTOR PRODUCTION (GO:0032760)	2.07×10^{-2}
COMMUNITY 2	ACTIN FILAMENT ORGANIZATION (GO:0007015)	1.40×10^{-2}
COMMUNITY 3	POSITIVE REGULATION OF ENDOTHELIAL CELL MIGRATION (GO:0010595)	8.14×10^{-3}
COMMUNITY 3	MAPK CASCADE (GO:0000165)	2.17×10^{-2}
COMMUNITY 3	POSITIVE REGULATION OF ERYTHROCYTE DIFFERENTIATION (GO:0045648)	3.87×10^{-2}
COMMUNITY 4	SKELETAL MUSCLE TISSUE DEVELOPMENT (GO:0007519)	7.48×10^{-3}
COMMUNITY 4	REGULATION OF RECEPTOR SIGNALING PATHWAY VIA JAK-STAT (GO:0046425)	1.42×10^{-2}
COMMUNITY 4	POSITIVE REGULATION OF REACTIVE OXYGEN SPECIES METABOLIC PROCESS (GO:2000379)	2.30×10^{-2}
COMMUNITY 5	POSITIVE REGULATION OF TUMOR NECROSIS FACTOR PRODUCTION (GO:0032760)	1.88×10^{-3}
COMMUNITY 5	POSITIVE REGULATION OF EPITHELIAL TO MESENCHYMAL TRANSITION (GO:0010718)	1.84×10^{-2}
COMMUNITY 5	POSITIVE REGULATION OF INFLAMMATORY RESPONSE (GO:0050729)	1.78×10^{-2}
COMMUNITY 6	POSITIVE REGULATION OF LEUKOCYTE MIGRATION (GO:0002687)	3.62×10^{-2}
COMMUNITY 6	GENERATION OF NEURONS (GO:0048699)	4.03×10^{-2}
COMMUNITY 6	POSITIVE REGULATION OF VASCULATURE DEVELOPMENT (GO:1904018)	4.03×10^{-2}
COMMUNITY 7	MITOTIC G2/M TRANSITION CHECKPOINT (GO:0044818)	1.24×10^{-1}
COMMUNITY 7	DOUBLE-STRAND BREAK REPAIR VIA HOMOLOGOUS RECOMBINATION (GO:0000724)	1.24×10^{-1}
COMMUNITY 7	REGULATION OF DNA METABOLIC PROCESS (GO:0051052)	1.24×10^{-1}
COMMUNITY 8	POSITIVE REGULATION OF TUMOR NECROSIS FACTOR PRODUCTION (GO:0032760)	5.40×10^{-3}
COMMUNITY 8	POSITIVE REGULATION OF REACTIVE OXYGEN SPECIES METABOLIC PROCESS (GO:2000379)	3.01×10^{-2}
COMMUNITY 8	INTRINSIC APOPTOTIC SIGNALING PATHWAY IN RESPONSE TO ENDOPLASMIC RETICULUM STRESS (GO:0070059)	4.27×10^{-2}