IDENTIFYING DONOR-ROBUST PERTURBATION TARGETS VIA SPARSE MANIFOLD CONTROL

Adithya V. Madduri
Harvard College
Department of Biomedical Informatics, Harvard Medical School
adithyamadduri@college.harvard.edu

Chirag J. Patel*
Department of Biomedical Informatics, Harvard Medical School
Chirag_Patel@hms.harvard.edu

ABSTRACT

A central challenge in applying machine learning to biological discovery is not only prediction, but also control: identifying small, interpretable sets of molecular targets whose perturbation can drive cells along meaningful biological trajectories. We introduce Sparse Linear Manifold Control (SLMC), a disease-aligned framework that isolates donor-robust axes of variation in single-cell transcriptomic data and casts target selection as a sparse reconstruction problem. Across diverse human datasets, the resulting objective exhibits strong diminishing returns, enabling simple greedy algorithms to recover compact, disease-relevant gene sets that generalize across donors. By exposing algorithmic structure in disease manifolds, SLMC bridges representation learning and experiment-ready perturbation design for interpretable and actionable biological discovery.

1 INTRODUCTION

A central challenge in applying machine learning to biological discovery is control: identifying small, interpretable sets of molecular targets whose perturbation can reliably move cells along biologically meaningful trajectories. In single-cell genomics, this challenge is amplified by donor heterogeneity, technical variation, and the combinatorial nature of genetic interventions. While modern representation learning methods can capture rich structure in high-dimensional transcriptomic data, translating these representations into actionable experimental hypotheses, specifically, deciding *which genes to perturb*, remains a major bottleneck.

Recent progress in generative and perturbation-aware modeling has enabled increasingly realistic simulation of cellular responses. Deep generative frameworks such as scGen and compositional perturbation autoencoders can predict transcriptome-wide effects of specified interventions (Lotfollahi et al., 2019; 2021), while regulatory network approaches aim to simulate perturbations through inferred gene interactions (Kamimoto et al., 2023). However, these methods generally assume that candidate perturbations are provided a priori. As a result, they do not directly address the upstream reasoning problem of identifying minimal, donor-robust gene sets that define disease-relevant directions of variation and are feasible for downstream experimentation.

In this work, we study sparse gene selection for disease-aligned representation learning in single-cell transcriptomics, motivated by the goal of identifying minimal perturbation targets that generalize across donors. We introduce *Sparse Linear Manifold Control* (SLMC), a framework that isolates donor-robust axes of disease-associated variation and formulates target selection as a sparse reconstruction problem. SLMC defines a scalar disease-aligned objective per cell by projecting transcriptomic data onto a donor-orthogonal direction in a structured PCA embedding, thereby reducing confounding effects while preserving disease-relevant signal.

*Corresponding author.

Our central contribution is an empirical analysis of the optimization geometry induced by this objective. Although optimal subset selection is NP-hard in general, we find that the SLMC reconstruction objective exhibits strong diminishing returns across diverse human single-cell datasets. Using randomized submodularity diagnostics, we observe consistently low violation rates, indicating that the objective is well-approximated by a submodular function in practice. This structure explains why simple greedy and forward–backward greedy algorithms recover compact gene sets with high fidelity, often outperforming continuous ℓ_1 relaxations under tight sparsity constraints. Across renal cell carcinoma and Alzheimer's disease datasets, SLMC-based selection identifies gene programs consisting of as few as 10–25 genes that reconstruct disease-aligned trajectories and generalize to held-out donors. By exposing algorithmic structure in disease-associated manifolds, this work provides a principled interface between representation learning and perturbation prioritization, supporting interpretable and experimentally actionable reasoning in single-cell genomics.

2 Problem Formulation and Disease-Aligned Objective

We address the problem of sparse gene selection in single-cell transcriptomics. Our objective is to identify a minimal set of genes S whose expression captures the transition from a diseased state toward a healthy reference, effectively providing a shortlist for in silico perturbation screening. Let $X \in \mathbb{R}^{n \times p}$ denote a standardized gene expression matrix with n cells and p genes. Each cell i is associated with a disease label $c_i \in \{\text{disease}, \text{healthy}\}$ and a donor/batch identifier d_i.

2.1 Disease Direction in a PCA Embedding

To capture global transcriptomic shifts rather than gene-level noise, we first project the data into a low-dimensional space:

$$Z = \mathrm{PCA}_d(X) \in \mathbb{R}^{n \times d}$$

Let $\mathcal{I}_D$ and $\mathcal{I}_H$ denote the indices of diseased and healthy cells. We define the raw disease direction v as the vector connecting the centroids of these two populations in PCA space:

$$v = \mu_D - \mu_H, \quad \text{where } \mu_D = \frac{1}{|\mathcal{I}_D|} \sum_{i \in \mathcal{I}_D} Z_i, \quad \mu_H = \frac{1}{|\mathcal{I}_H|} \sum_{i \in \mathcal{I}_H} Z_i$$

This vector v represents the primary axis of dysregulation. However, in multi-donor datasets, v is often confounded by technical batch effects or donor-specific biological variation.

2.2 Donor-Subspace Orthogonalization

To ensure our selection is clinically relevant and not driven by batch artifacts, we perform subspace pruning. We identify principal components (PCs) that represent "nuisance" variation. For each PC coordinate $Z_{\cdot k}$, we compute the fraction of variance explained (R^2) by donor identity and disease status. PCs are treated as donor-confounded if they meet the following criteria:

- High donor association: $R^2(\text{donor}) > \tau_{\text{donor}}$
- Low disease association: $R^2(\text{disease}) < \tau_{\text{disease}}$

Both R^2 quantities are computed directly from the data as the fraction of variance of each PC explained by the categorical donor or disease labels, so the identification of nuisance PCs is entirely data-driven rather than based on fixed component indices. Let $U \in \mathbb{R}^{d \times r}$ be the matrix whose columns are the r PCs satisfying these criteria. We define the donor-orthogonal disease direction $v_\perp$ by projecting v onto the orthogonal complement of the donor subspace:

$$v_\perp = (I - UU^\top)v$$

This operation selectively removes only those principal component directions that are empirically dominated by donor effects and weakly associated with disease, preserving disease-relevant variation while suppressing donor-specific structure. The thresholds τ_{donor} and τ_{disease} are fixed hyperparameters, but the set of removed components depends on the dataset-specific variance decomposition of each PC. Finally, we assign each cell a scalar disease-aligned target score $y_i = Z_i v_\perp$, representing the "disease-ness" of a cell after removing donor-specific confounders.

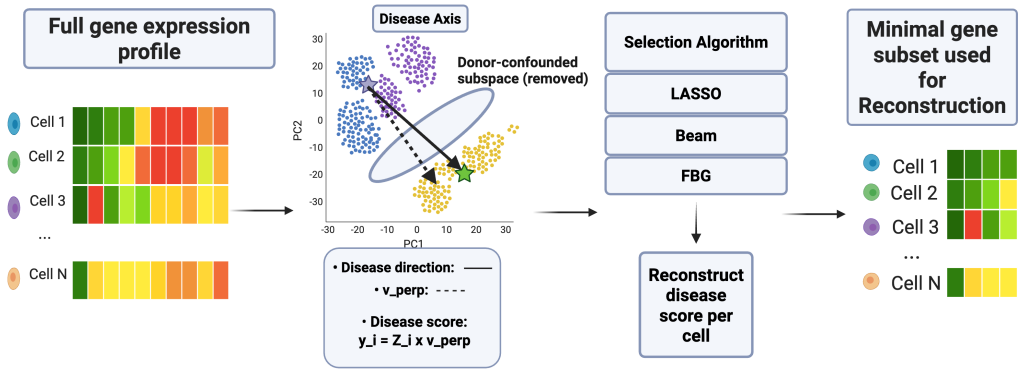

Figure 1: We define a donor-orthogonal disease axis in PCA space and project cells onto this axis to obtain a scalar disease score. Sparse gene selection methods are then used to identify small gene sets that reconstruct the disease score, yielding compact and interpretable representations of disease-aligned variation.

2.3 Sparse Reconstruction Objective

We formulate gene selection as a combinatorial optimization problem. For a subset of genes $S \subseteq \{1, \ldots, p\}$, we let X_S denote the corresponding columns of X. We evaluate S by its ability to reconstruct the target score y via ridge regression:

$$\hat{y}_S = X_S \beta, \quad \text{where } \beta = \arg\min_{\beta} \|y - X_S\beta\|_2^2 + \alpha\|\beta\|_2^2$$

Our objective function $f(S)$ is defined as the negative Mean Squared Error (MSE):

$$f(S) = -\frac{1}{n}\|y - \hat{y}_S\|_2^2$$

The goal is to find $\max_{|S| \leq k} f(S)$. While this is NP-hard, the efficiency of greedy approximations depends heavily on the submodularity of $f(S)$.

3 Related Works

3.1 Computational Modeling of Cellular Perturbations

Identifying genes that drive cellular state transitions is a foundational goal in systems biology and a key component of generative single-cell modeling. Deep generative frameworks such as scGen and Compositional Perturbation Autoencoders (CPA) can predict whole-transcriptome responses to specified or combinatorial perturbations, enabling in silico simulation of cell-state shifts across conditions (Lotfollahi et al., 2019; 2021). Gene regulatory network approaches such as CellOracle instead simulate the consequences of perturbing specific regulators based on inferred transcription factor–target interactions (Kamimoto et al., 2023). While these methods focus on accurately predicting full transcriptomic responses to specified perturbations, we instead aim to identify a small, donor-robust set of genes whose joint perturbation moves cells along a disease-aligned trajectory.

3.2 Foundations of Submodular Feature Selection

Selecting a subset of features to maximize a utility function is combinatorial and NP-hard in general, motivating the use of tractable approximation algorithms for subset selection (Das & Kempe, 2011). When the objective exhibits diminishing returns, greedy selection admits strong approximation guarantees characterized by the submodularity ratio and curvature, and least-squares–type objectives

satisfy a corresponding weak submodularity under standard restricted convexity and smoothness conditions that yields constant-factor guarantees for greedy feature selection even without exact submodularity (Das & Kempe, 2011; Bian et al., 2017; Elenberg et al., 2018). In genomics, submodular optimization has been successfully applied to choose small, diverse panels of assays under strict experimental budgets (Wei et al., 2016). We build on these foundations by showing empirically that the SLMC ridge reconstruction objective exhibits near-diminishing returns, helping explain the strong performance of simple greedy gene selection.

3.3 Single-Cell Trajectories and Disease Manifolds

Recent work constructs batch-corrected latent manifolds to enable comparisons of cellular states across donors and datasets (Danino et al., 2024). Complementary trajectory methods show that meaningful cross-donor progressions require explicitly modeling between-sample variability to avoid donor-specific artifacts (Hou et al., 2023). However, these approaches typically represent disease progression using dense latent coordinates rather than small, experimentally actionable gene sets. In this work, we instead ask whether a sparse subset of genes can reconstruct a donor-robust disease-aligned axis that generalizes to unseen donors.

4 Methods

4.1 Empirical Submodularity Diagnostics

A set function is submodular if it exhibits "diminishing returns": adding a gene to a small set provides more value than adding it to a larger set. To verify this property, we perform randomized triplet tests of the inequality:

$$\Delta(g \mid B) \geq \Delta(g \mid A) \quad \text{for } B \subset A \text{ and } g \notin A$$

where $\Delta(g \mid S) = f(S \cup \{g\}) - f(S)$ is the marginal gain of adding gene g.

External-Manifold Diagnostic: To establish a baseline for single-cell manifolds, we split genes into "target" and "feature" sets. We reconstruct the PCA manifold of the target genes using the feature genes. This determines if the ridge reconstruction objective is inherently submodular in transcriptomic data.

Disease-Aligned Diagnostic: We repeat the test using the specific target y (from Section 2.2), for the two datasets used in our experiment. We report the Violation Rate (frequency where $\Delta_B < \Delta_A$) and Violation Magnitude.

4.2 Datasets

We analyzed a total of five publicly available human single-cell datasets to assess the behavior of our external-manifold diagnostics across heterogeneous biological contexts. These datasets included: (i) GSM8652069, a healthy peripheral blood mononuclear cell (PBMC) single-cell multiomic dataset from human blood; (ii) GSE314072, a renal cell carcinoma (RCC) single-cell study profiling tumor-infiltrating T cells; (iii) GSE308624, a spatial single-cell transcriptomics dataset of H. pylori–associated gastric cancer (Chen et al., 2025); (iv) GSE227734, a single-cell cardiomyocyte dataset investigating hypertrophic cardiomyopathy (HCM) (Chen et al., 2024); and (v) GSE138852, a single-nucleus RNA-seq atlas of the human cortex in Alzheimer's disease (Grubman et al., 2019). We selected two of these datasets, GSE314072 (RCC single-cell RNA-seq) and GSE138852 (Alzheimer's disease single-cell RNA-seq) for analysis and validation using the experimental protocol outlined in 4.5. The RCC dataset (GSE314072) comprises 105,553 single cells, including 59,982 RCC cells and 45,571 healthy cells, while the Alzheimer's disease dataset (GSE138852) consists of 13,214 nuclei, including 6,673 AD and 6,541 control nuclei. We lastly, selected GSE314072 to analyze whether selected programs generalize to unseen donors.

4.3 Sparse Gene Selection Algorithms

Motivated by the empirical evidence of near-submodularity (violation rates $< 10\%$), we compare four sparse gene selection strategies:

- **Forward Greedy (OMP-style):** Iteratively adds the gene g that maximizes the marginal gain $\Delta(g \mid S) = f(S \cup \{g\}) - f(S)$. This procedure is analogous to orthogonal matching pursuit in linear models, applied here to ridge-based reconstruction of the disease-aligned target.
- **Forward–Backward Greedy (FBG):** Extends forward greedy selection with a backward pruning step. After each addition, previously selected genes are removed if their removal does not decrease the objective beyond a small tolerance, yielding more parsimonious sets when redundancy arises.
- **Beam Search:** Maintains a fixed-width set of candidate solutions at each sparsity level, expanding multiple hypotheses in parallel. This method explores potential non-greedy interactions and serves as a stress test for the greedy assumption.

4.4 LASSO BASELINE (CONTINUOUS ℓ_1 RELAXATION)

As a continuous sparsity baseline, we evaluate LASSO regression, which replaces the combinatorial subset constraint with an ℓ_1 penalty (Tibshirani, 1996). Specifically, we solve

$$\hat{\beta}(\alpha) = \arg\min_{\beta} \ \|y - X\beta\|_2^2 + \alpha\|\beta\|_1,$$

where X denotes the standardized gene expression matrix and y is the disease-aligned target. All genes are standardized to zero mean and unit variance prior to fitting.

We sweep the regularization parameter α over a logarithmically spaced grid of 50 values, ranging from the smallest value that yields a nonzero solution to a value that sets all coefficients to zero. Each α induces a model with a variable number of nonzero coefficients, corresponding to an implicit gene set size $k(\alpha)$.

To enable fair comparison with combinatorial methods at fixed gene budgets $k \in \{10, 25, 50, 100\}$, we select, for each target budget k, the LASSO model with $k(\alpha) \leq k$ that achieves the lowest reconstruction error. No refitting is performed after feature selection; coefficients are taken directly from the LASSO solution. Performance is evaluated using the same reconstruction metrics (MSE and R^2) as for greedy and beam search methods.

4.5 EXPERIMENTAL PROTOCOL

We evaluated all methods across five diverse single-cell datasets. For each dataset, we follow a consistent experimental protocol:

- **Sparsity sweep:** We evaluate gene set sizes $k \in \{10, 25, 50, 100\}$.
- **Robustness:** Each experiment is repeated across five random seeds, controlling randomness in PCA initialization, gene splits, and submodularity diagnostics.
- **Metrics:** We report mean squared error (MSE), which corresponds directly to the optimized objective, as well as the coefficient of determination (R^2).

4.6 DONOR-HELD-OUT GENERALIZATION PROTOCOL

To evaluate whether sparse gene sets selected under the SLMC objective generalize beyond donor-specific structure, we adopt a donor-held-out evaluation scheme. For each dataset and random seed, donors are randomly partitioned into disjoint training (70%) and test (30%) sets, and all cells from a given donor are assigned exclusively to one split.

All preprocessing steps that could induce information leakage—including gene standardization, PCA embedding, donor-subspace identification, and construction of the disease-aligned target—are performed using training donors only. The learned PCA transformation and donor-orthogonal disease direction are then applied to held-out donors to compute test targets.

Sparse gene selection is performed exclusively on the training split. Generalization performance is assessed by evaluating the ridge reconstruction error of the learned gene sets on held-out donors. This protocol ensures that reported test performance reflects true out-of-donor generalization rather than memorization of donor-specific effects.

5 RESULTS

5.1 NEAR-SUBMODULARITY OF SINGLE-CELL GENE RECONSTRUCTION OBJECTIVES

We first empirically evaluated whether the ridge reconstruction objective used for gene selection satisfies diminishing returns across diverse single-cell RNA-seq datasets. Across five datasets (GSE138852, GSE314072, GSE308624, GSM8652069, GSE227734), we conducted 2,000 randomized triplet tests per seed and five independent random seeds per dataset.

Across datasets, the mean violation rate ranged from 2.12% to 9.06%, with low variance across seeds (standard deviation $\leq$ 1.44%). Specifically, mean violation rates were 2.12% (GSE314072), 3.62% (GSE308624), 5.22% (GSE227734), 5.75% (GSE138852), and 9.06% (GSM8652069). In all datasets, the mean marginal difference $\Delta(g \mid S') - \Delta(g \mid S)$ was negative, indicating systematic diminishing returns. Overall, these results indicate that the reconstruction objective is approximately submodular, justifying greedy and forward-backward selection strategies.

Dataset	Mean violation rate (%)	Std (%)
GSE138852	5.75	0.60
GSE227734	5.22	0.40
GSE308624	3.62	0.87
GSE314072	2.12	0.48
GSM8652069	9.06	1.44

Table 1: Submodularity violation rates across datasets (mean $\pm$ std).

5.2 PREDICTIVE PERFORMANCE OF FBG VERSUS BASELINES

We compared Forward–Backward Greedy (FBG), pure Greedy, Beam Search, and Lasso on two representative datasets (GSE314072 and GSE138852) across gene budgets $k \in \{10, 25, 50, 100\}$ and five random seeds.

On the RCC dataset (GSE314072), FBG, Greedy, and Beam exhibited nearly identical performance across all budgets. At $k = 10$, all three methods achieved the same performance (MSE 268.866, R^2 0.859), while Lasso performed substantially worse (MSE 685.213, R^2 0.640). At $k = 50$, FBG achieved an MSE of 99.804 compared to 100.657 for Beam, with identical R^2 of 0.947. At $k = 100$, FBG achieved an MSE of 57.574 and R^2 of 0.970, marginally improving over Beam (MSE 59.404).

On AD dataset (GSE138852), we observed patterns highly consistent with the RCC dataset. At $k = 25$, FBG achieved a reconstruction MSE of 113.661 ($R^2 = 0.916$), outperforming Beam Search (MSE 115.588). This performance gap persisted at higher gene budgets; by $k = 100$, FBG reached an MSE of 43.105 ($R^2 = 0.968$), while Beam Search trailed at 48.813. Across all sparsity levels, the continuous LASSO baseline underperformed significantly, yielding MSE values 2–5$\times$ larger than combinatorial methods at low k.

Taken together, these results indicate that forward greedy selection is sufficient to attain near-optimal predictive performance under the reconstruction objective studied here. FBG and Beam search do not provide systematic improvements over pure Greedy, but closely track its performance, consistent with an objective that exhibits strong diminishing returns. In contrast, convex sparsity-based selection via Lasso performs substantially worse at small gene budgets.

5.3 MARGINAL GAIN DECAY AND DIMINISHING RETURNS DURING SELECTION

We analyzed the trajectory of marginal gains during FBG selection on GSE314072 at $k = 50$. The mean marginal gain at step 1 was 1.03×10^3, dropping to 2.68×10^2 by step 2 and 1.11×10^2 by step 3. By step 10, the marginal gain had decreased to 1.62×10^1.

When normalized by the first step, the relative marginal gain dropped to 0.26 at step 2, 0.11 at step 3, and 0.016 by step 10. At the midpoint (step 25), the relative gain was 0.004, and by the final step it was 0.001. Across 49 transitions, only 5 monotonicity violations were observed, confirming near-monotonic decay. This quantitative decay strongly supports the presence of diminishing returns consistent with approximate submodularity.

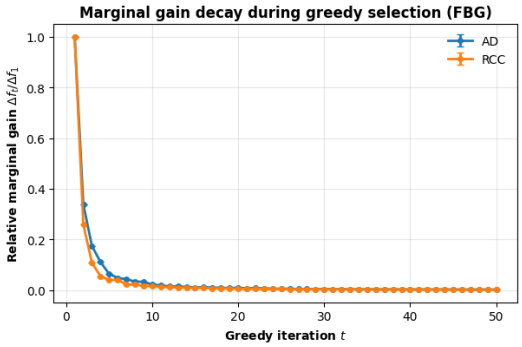

Figure 2: Marginal gain decay during forward–backward greedy (FBG) selection at $k = 50$. Relative marginal gains decrease rapidly across greedy iterations, with near-monotonic decay across seeds and datasets, indicating strong diminishing returns consistent with approximate submodularity.

5.4 Out-of-donor generalization testing

We evaluated out-of-donor generalization of sparse gene sets selected under the SLMC objective using donor-held-out evaluation on two biologically distinct datasets: RCC (GSE314072) and Alzheimer's disease (AD; GSE138852). Across five random donor splits and gene budgets $k \in \{10, 25, 50, 100\}$, test performance increased monotonically with k for all methods, indicating that larger gene sets consistently improve reconstruction of the disease-aligned axis under donor hold-out. In RCC, Forward–Backward Greedy (FBG) closely matched beam search across all sparsity levels, while substantially outperforming LASSO under tight sparsity. In AD, overall generalization was weaker and variance across donor splits was higher, reflecting greater biological heterogeneity; nevertheless, FBG again outperformed LASSO at all budgets and consistently exceeded beam search at small and intermediate k, with all methods converging at larger k. These results show that greedy optimization under the SLMC objective yields sparse gene sets that generalize robustly across donors while preserving the geometry of the disease-aligned trajectory.

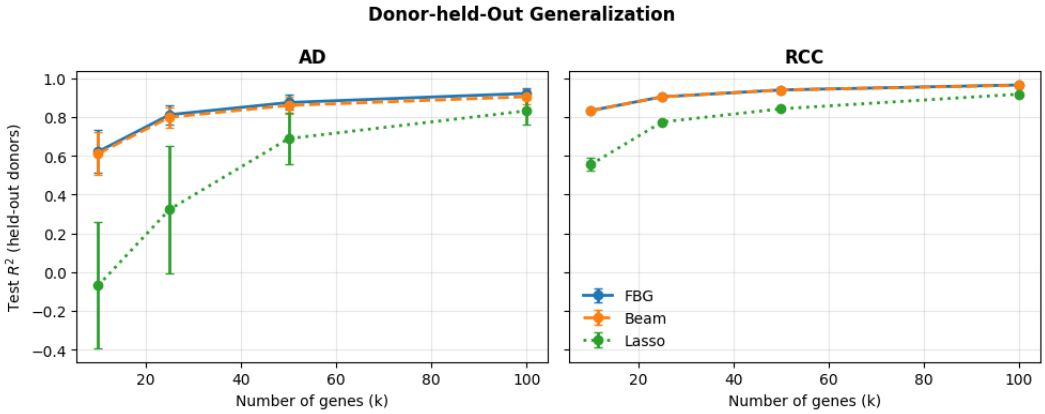

Figure 3: Out-of-donor generalization under the SLMC objective. Test R^2 on held-out donors as a function of gene budget k for FBG, Beam Search, and LASSO, averaged across five donor splits.

6 Discussion

Across diverse single-cell datasets, we find that the SLMC reconstruction objective, while combinatorial and NP-hard in principle, exhibits strong diminishing returns in practice. Low submodularity violation rates, near-monotonic decay of marginal gains, and the close alignment between greedy,

forward–backward greedy, and beam search performance all point to a common conclusion: disease-aligned gene selection under SLMC is governed by a highly regular optimization geometry, in which simple greedy procedures are not merely efficient approximations, but empirically reliable methods for recovering compact gene programs.

Crucially, this behavior arises from defining reconstruction with respect to a donor-orthogonal disease axis rather than from generic properties of high-dimensional feature selection. By suppressing donor-driven variation and emphasizing disease-associated structure, SLMC yields objectives whose informative signal is concentrated in early selections and rapidly saturates thereafter. This explains why small gene sets, often 10–25 genes, are sufficient to reconstruct disease-aligned trajectories, and why performance generalizes robustly to unseen donors.

The consistency between optimization diagnostics, predictive performance, and biological validation further supports this interpretation. Greedy-selected gene sets recover canonical disease-associated genes, achieve GWAS overlap comparable to more expensive search procedures, and preserve reconstruction fidelity under donor hold-out, despite mild deviations from exact submodularity. Together, these findings indicate that approximate submodularity is an empirically meaningful property of disease-aligned single-cell manifolds that can be exploited to obtain interpretable and experimentally actionable gene programs.

6.1 Limitations

Our work has several limitations. First, SLMC relies on a linear reconstruction objective in a PCA embedding and therefore cannot capture nonlinear gene interactions or branching disease trajectories. In datasets where such structure is essential, this limitation would manifest as poor reconstruction and early performance saturation, signaling that SLMC's assumptions are inadequate. Second, while PCA effectively denoises and captures global variation, it may discard biologically relevant signal residing in lower-variance components. Donor-subspace orthogonalization mitigates dominant confounding effects but does not guarantee preservation of all disease-relevant variation. Third, GWAS overlap is used here as descriptive biological context rather than inferential validation; overlap should not be interpreted as evidence of causality, and lack of overlap—particularly in under-characterized diseases—does not imply biological irrelevance. Finally, cell-type–specific stratification may be important in some diseases, however we opted against it since doing so would substantially reduce effective sample sizes and stability under tight sparsity constraints.

6.2 Future Directions

Several extensions follow naturally from this work. Methodologically, incorporating nonlinear or kernelized reconstruction objectives could capture higher-order interactions while retaining explicit sparsity constraints, provided donor robustness and tractable optimization can be preserved. Alternative embeddings that better retain rare or nonlinear variation may further improve performance in heterogeneous diseases. Biologically, applying SLMC within well-powered cell types, or to longitudinal and perturbative datasets, could clarify how sparse gene programs evolve along disease progression rather than merely separating disease states. Finally, coupling SLMC-selected gene sets with downstream perturbation models or experimental screening frameworks would enable direct evaluation of whether disease-aligned reconstruction fidelity translates into effective control of cellular state.

7 Conclusion

We presented Sparse Linear Manifold Control as a framework for identifying compact, donor-robust gene programs that reconstruct disease-aligned variation in single-cell transcriptomic data. Empirically, the resulting objective exhibits strong diminishing returns across diverse datasets, explaining why simple greedy selection reliably recovers high-fidelity gene sets under tight sparsity constraints. By exposing geometric structure in disease-associated manifolds, SLMC provides a principled and interpretable interface between representation learning and target prioritization. More broadly, this work illustrates how understanding optimization geometry can render otherwise intractable selection problems practically useful for biological discovery.

REFERENCES

Céline Bellenguez, Fahri Küçükali, et al. New insights into the genetic etiology of alzheimer's disease and related dementias. *Nat. Genet.*, 54(4):412–436, April 2022.

Andrew An Bian, Joachim M Buhmann, Andreas Krause, and Sebastian Tschiatschek. Guarantees for greedy maximization of non-submodular functions with applications. *arXiv [cs.DM]*, March 2017.

Bonan Chen, Hongzhen Tang, Xiaohong Zheng, Fuda Xie, Peiyao Yu, Yang Lyu, Tiejun Feng, Jialin Wu, Jingya Liu, Yi Xu, Alvin H K Cheung, Canbin Fang, Zhangding Wang, Shouyu Wang, Justin Chak Ting Cheung, Yujuan Dong, Ruoxi Tian, Yigan Zhang, Cheng Lu, Chi Chun Wong, Jun Yu, William K K Wu, Elke Burgermeister, Man Tong, Fengbin Zhang, Wei Kang, Kam Tong Leung, and Ka Fai To. Spatial and functional dissection of cancer-associated fibroblasts-mediated immune modulation in H. pylori-associated gastric cancer. *Mol. Cancer*, 24(1):282, November 2025.

Shi Chen, Kui Wang, Jingyu Wang, Xiao Chen, Menghao Tao, Dan Shan, Xiumeng Hua, Shengshou Hu, and Jiangping Song. Profiling cardiomyocytes at single cell resolution reveals COX7B could be a potential target for attenuating heart failure in cardiac hypertrophy. *J. Mol. Cell. Cardiol.*, 186:45–56, January 2024.

Zhiping Chen and Lanfeng Wang. The clinical significance of UBE2C gene in progression of renal cell carcinoma. *Eur. J. Histochem.*, 65(2):3196, March 2021.

Reut Danino, Iftach Nachman, and Roded Sharan. Batch correction of single-cell sequencing data via an autoencoder architecture. *Bioinform. Adv.*, 4(1):vbad186, 2024.

Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *arXiv [stat.ML]*, February 2011.

Ethan R Elenberg, Rajiv Khanna, Alexandros G Dimakis, and Sahand Negahban. Restricted strong convexity implies weak submodularity. *aos*, 46(6B):3539–3568, December 2018.

Francesca Fernandez-Enright and Jessica L Andrews. Lingo-1: a novel target in therapy for alzheimer's disease? *Neural Regen. Res.*, 11(1):88–89, January 2016.

Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K Simmons, Sam Buckberry, Dulce B Vargas-Landin, Daniel Poppe, Jahnvi Pflueger, Ryan Lister, Owen J L Rackham, Enrico Petretto, and Jose M Polo. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.*, 22(12):2087–2097, December 2019.

Wenpin Hou, Zhicheng Ji, Zeyu Chen, E John Wherry, Stephanie C Hicks, and Hongkai Ji. A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nat. Commun.*, 14(1):7286, November 2023.

Rosemary J Jackson, Bradley T Hyman, and Alberto Serrano-Pozo. Multifaceted roles of APOE in alzheimer disease. *Nat. Rev. Neurol.*, 20(8):457–474, August 2024.

Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, February 2023.

Rajiv Khanna, Ethan Elenberg, Alexandros G Dimakis, Sahand Negahban, and Joydeep Ghosh. Scalable greedy feature selection via weak submodularity. *arXiv [stat.ML]*, March 2017.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scGen predicts single-cell perturbation responses. *Nat. Methods*, 16(8):715–721, August 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, pp. 2021.04.14.439903, April 2021.

Alexandru Nesiu, Anca Maria Cimpean, Raluca Amalia Ceausu, Ahmed Adile, Ioan Ioiart, Camillo Porta, Michele Mazzanti, Tommaso Ciro Camerota, and Marius Raica. Intracellular chloride ion channel protein-1 expression in clear cell renal cell carcinoma. *Cancer Genomics Proteomics*, 16 (4):299–307, July 2019.

Xiangyu Qiu, Zhaoan Yu, Xiaoqing Lu, Xin Jin, Jinrong Zhu, and Rongxin Zhang. PD-1 and LAG-3 dual blockade: emerging mechanisms and potential therapeutic prospects in cancer. *Cancer Biol. Med.*, 21(11):970, December 2024.

Jiamou Ren, Shuli Zhang, Xiaoling Wang, Yuxin Deng, Yi Zhao, Yan Xiao, Jian Liu, Liangzhao Chu, and Xiaolan Qi. MEF2C ameliorates learning, memory, and molecular pathological changes in alzheimer's disease in vivo and in vitro. *Acta Biochim. Biophys. Sin. (Shanghai)*, 54(1):77–90, January 2022.

Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, Arwa Ibrahim, Yue Ji, Sajo John, Elizabeth Lewis, Jacqueline A L MacArthur, Aoife McMahon, David Osumi-Sutherland, Kalliope Panoutsopoulou, Zoë Pendlington, Santhi Ramachandran, Ray Stefancsik, Jonathan Stewart, Patricia Whetzel, Robert Wilson, Lucia Hindorff, Fiona Cunningham, Samuel A Lambert, Michael Inouye, Helen Parkinson, and Laura W Harris. The NHGRI-EBI GWAS catalog: knowledgebase and deposition resource. *Nucleic Acids Res.*, 51(D1):D977–D985, January 2023.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288, January 1996.

Kai Wei, Maxwell W Libbrecht, Jeffrey A Bilmes, and William Stafford Noble. Choosing panels of genomics assays using submodular optimization. *Genome Biol.*, 17(1):229, November 2016.

A Appendix

A.1 Code and Data Availability

All code used in this study is available at: `https://github.com/AdiVM/SLMC_single-cell`. The single-cell data used in this study are publicly available human transcriptomic datasets generated by prior studies and accessible through the Gene Expression Omnibus (GEO). Analyses were performed using renal cell carcinoma single-cell RNA-seq data (GEO accession: `GSE314072`) and Alzheimer's disease single-nucleus RNA-seq data from human cortex (GEO accession: `GSE138852`), with additional publicly available datasets used for cross-context diagnostic evaluation (GEO accessions: `GSM8652069`, `GSE308624`, and `GSE227734`). All datasets contain de-identified human samples and are available under standard public-use terms via GEO.

A.2 Theoretical Justification of Near Submodularity

Greedy subset selection for linear reconstruction objectives is not submodular in general; however, such objectives are known to satisfy weak submodularity under standard restricted eigenvalue conditions on the design matrix (Das & Kempe, 2011; Elenberg et al., 2018). Weak submodularity is quantified via the submodularity ratio, which measures the extent to which a set function exhibits diminishing returns. For objectives arising from (regularized) least-squares regression, prior work shows that the submodularity ratio is lower bounded by quantities depending on the restricted strong convexity (RSC) and restricted strong smoothness (RSS) constants of the objective over sparse supports (Elenberg et al., 2018; Khanna et al., 2017). Concretely, if the underlying ridge regression loss is m-restricted strongly convex and M-restricted strongly smooth in the coefficients on supports of size at most k, then the induced set function $f(S)$ has submodularity ratio bounded below by m/M (Elenberg et al., 2018; Khanna et al., 2017). These conditions are strictly weaker than exact submodularity and hold for a broad class of design matrices encountered in practice, provided that the covariance structure is well-conditioned on sparse subsets.

In SLMC, the ridge-regularized least-squares loss

$$\ell_S(\beta) = \|y - X_S\beta\|_2^2 + \alpha\|\beta\|_2^2$$

satisfies the RSC/RSS conditions whenever the regularized Gram matrix $X_S^\top X_S + \alpha I$ is uniformly well-conditioned over sparse supports (Elenberg et al., 2018; Khanna et al., 2017). Under these conditions, the associated set function

$$f(S) = y^\top X_S (X_S^\top X_S + \alpha I)^{-1} X_S^\top y$$

is weakly submodular (Elenberg et al., 2018; Khanna et al., 2017). Ridge regularization therefore stabilizes the spectral properties governing the submodularity ratio by preventing arbitrarily small or large marginal gains induced by collinearity. Importantly, this theory does not imply that the SLMC objective is exactly submodular, nor does it guarantee that the submodularity ratio is close to one in the worst case. Rather, it establishes that greedy algorithms admit constant-factor approximation guarantees of the form $f(S_{\text{greedy}}) \geq (1 - e^{-\gamma}) f(S^\star)$, where γ is the submodularity ratio (Das & Kempe, 2011; Bian et al., 2017). Our empirical results in Section 5 demonstrate that, for disease-aligned single-cell manifolds, observed violation rates and magnitudes are small, indicating that the effective submodularity ratio is high in practice—consistent with, but not implied by, the above theoretical bounds.

A.3 Complexity Analysis of Selection Algorithms

We compare the computational cost of SLMC-FBG to combinatorial search baselines. Let n denote the number of cells, p the number of candidate genes (capped at 5000 in our implementation), k the target sparsity level, and B the beam width.

- **Forward Greedy / FBG:** At each of the k selection steps, the algorithm evaluates all remaining p genes, resulting in $O(k \cdot p)$ ridge regression fits. Each ridge fit on a subset of size $s \leq k$ has cost $O(ns^2)$ (ignoring lower-order terms), yielding an overall complexity of $O(n \cdot k^3 \cdot p)$ in the worst case. Although FBG includes a backward pruning step, we empirically observe that backward removals are rarely triggered, making its runtime effectively identical to pure forward greedy selection in practice.
- **Beam Search:** Unrestricted beam search would explore $O(k \cdot p \cdot B)$ candidates. To ensure tractability, our implementation restricts expansion to a candidate pool of size $M \ll p$ based on initial correlation with the target, reducing the cost to $O(k \cdot M \cdot B)$ ridge fits. Despite this heuristic restriction, FBG—which evaluates the full feature space at each step—matches or exceeds Beam Search performance under the SLMC objective, indicating that greedy trajectories are sufficient in this near-submodular regime.
- **LASSO (continuous ℓ_1 relaxation):** Our LASSO baseline performs a sweep over A regularization values (with $A = 50$ in our experiments), fitting a convex ℓ_1-penalized regression model at each value. Each fit has cost approximately $O(n \cdot p)$ per coordinate descent pass, yielding a total complexity of $O(A \cdot n \cdot p)$ for the sweep. While individual LASSO fits are computationally efficient, the method does not directly optimize for a fixed sparsity level k and requires post-hoc model selection to match target budgets, in contrast to the explicit k-controlled selection used by greedy and beam-based methods.

A.4 Out-of-donor generalization across datasets

We first evaluated whether sparse gene sets selected under the SLMC objective generalize to unseen donors using the RCC dataset (GSE314072). Across five random seeds and gene budgets $k \in \{10, 25, 50, 100\}$, combinatorial methods exhibited strong out-of-donor generalization, with test R^2 increasing monotonically as a function of k. Forward–Backward Greedy (FBG) consistently matched beam search performance on held-out donors. Averaged across seeds, FBG achieved test R^2 = 0.833 at k=10, 0.904 at k=25, 0.940 at k=50, and 0.966 at k=100, closely tracking beam search at all sparsity levels. In contrast, LASSO-based selection generalized substantially worse under tight sparsity constraints. At k=10, LASSO achieved mean test R^2 = 0.55, compared to 0.83 for FBG. Although LASSO performance improved with increasing sparsity, it remained consistently below combinatorial methods across all budgets. This gap was most pronounced at small k, indicating that continuous ℓ_1 relaxation underfits when the goal is to identify minimal gene sets that generalize across donors. Notably, the parity between FBG and beam search persisted under donor-held-out evaluation, despite the presence of mild submodularity violations measured on training data. This suggests that the near-submodular structure of the SLMC reconstruction objective reflects intrinsic

geometry of the disease-aligned manifold rather than donor-specific artifacts. As a result, simple greedy selection is sufficient to identify gene sets that generalize across donors, without requiring more expensive combinatorial search. Together, these results demonstrate that the geometric structure underlying the SLMC objective not only supports efficient greedy optimization, but also yields sparse gene sets that generalize robustly across donors.

We next evaluated out-of-donor generalization of sparse gene sets under the SLMC objective using the Alzheimer's disease dataset (GSE138852). As in the RCC setting, performance was assessed across five random donor-held-out splits and gene budgets $k \in \{10, 25, 50, 100\}$. In contrast to RCC, overall generalization performance was lower and exhibited substantially higher variability across donor splits, reflecting the greater heterogeneity of AD samples.

For the combinatorial methods, test performance increased monotonically with increasing gene budget. Forward–Backward Greedy (FBG) achieved mean test $R^2 = 0.56$ at $k = 10$, improving to 0.79 at $k = 25$, 0.85 at $k = 50$, and 0.91 at $k = 100$. Beam search followed a similar trend but consistently underperformed FBG at all sparsity levels, with mean test $R^2 = 0.55$ at $k = 10$, 0.77 at $k = 25$, 0.84 at $k = 50$, and 0.88 at $k = 100$. Notably, both methods exhibited large standard deviations at small k, indicating sensitivity to donor composition under severe sparsity.

LASSO-based selection generalized poorly under tight sparsity constraints in the AD setting. At $k = 10$, LASSO achieved negative mean test R^2 (-0.25), indicating systematic underfitting on held-out donors. Performance improved with increasing sparsity, reaching mean test $R^2 = 0.13$ at $k = 25$, 0.61 at $k = 50$, and 0.79 at $k = 100$, but remained substantially below combinatorial methods across all budgets. This pronounced gap at low k highlights the difficulty of recovering donor-robust signal in AD using continuous ℓ_1 relaxation when constrained to minimal gene sets.

As in RCC, FBG required no backward deletions across all runs, while beam search exhibited no gain monotonicity violations by construction. Despite the presence of mild submodularity violations in the FBG forward gains (185 total across seeds and budgets), greedy selection consistently outperformed or matched beam search in terms of test performance. Together, these results indicate that although AD presents a more challenging and heterogeneous generalization setting than RCC, the SLMC objective retains sufficient near-submodular structure for simple greedy optimization to recover sparse gene sets that generalize across donors, particularly at moderate sparsity levels.

A.5 Held-out Donor Reconstruction Fidelity

To complement the aggregate donor-held-out generalization metrics reported in Section 5.4, we provide a qualitative and distribution-level assessment of reconstruction fidelity on unseen donors. Specifically, we visualize the true donor-orthogonal disease scores y_{test} against their sparse reconstructions $\hat{y}_{\text{test}}$, pooling cells across five independent random donor-held-out splits (seeds) and multiple sparsity levels $k \in \{10, 25, 50, 100\}$. All quantities are computed using PCA embeddings, donor-subspace identification, and disease-aligned directions learned exclusively on training donors, and applied without refitting to held-out donors. Note that, unlike the split-averaged R^2 reported above, reconstruction fidelity here is computed by pooling held-out cells across donor splits, yielding a single global R^2 that reflects cell-level alignment rather than split-level variability.

Across both datasets, these visualizations confirm that high test R^2 reflects faithful recovery of the disease-aligned axis itself—preserving both ordering and scale—rather than trivial rescaling or variance shrinkage. In the Alzheimer's disease dataset, reconstruction fidelity increases monotonically with sparsity, achieving pooled held-out R^2 values of 0.81, 0.91, 0.94, and 0.96 at $k = 10, 25, 50$, and 100, respectively, while maintaining strong disease discrimination (AUROC$(y_{\text{test}}) = 0.93$; AUROC$(\hat{y}_{\text{test}}) = 0.86$–$0.93$). A similar pattern is observed in renal cell carcinoma, where held-out R^2 increases from 0.85 at $k = 10$ to 0.97 at $k = 100$, with AUROC preserved between the true and reconstructed scores (AUROC$(y_{\text{test}}) = 0.91$; AUROC$(\hat{y}_{\text{test}}) = 0.86$–$0.90$).

Together, these results demonstrate that sparse gene sets selected under the SLMC objective not only generalize across donors in aggregate metrics, but also reconstruct the same donor-orthogonal disease trajectory at the level of individual cells, supporting the interpretability and validity of the LOPO generalization analysis.

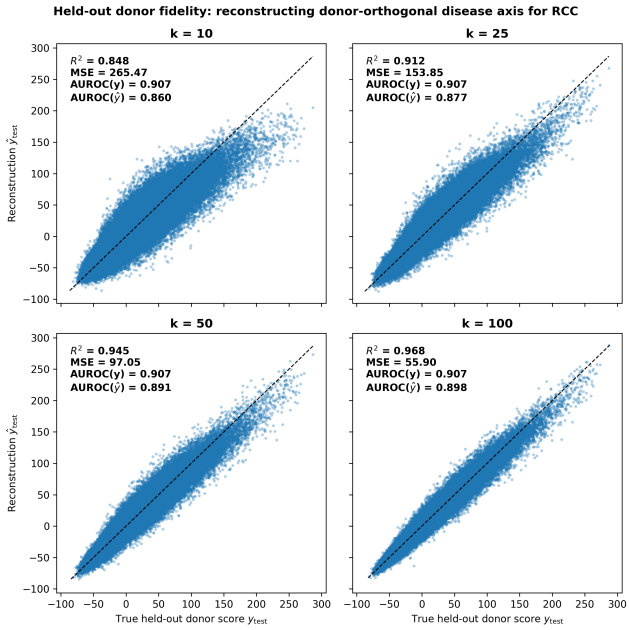

Figure 4: Held-out donor reconstruction fidelity for renal cell carcinoma (GSE314072). True donor-orthogonal disease scores y_{test} are plotted against sparse reconstructions $\hat{y}_{\text{test}}$ for $k \in \{10, 25, 50, 100\}$, pooled across five donor-held-out splits.

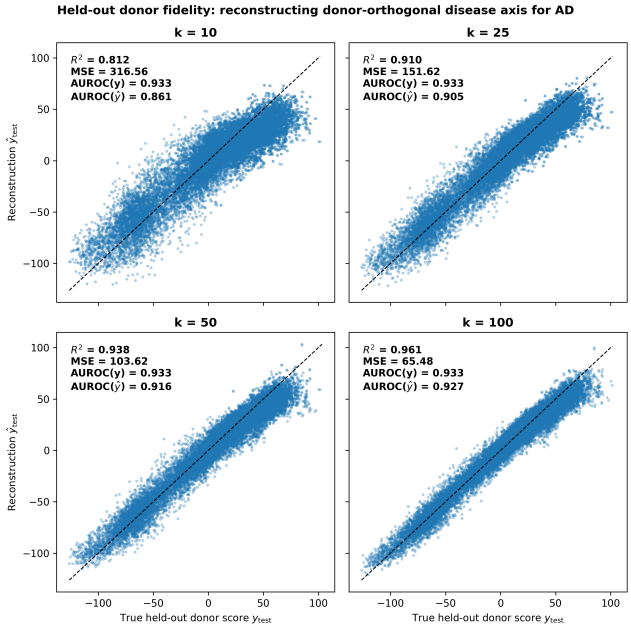

Figure 5: Held-out donor reconstruction fidelity for Alzheimer's disease (GSE138852). Sparse reconstructions preserve the ordering and scale of the donor-orthogonal disease axis across gene budgets and donor splits, despite increased biological heterogeneity.

A.6 Biological Validation via GWAS and Literature Overlap

To assess the biological relevance of the identified gene programs, we evaluated the overlap between genes selected across all five random donor-held-out splits ($k = 50$) and curated sets of

disease-associated genes from Genome-Wide Association Studies (GWAS) (Sollis et al., 2023). In the Alzheimer's Disease (AD) context, Forward-Backward Greedy (FBG) identified 26 GWAS-associated genes, achieving parity with the more computationally intensive Beam Search (26 genes) and outperforming LASSO (23 genes). Notably, FBG consistently prioritized "gold standard" AD risk genes, including **APOE**, **MEF2C**, **NRXN1**, and **RBFOX1** (Ren et al., 2022; Bellenguez et al., 2022). High-confidence drivers such as **LINGO1** and **APOE** exhibited higher selection frequencies across donor-held-out splits in the FBG model compared to both Beam Search and LASSO (Fernandez-Enright & Andrews, 2016; Jackson et al., 2024). This demonstrates that the approximately submodular geometry of the SLMC objective allows greedy trajectories to capture the core genetic architecture of the disease manifold as effectively as complex procedures, while remaining invariant to inter-donor variation.

In the Renal Cell Carcinoma (RCC) context, while the current GWAS catalog remains relatively underdeveloped for this disease—yielding zero overlaps for all models at $k = 50$—the gene programs identified by FBG are highly enriched for emerging therapeutic targets and biomarkers. Specifically, FBG identified **LAG3**, a primary immune checkpoint protein currently the subject of Phase II clinical trials for dual blockade with PD-1 in RCC patients (Qiu et al., 2024). Other identified genes include **TNFRSF9**, a co-stimulatory receptor on CD8+ T cells, and **UBE2C**, a member of the ubiquitin modification system recently shown to promote tumor cell proliferation and predict poor overall survival in RCC (Chen & Wang, 2021). Additionally, FBG prioritized **CLIC1**, which has been correlated with tumor grade, metastasis, and poor prognosis in clear-cell RCC (Nesiu et al., 2019). The robust identification of these clinically significant candidates under strict donor-held-out generalization testing underscores the utility of the SLMC framework for prioritizing interpretable gene sets that reflect actionable biological trajectories.

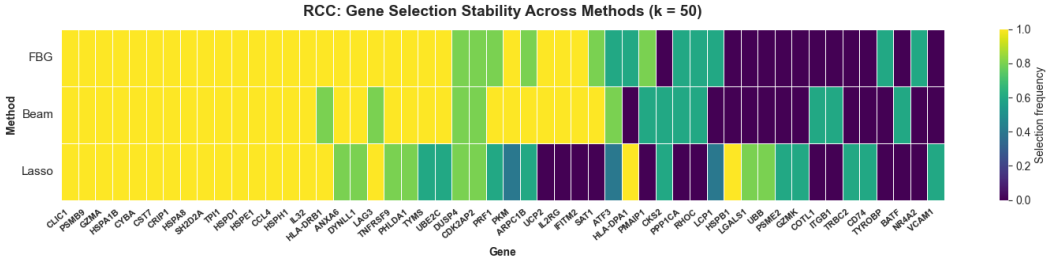

Figure 6: Gene selection stability across methods in RCC at $k = 50$. Each column corresponds to a gene, and color intensity indicates selection frequency across five donor-held-out splits.

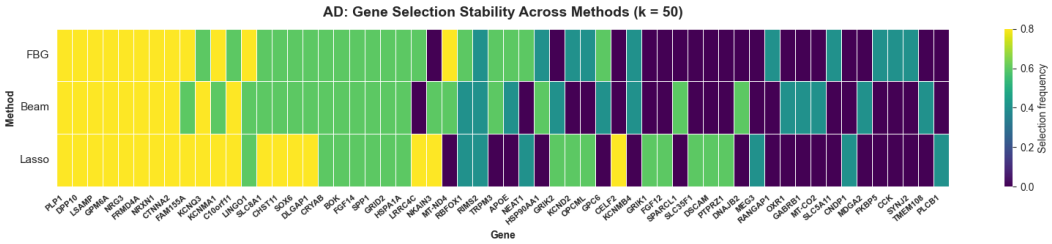

Figure 7: Gene selection stability across methods in Alzheimer's disease at $k = 50$. Each column corresponds to a gene, and color intensity indicates selection frequency across five donor-held-out splits.

A.7 Selection Stability via Jaccard Similarity

We evaluated selection stability using pairwise Jaccard similarity of the full selected gene sets across five random seeds under the out-of-donor protocol.

For the RCC dataset (GSE314072), stability was moderate and consistent across sparsity levels. At $k = 10$, the mean Jaccard similarity was 0.523 (std 0.118, range 0.333–0.667). Increasing the budget to $k = 50$ yielded a similar but slightly more concentrated overlap, with mean Jaccard similarity 0.545 (std 0.055, range 0.471–0.639).

In contrast, the AD dataset (GSE138852) exhibited lower and more variable stability. At $k = 10$, the mean Jaccard similarity was 0.382 (std 0.235, range 0.250–1.000), reflecting strong sensitivity of small gene sets to donor splits and initialization. At $k = 50$, the mean Jaccard similarity was 0.378 (std 0.194, range 0.250–0.887), again indicating limited but highly variable overlap between runs. The presence of occasional very high overlaps (e.g., Jaccard ≈ 0.89 or 1.00 for specific seed pairs) alongside many low-overlap pairs suggests that multiple distinct but similarly predictive gene programs can be recovered in the AD setting.

Overall, these results show that FBG selection is reproducible but not identical across random initializations. Stability is higher and more uniform in RCC, whereas AD exhibits greater heterogeneity, with several sparse gene sets capable of reconstructing the disease-aligned axis.

A.8 Optimization Diagnostics on Full-Dataset Fits

In this section we analyze the optimization behavior of FBG when fit on the full dataset without out-of-donor holdout. Unlike the main experiments, which evaluate generalization across unseen donors, these diagnostics are designed to probe intrinsic properties of the reconstruction objective and greedy search dynamics (e.g., stability, effective submodularity, and redundancy along the selection path). Evaluating on the full dataset removes cross-donor sampling noise and allows us to directly study how the objective behaves as genes are added or removed.

A.8.1 Role of backward steps in FBG

To assess the contribution of backward elimination in Forward–Backward Greedy (FBG), we explicitly tracked backward activity at both the step level and the run level. Across all random seeds and gene budgets ($k \in \{10, 25, 50, 100\}$), no forward step triggered a backward removal, and no run exhibited any backward elimination events. As a result, the total number of genes removed per run was exactly zero in all experiments. Rather than indicating a failure of the backward mechanism, this behavior provides empirical evidence that the reconstruction objective behaves as effectively submodular in the regimes considered. In particular, once a gene is selected by the forward step, its marginal contribution is never outweighed by that of a later-added gene, making backward correction unnecessary. This observation is consistent with the low submodularity violation rates and small violation magnitudes observed in our empirical submodularity diagnostics.

Together, these results indicate that the greedy forward path is locally stable under the objective studied here: selections made by forward greedy are never reversed by backward pruning. Consequently, FBG closely matches pure forward greedy selection across both the RCC and AD datasets, not because backward steps are disabled, but because the objective itself renders them redundant.

A.8.2 Robustness to submodularity violations

We stratified runs by median violation rate and compared FBG and Beam at $k = 50$. In the low-violation regime, FBG achieved mean MSE 99.72 compared to 100.60 for Beam. In the high-violation regime, FBG achieved mean MSE 99.93 compared to 100.74 for Beam. This trend was even more pronounced in the AD dataset, where FBG maintained a $\sim$2.3 MSE lead over Beam search regardless of violation frequency. The performance gap remained consistent across regimes, indicating that FBG's performance is robust to moderate deviations from exact submodularity.

A.8.3 Gene-level contribution structure and redundancy

We examined gene-level contributions along the FBG trajectory. Early selections yielded large marginal gains with moderate redundancy: mean relative gain over steps 1–10 was 0.158, with mean maximum correlation to previously selected genes of 0.390. In contrast, late selections (steps 41–50) exhibited mean relative gain of 0.002 and lower redundancy (mean correlation 0.316). Similar decay trajectories were observed in the AD context.

This indicates that early selections capture most of the predictive signal, while later additions contribute marginal refinements without excessive redundancy.