GGeraPHF: Graph Generative Poisson Hierarchical Factorization

Mingxuan Zhang [*1] **Kevin Hoffer-Hawlik** [*2] **Benjamin Izar** [3] **Elham Azizi** [2]

Abstract

Gene regulatory network (GRN) models provide insight into mechanisms underlying cellular function. While previous methods have attempted to infer GRNs from single cell RNA sequencing (scRNA-seq) data, they are limited in interpretability, and do not explain how GRNs impact cell state transitions and plasticity. We developed Stochastic Block Graph Generative Poisson Hierarchical Factorization, or GGeraPHF, to identify regulons, defined as communities of genes, driving cell plasticity. GGeraPHF combines tasks of community detection, graph structure learning, matrix factorization, and low rank estimation to learn an interpretable, joint cell and gene latent space from scRNA-seq and GRN data. We applied GGeraPHF to a simulated dataset and real tumor scRNA-seq data. GGeraPHF faithfully reconstructs expression matrices, refines GRNs as densely connected communities, and successfully associates them to heterogeneous cell populations.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has enabled the characterization of cell heterogeneity and plasticity at high resolution in complex diseases such as cancer. scRNA-seq also provides unique opportunities to study the role of gene regulatory circuitry at a fine scale, as alteration of gene regulatory relationships are often driving forces behind cell function and fate. Gene regulatory networks (GRNs) are popular models for disentangling interactions between transcription factors (TFs) and target genes in dynamic biological systems, and an active field of research is reverse-engineering GRNs from scRNA-seq (Mercatelli, 2020). Identification of regulons, or jointly co-regulated genes, driving cell plasticity would be a boon for uncovering mechanisms involved in therapeutic response, such as cancer and immune cell plasticity during/after immunotherapy in melanoma (Madhamshettiwar, 2012; Marusyk, 2012). However, biological and technical noise and sparsity from high drop-out rates cause poor performance for most GRN inference methods, which were originally designed for bulk RNA sequencing data. More importantly, current methods are not capable of pinpointing regulatory mechanisms that drive cell plasticity (Mercatelli, 2020; Iglesias-Martinez, 2021; Osorio, 2020; Keyl, 2023). In particular, identifying topological structures of GRNs (e.g., hubs, communities) associated with heterogeneous and altered cell states in disease can reveal novel drug targets for reversing abnormal cell plasticity. We propose a novel framework with joint Bayesian modeling of scRNA-seq and GRN data, to identify GRN structures and regulons explaining cell state transitions.

We present Stochastic Block **G**raph **Ge**ne**ra**tive **P**oisson **H**ierarchical **F**actorization, or GGeraPHF, a Bayesian hierarchical model to achieve both goals of generating refined GRN structure using scRNA-seq and associating GRN topologies to cell sub-populations to identify interpretable regulons driving cell plasticity. Novelly, GGeraPHF combines tasks of community detection, graph structure learning, and matrix factorization and low rank estimation. Matrix factorization allows the model to learn joint gene and cell factors. A stochastic block graph is used to leverage the joint factors in defining community structures of genes, interpretable as regulons. Densely-connected communities are then assigned to sub-populations of cells, for which up-regulation of key regulons may drive cell plasticity. Finally, inter-community edges are penalized and intra-community edges are rewarded, thus preserving only meaningful community interactions and refining expected network topology. We apply GGeraPHF to both a simulated scRNA-seq dataset with pre-defined ground-truth co-expression patterns as well as a clinical dataset from melanoma patients treated with immunotherapy. We show GGeraPHF holds promise in disentangling GRN and cell fate dynamics from scRNA-seq.

[*]Equal contribution [1]Department of Systems Biology, Columbia University Medical Center, New York, NY, USA [2]Department of Biomedical Engineering and Irving Institute for Cancer Dynamics, Columbia University, New York, NY USA [3]Columbia Center for Translational Immunology, Department of Medicine, Division of Hematology and Oncology, and Program for Mathematical Genomics, Columbia University Medical Center, New York, NY USA. Correspondence to: Mingxuan Zhang <mingxuan.zhang@columbia.edu>, Kevin Hoffer-Hawlik <kh3205@columbia.edu>.

The 2023 ICML Workshop on Computational Biology. Honolulu, Hawaii, USA, 2023.

2. Methods

2.1. The GGeraPHF Model

Suppose we have a scRNA-seq dataset with $\mathbf{C}$ cells and $\mathbf{G}$ genes, unique molecular identifier (UMI) counts X is a $\mathbf{C} \times \mathbf{G}$ matrix with integer entries. We also consider an input GRN $N = \{E, V\}$ representing prior knowledge on possible regulatory links, where E represents the set of genes and V represents edges connecting pairs of genes that regulate each other. We define 'factors' as sets of co-expressed genes associated with clusters of cells. If we have k factors, our objective is to learn a cell loading $W \in \mathbf{R}^{\mathbf{C}\times\mathbf{k}}$ and a gene loading $H \in \mathbf{R}^{\mathbf{G}\times\mathbf{k}}$ where the low-rank count matrix $\hat{X} = WH^T$ resembles the input matrix X. We also aim to jointly learn a refined GRN $A_G \in \mathbf{R}^{\mathbf{G}\times\mathbf{G}}$ where the factors correspond to community structures of the graph and the weights indicate the probability of two genes regulating each other.

We regularize the model to preserve the topological structure of the input graph. To achieve this, we designed the following generative model inspired by the ideas of hierarchical Poisson factorization (Gopalan, 2015) and stochastic block models (Lee, 2019) (Fig. 1):

$$\beta_i \sim Gamma(a', b'), \beta_g \sim Gamma(c', d')$$
$$W_{ik} \sim Gamma(a, \beta_i), H_{gk} \sim Gamma(c, \beta_g)$$
$$x_{ig}|W_{ik}, H_{gk} \sim Poisson(W_{ik}H_{gk}^T)$$
$$\pi_k \sim Dirichlet(\mathbf{E}_g[H_{gk}]), z_{gk} \sim Categorical(\pi_k)$$
$$\eta_{kk} \sim Beta(\alpha, \beta)$$
$$\Lambda = z_{gk}\eta_{kk}z_{gk}^T, \Lambda \in \mathbf{R}^{G\times G}$$
$$A_G|\Lambda \sim Bernoulli(\Lambda)$$

$a', b', c', d', a, c, \alpha, \beta$ are hyperparameters. W_{ik} are cell weights across k factors based on cell budget β_i. H_{gk} are gene weights across k factors based on gene budget β_g. π is the factor membership assignment probability parameterized by a Dirichlet distribution with Gamma prior given by the mean gene weights in each factor. z is the per-gene factor membership sampled from a Categorical distribution. η is the block matrix sampled from Beta distributions where the diagonal entries represent the probability of within-factor connections, and the off-diagonal entries represent the probability of cross-factor connections, to capture cascades of TFs between communities. z and η jointly determine the edge probability between pairs of genes, and the generated GRN A_G is sampled from a Bernoulli distribution.

2.2. Hyperparameter Initialization and Regularization

We choose to initialize hyperparameters b' and d' to preserve the variance-to-mean ratio of total UMI counts per cell or gene in the sampling distributions of gene/cell budgets (Mendes Levitin, 2019). Specifically, we set b' and d' as:

$$b' = \frac{\mathbf{Var}[\sum_g x_{ig}]}{\mathbf{E}[\sum_g x_{ig}]}, d' = \frac{\mathbf{Var}[\sum_i x_{ig}]}{\mathbf{E}[\sum_i x_{ig}]}$$

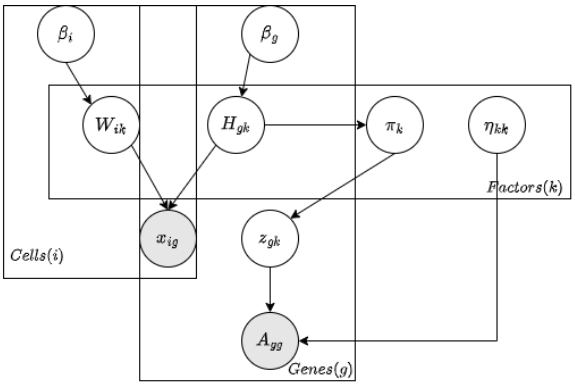

Figure 1. Graphical representation of the model. Circles represent latent variables. Colored circles represent observations.

where x_{ig} are individual entries of count matrix X. To enforce sparsity, we initialize the Gamma shape parameters a', c' and a, c as 1.0 and 0.3, respectively. The Beta distribution parameters α, β for η are initialized to non-informative values 1.0 and 1.0. Since we expect the generated GRN to have densely connected communities and loose connections between communities, we learn a sparsity regularizer ρ and construct a mask $M \in \mathbf{R}^{k\times k}$. Assuming $\rho \sim Beta(\gamma, 5\gamma)$, the mask matrix is constructed as:

$$M_{ij} = \begin{cases} 1, & \text{if } i = j \\ \rho, & \text{if } i \neq j \end{cases}$$

Then, the regularized Bernoulli rate is given by $\Lambda = z(\eta \odot M)z^T$. This regularization method only allows strong signals of cross-community connection to be preserved, which prunes the input GRN and refines the graph's topological structure according to observed phenotypic states.

2.3. Model Inference

Inference is conducted by Markov Chain Monte Carlo (MCMC) sampling. The traditional Metropolis-Hasting (MH) algorithm does not scale well for high dimensional distributions due to the random walk nature of its movement. Therefore, we perform parameter inference under the Hamiltonian Monte Carlo (HMC) framework, where gradient information of the target distribution is used to guide the sampler movement and make distant proposals with high acceptance probabilities. Our model contains both discrete and continuous latent variables, which renders traditional HMC-based algorithms such as No U-Turn Sampling (NUTS) ineffective. To overcome the mixed nature of the latent space, we apply Gibbs sampling at discrete sites and NUTS at continuous sites to sample from the target distribution. The models are trained with 4 chains with a warm-up distance of 100 and a sampling distance of 400. The sampling converged with an $\hat{R}$ value less than 1.05, indicating that the chains are well-mixed.

2.4. Data Preparation

2.4.1. Simulated scRNA-seq Data

We simulated an scRNA-seq dataset with known co-expression patterns and a known set of gene factors associated with cell grouping patterns, to test GGeraPHF performance in learning and reconstructing interpretable joint cell and gene latent spaces. We used ESCO (Tian, 2021) which simulates scRNA-seq by incorporating variation in expression from cell heterogeneity (i.e., differentially expressed genes or DEGs), intrinsic variation in gene expression between similar cell types, technical noise, and co-expression patterns using a Gaussian copula. We generated data for 2000 cells and 200 genes with global DEG probability of 0.5 (for 100 DEGs total) and group-specific DEG probability of 0.3. Cells were divided into three groups with 60% of cells in group 1 and 20% in groups 2 and 3. To create GGeraPHF input, we filtered the count matrix for the 100 DEGs and constructed a GRN by calculating empirical correlation between gene pairs for the 100 DEGs.

2.4.2. Melanoma Clinical scRNA-seq Data

We also tested GGeraPHF on a recently published clinical dataset consisting of scRNA-seq of tumor samples from melanoma patients treated with immunotherapy (Wang, 2023). Data are from biopsies from a single patient before and while receiving immunotherapy (about 8,000 cells total). We performed feature selection on the union of the top 3,000 highly variable genes (HVGs) across both samples and a list of about 1,900 known TFs in humans. The unnormalized count matrix was input into GGeraPHF. After normalizing and log transforming counts, we attempted GRN inference with GRNBoost from the SCENIC pipeline (Aibar, 2017), but the result GRN was densely connected with no local structure. Subsequently, we constructed a filtered covariance matrix between each HVG pair (binarized by non-zero covariance interactions and only keeping TF-involved interactions) as the initial GRN for GGeraPHF. Alternative graph inputs to GGeraPHF can be obtained using other GRN inference methods (Lachmann, 2016; Passemiers, 2022) or causal graphs (Squires, preprint; Lopez, preprint).

3. Results

3.1. Simulated scRNA-seq Data

GGeraPHF correctly reconstructed input scRNA-seq counts and the structure of the ground-truth GRN, while recovering factors representing densely-connected communities. Reconstructed counts and GRN were sampled from GGeraPHF following its generative process with parameters that are posterior means. Reconstructed and original counts had tight linear fit with slope close to 1 ($R^2 = 0.990$, regression line $y = 0.984x + 0.066$), confirming that GGeraPHF accurately reconstructs UMI counts. The ground-truth GRN contained three densely-connected modules representing three sets of DEGs across three distinct cell groups, which GGeraPHF recovered with high fidelity and refined structure (Supp Fig. 6).

Additionally, GGeraPHF summarized groups of genes from densely-connected modules into factors describing cell heterogeneity, and it correctly assigned factors to distinct subgroups of cells (Fig. 2). The simulated dataset was designed such that specific cell groups were defined by sets of DEGs. Hence, we computed co-expression of genes within factors and observed selective enrichment in their respective cell group (Fig. 2), indicating that learned factors estimate ground truth DEG sets defining cell heterogeneity.

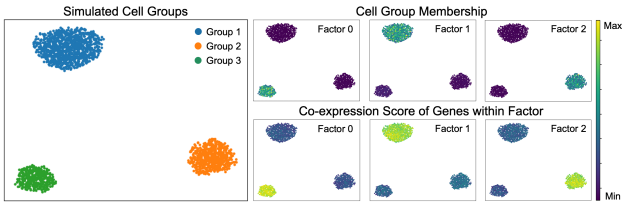

Figure 2. Left: UMAP of ESCO simulated scRNA-seq data based on 100 DEGs, visualizing pre-defined cell groups. Top right: GGeraPHF correctly associates learned factors to the ground truth cell groups. Cells are colored by the cell weights normalized by the cell capacity. Bottom right: GGeraPHF learns factors that resemble ground truth DEG sets. Cells are colored by co-expression scores given by the sum of gene expression values of genes in each factor.

3.2. Melanoma Clinical scRNA-seq Data

To investigate the effect of GGeraPHF on real scRNA-seq data, we tested the model on the melanoma clinical dataset and compared the performance to a naive Hierarchical Poisson Factorization (HPF) model (Mendes Levitin, 2019). GGeraPHF achieved better reconstruction performance ($R^2 = 0.775$, regression line $y = 0.677x + 0.113$) than HPF ($R^2 = 0.734$, regression line $y = 0.605x + 0.152$). GGeraPHF also generated a GRN capturing both global and local topology of the input network (Fig. 3 left, middle).

The melanoma dataset contained cells with no clear prior relationship between gene modules and cell groupings across samples. Nonetheless, GGeraPHF uncovered interpretable factors and associated them to distinct groups of cells, while HPF failed on these tasks (Fig. 4).

To further interpret factors learned by GGeraPHF, we examined tumor clonality predicted by InferCNV (Patel, 2014). The factors were associated with clones 1 and 3, which are two clones that exist across both samples. Factor 1 associated with on-treatment clone 1 cells, which retained its rough size. Factor 2 represented both clone 1 and clone 3 pre-treatment cells. Factor 3 was enriched in on-treatment

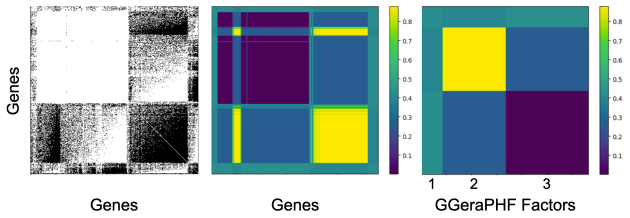

Figure 3. Left: Hierarchically sorted binary adjacency matrix of the original GRN from melanoma scRNA-seq data. Two genes can be connected if their empirical covariance is non-zero and at least one is a TF. Middle: Adjacency matrix of the generated GRN, with genes ordered as in the left figure and weights corresponding to learned edge probabilities. Right: Adjacency matrix of the generated GRN, with genes sorted by corresponding learned factors.

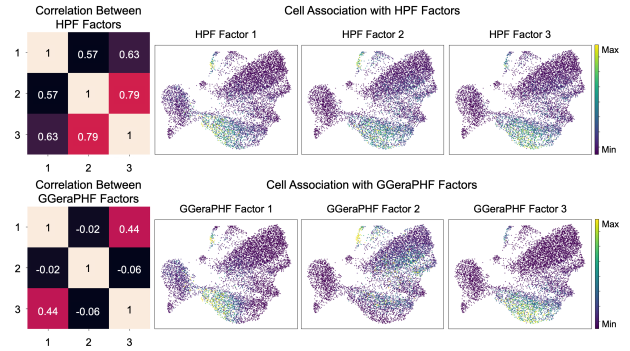

Figure 4. Top: Correlations between factors identified by HPF model on melanoma dataset, and enrichment of HPF factors across melanoma cells. Bottom: Correlations between factors identified by GGeraPHF model on melanoma dataset, and enrichment of GGeraPHF factors across melanoma cells.

clone 3 cells, which expanded significantly between pre- and on-treatment sample timepoints (Fig. 5). Genes in the factors also displayed different regulatory patterns corresponding to respective clonal dynamics. Factor 1 represented a set of TFs with global regulation effect. Factor 2 was a self-regulating gene module but had weak regulatory effect on Factor 3 genes. Factor 3 genes were most likely targets with no interactions between each other (Fig. 3 right).

Finally, we performed gene set enrichment analysis (GSEA) on the factors to support that the learned modules characterizing tumor heterogeneity were also biologically relevant (Subramanian, 2005; Fang, 2023). Most factors had significant enrichment for gene sets with false discovery rate (FDR) below 0.250, including genes from both well-profiled (e.g., IL-6/JAK/STAT3, PI3K/AKT/mTOR signaling) and under-studied pathways (e.g., cholesterol homeostasis) in melanoma (Gu, 2022) (Supp Fig. 7). Taken together, when applied to the melanoma dataset, GGeraPHF links gene regulation to tumor heterogeneity using interpretable factors, and these factors represent regulons driving cell plasticity through various pathways in melanoma.

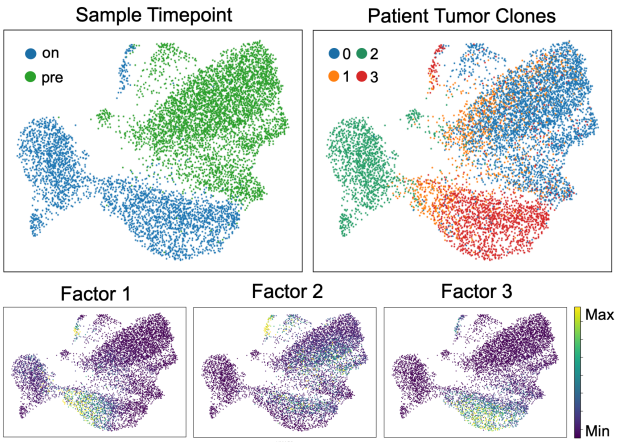

Figure 5. Top left: UMAP of melanoma scRNA-seq data after feature selection, visualizing patient samples before and on immunotherapy. Top right: Tumor clonality, inferred using InferCNV. Bottom: GGeraPHF factor association to heterogenous cell groups. Each factor correspond to a row of the learned cell weight matrix, and cells are colored by the cell weights normalized by the learned cell capacity.

4. Conclusion and Applications

GGeraPHF shows promise in disentangling gene regulatory elements and their role in transcriptional fate. We tested the model on simulated and real tumor sample scRNA-seq data. We showed that on both datasets, GGeraPHF learned interpretable factors corresponding to GRN neighborhoods and mapped topological structures to heterogenous cell groups. Future work includes fine-tuning the graph learning process to work on directed causal graphs, and further interpreting network structures within learned communities. Additionally, we are working to expand this model to additional patient tumor and immune cell data. By expanding our focus to both tumor and immune cells, we hope to gain a comprehensive understanding of mechanisms associated with melanoma progression and effector cell heterogeneity driving patient response or resistance to cancer immunotherapy. Especially as clinical scRNA-seq datasets become increasingly available across cancer types and treatment conditions, GGeraPHF and its ability to identify key cell state-specific regulons can bolster the importance of well-characterized pathways as well as uncover understudied biological mechanisms, towards improving cancer therapeutics and patient outcomes.

References

Aibar, S. SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, 2017. doi: 10.1038/nmeth.4463.

Fang, Z. GSEApy: a comprehensive package for performing

gene set enrichment analysis in python. *Bioinformatics*, 2023. doi: 10.1093/bioinformatics/btac757.

Gopalan, P. Scalable recommendation with hierarchical poisson factorization. *Association for Uncertainty in AI Proceedings*, 2015.

Gu, J. Cholesterol homeostasis and cancer: a new perspective on the low-density lipoprotein receptor. *Cellular Oncology*, 2022. doi: 10.1007/s13402-022-00694-5.

Iglesias-Martinez, L. KBoost: a new method to infer gene regulatory networks from gene expression data. *Scientific Reports*, 2021. doi: 10.1038/s41598-021-94919-6.

Keyl, P. Single-cell gene regulatory network prediction by explainable AI. *Nucleic Acids Research*, 2023. doi: 10.1093/nar/gkac1212.

Lachmann, A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 2016. doi: doi.org/10.1093/bioinformatics/btw216.

Lee, C. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 2019. doi: 10.1007/s41109-019-0232-2.

Lopez, R. Large-scale differentiable causal discovery of factor graphs. preprint. doi: 10.48550/arXiv.2206.07824.

Madhamshettiwar, P. B. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome medicine*, 2012. doi: 10.1186/gm340.

Marusyk, A. Intra-tumour heterogeneity: a looking glass for cancer? *Nature reviews. Cancer*, 2012. doi: 10.1038/nrc3261.

Mendes Levitin, H. De novo gene signature identification from single-cell RNA-seq with hierarchical poisson factorization. *Molecular Systems Biology*, 2019. doi: 10.15252/msb.20188557.

Mercatelli, D. Gene regulatory network inference resources: A practice overview. *BBA - Gene Regulatory Mechanisms*, 2020. doi: 10.1016/j.bbagrm.2019.194430.

Osorio, D. scTenifoldNet: A machine learning workflow for constructing and comparing transcriptome-wide gene regulatory networks from single-cell data. *Cell Press Patterns*, 2020. doi: 10.1016/j.patter.2020.100139.

Passemiers, A. Fast and accurate inference of gene regulatory networks through robust precision matrix estimation. *Bioinformatics*, 2022. doi: 10.1093/bioinformatics/btac178.

Patel, A. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 2014. doi: 10.1126/science.1254257.

Squires, C. Permutation-based causal structure learning with unknown intervention targets. preprint. doi: 10.48550/arXiv.1910.09007.

Subramanian, A. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 2005. doi: 10.1073/pnas.0506580102.

Tian, J. ESCO: single cell expression simulation incorporating gene co-expression. *Bioinformatics*, 2021. doi: 10.1093/bioinformatics/btab116.

Wang, Y. Multimodal single-cell and whole-genome sequencing of small, frozen clinical specimens. *Nature Genetics*, 2023. doi: 10.1038/s41588-022-01268-9.

A. Supplementary Figures

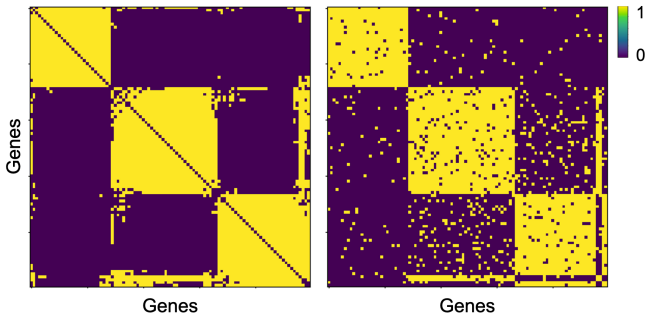

Figure 6. Left: Binary adjacency matrix of original GRN from simulated scRNA-seq data, hierarchically sorted. Right: Binary adjacency matrix of reconstructed GRN, with unchanged sorting.

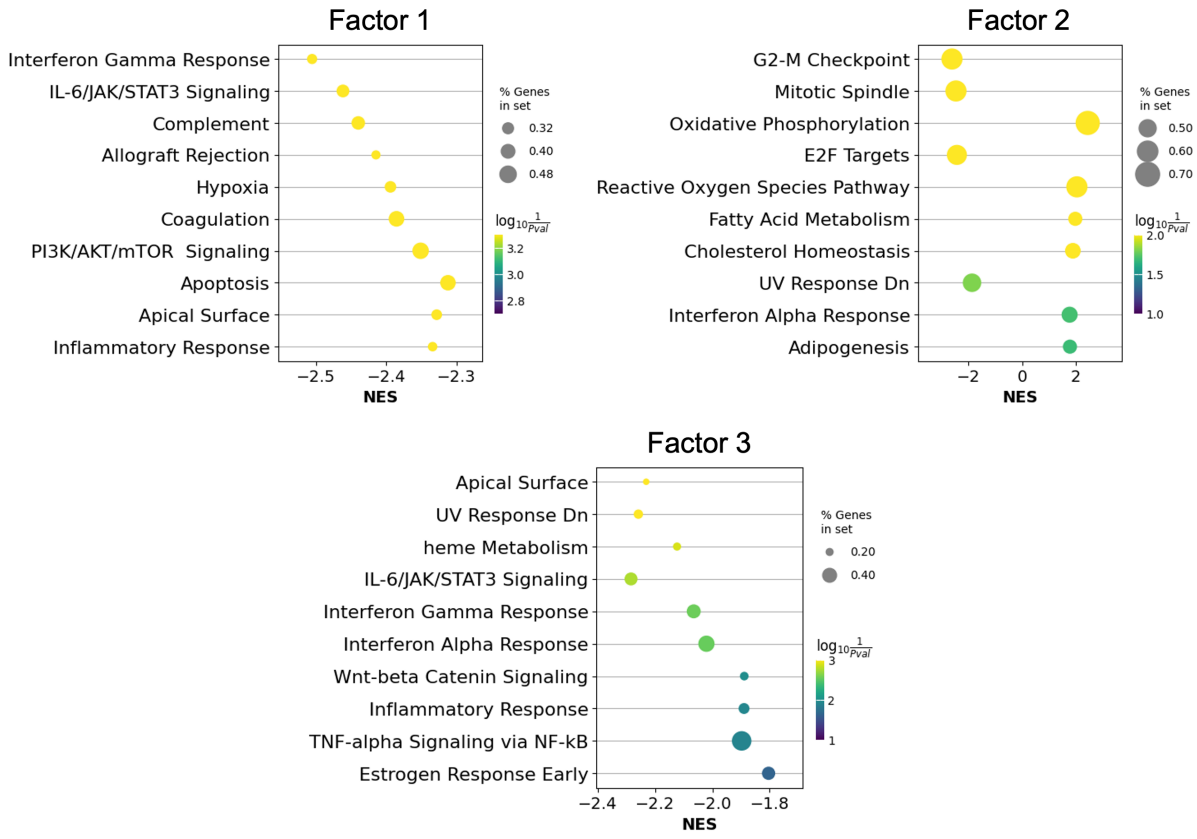

Figure 7. Top enriched pathways from gene set enrichment analysis (GSEA) performed for each GGeraPHF factor. Cells were assigned to Factor 0, 1, or 2 based on learned cell weights, and signal-to-noise ratio was used to rank genes from cells, using factor membership as the condition. Gene enrichment was identified using hallmark gene sets, and false discovery rate cutoff of 0.25 was used to select for enriched gene sets.