Sample Report

CellOracle™ Differentiation Optimization Analysis

CAPYBIO™

Table of Contents

Executive Summary

Project Details

- Project ID: CAPY-2026-003
- Date: March 2026
- Sample(s): Three samples (Day 0, Day 7, and FACS-sorted Day 14) of direct reprogramming of mouse cardiac fibroblasts to induced cardiomyocytes.
- Cells Analyzed: 10,606
- Number of clusters: 11
- Key cell types: Fibroblasts, Cardiomyocytes, Off-target
- Target cell type for in silico simulation: Cardiomyocytes
- Base GRN used: CapyBio Human GRN Database

Primary Conclusion

CellOracle GRN analysis of iCM (induced Cardiomyocyte) reprogramming between Day 0 and Day 14 reveals substantial rewiring of transcription factor (TF) regulatory programs across cell states. Egr1 and Mef2c emerge as the most central cardiomyocyte regulators. *In silico* perturbation simulations identify Ets1, Egr1, Foxs1, and Maff as top anti-cardiac candidate TFs and Klf5 and Mef2c as top pro-cardiac candidate TFs. All predictions are computational and should be validated experimentally.

Key Results

- Egr1 and Mef2c show the highest centrality in the cardiomyocyte network across both degree and eigenvector metrics, consistent with central roles in the cardiac regulatory hierarchy.
- GRN patterns indicate that the off-target population occupies an intermediate regulatory space, having partially exited the fibroblast identity without fully committing to a cardiac program.
- Knockout of Ets1, Egr1, Foxs1, and Maff produces the strongest predicted pro-cardiomyocyte shifts; Mef2c knockout produces the strongest anti-cardiomyocyte effect.
- Klf5 and Mef2c overexpression are the strongest predicted pro-cardiac candidates; the majority of simulated TFs oppose cardiac differentiation upon overexpression.
- The off-target population retains a divergent regulatory program with elements of both fibroblast and cardiomyocyte fates, consistent with incomplete reprogramming.

Recommended Next Steps

- Experimentally validate the top knockout candidates (Ets1, Foxs1, Maff) and overexpression candidates (Klf5) in follow-up reprogramming experiments.
- Consider combinatorial perturbations of the top candidates to assess synergistic effects.
- Consider re-running CapyBio's CellOracle Differentiation Optimization Analysis with the off-target population as the reference trajectory to identify strategies for reducing off-target reprogramming.
- Consider running CapyBio's Capybara™ Cell Benchmarking Analysis alongside perturbation experiments to quantify changes in cell identity and confirm on-target improvements.

Background

Overview of User-Provided Data

The submitted dataset consists of 10,606 cells obtained across 3 samples, Day 0, Day 7, and FACS-sorted Day 14. Figure 1 shows a 2D visualization of the data, highlighting the sample of origin (left), key cell clusters (center), and a pseudotime ordering (right) as defined by you. The pseudotime ordering is used to rank cells along the fibroblast-to-cardiomyocyte reprogramming path. This ordering serves as the reference cardiomyocyte differentiation trajectory in the *in silico* perturbation analyses below, defining the direction of progress toward the cardiomyocyte fate against which each perturbation is scored.

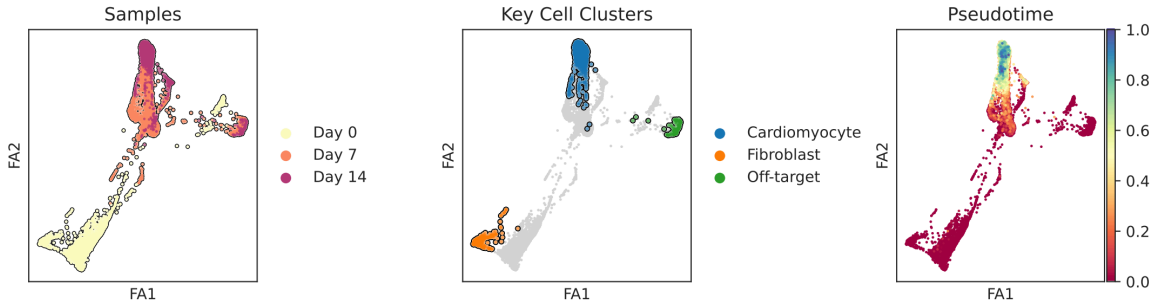

Figure 1. Quality control metrics for all cells in the dataset. Left: number of genes detected per cell. Center: total UMI (Unique Molecule Index) counts per cell. Right: percentage of reads mapping to mitochondrial genes.

Gene Regulatory Network Overview

To understand how gene regulation differs across cell states, we constructed gene regulatory networks (GRNs) for each cluster using CellOracle. Each network models which transcription factors regulate which target genes within that cell state. Briefly, a CapyBio-curated base network of transcription factor-target gene relationships was used as the regulatory scaffold, then refined using the single-cell gene expression data to estimate the strength of each regulatory connection within each cluster. This produces a cluster-specific regulatory map that captures how transcriptional control differs between cell states. A total of 160 transcription factors and 2,394 target genes with sufficient regulatory evidence were retained for downstream analysis.

Network Centrality Analysis

Network centrality metrics rank transcription factors by how influential they are within a cluster's GRN. Two metrics are reported to provide a more robust perspective. TFs that score highly on both are the most confident hits.

Degree centrality measures how many target genes a TF directly regulates in that cluster. A high score means the TF is a broad regulator.

Eigenvector centrality weighs those connections by the influence of the TFs on the other end. A high score means the TF regulates other influential regulators, not just many genes.

Comparing centrality profiles across clusters reveals how regulatory architecture changes with cell identity. Clusters with similar profiles share a common set of master regulators and are likely related cell states. Clusters with divergent profiles are controlled by distinct regulatory programs, indicating a fundamental shift in cell identity. TFs that are highly central in one cluster but not others are candidate drivers of that cluster's identity and the most informative targets for perturbation. In the analyses that follow, TF centrality is compared across two or more clusters to identify patterns of gene regulatory changes across different cell states in the data. Note that centrality reflects regulatory influence only, not its direction: a highly central TF in a given cell type may act as either an activator or a repressor of the respective identity, and the *in silico* perturbation analyses (shown in later sections) are needed to distinguish the two modes of action.

Results

Degree Centrality Across Clusters

Figure 2 shows a clustered heatmap of the top 20 most variable TFs by degree centrality across all clusters. TFs separate into two broad groups along the cardiac vs non-cardiac axis, with Egr1 and Mef2c showing consistently elevated centrality in the cardiomyocyte cluster. The fibroblast cluster groups closer to cardiomyocytes than to the off-target population, consistent with fibroblasts serving as the starting population for reprogramming and retaining elements of the regulatory machinery from which cardiac identity emerges. The off-target population is anchored by a distinct set of highly central TFs, indicating a divergent regulatory program.

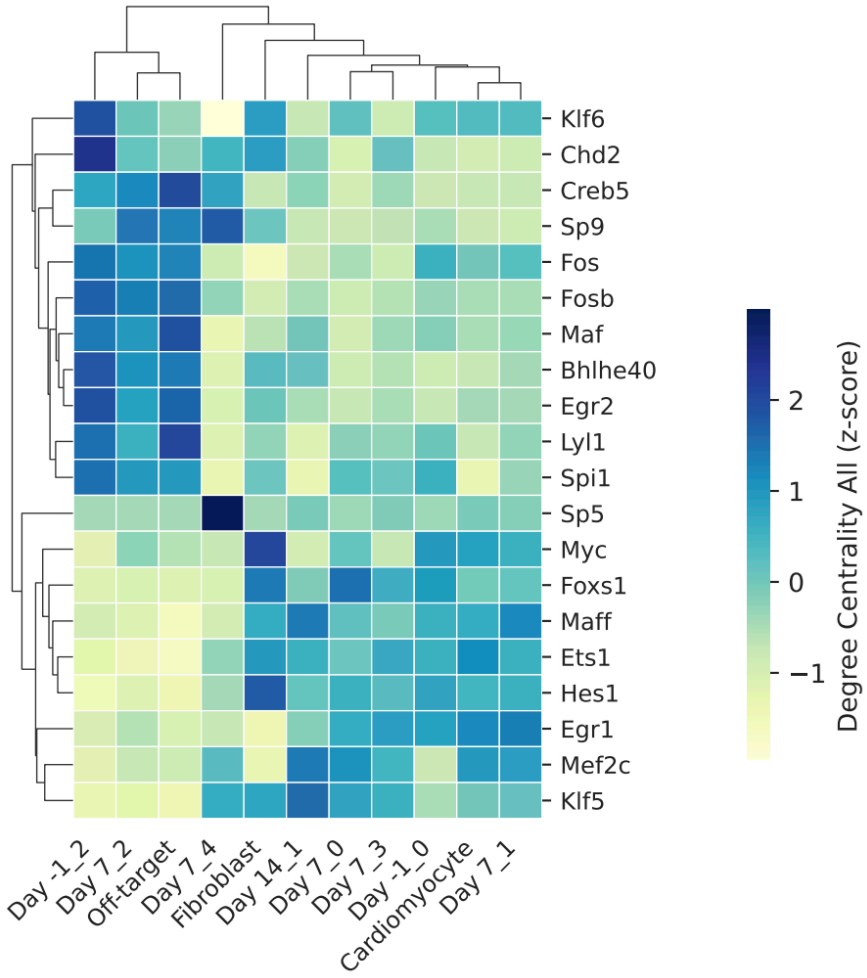

Figure 2. Clustered heatmap of the top 20 most variable TFs by degree centrality across all clusters. Each row represents a transcription factor and each column a cluster. Color indicates z-scored degree centrality, where darker blue reflects higher relative centrality within that cluster.

Eigenvector Centrality Across Clusters

Figure 3 shows a clustered heatmap of the top 20 most variable TFs by eigenvector centrality across all clusters. The overall cluster structure mirrors the degree centrality analysis, with the same broad separation between cardiac and non-cardiac states. Egr1 and Mef2c again show very high relative centrality in the cardiomyocyte cluster, and the off-target population remains distinct from both cardiomyocytes and fibroblasts. Egr1 stands out particularly strongly under the eigenvector metric, indicating that it not only regulates many cardiac targets but also sits among other influential regulators in the cardiomyocyte network.

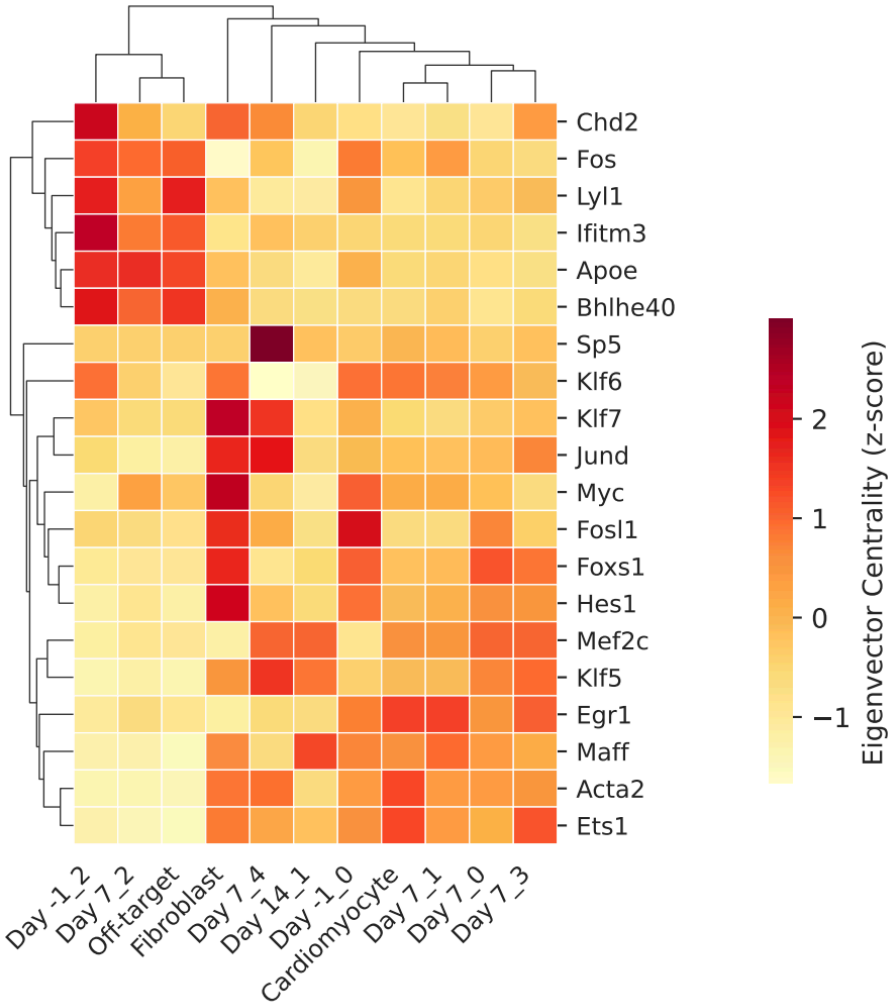

Figure 3. Clustered heatmap of the top 20 most variable TFs by eigenvector centrality across all clusters. Each row represents a transcription factor and each column a cluster. Color indicates z-scored eigenvector centrality, where warmer colors reflect higher relative centrality within that cluster.

Cardiomyocyte vs Fibroblast Cluster

The scatterplots in Figure 4 compare TF centrality between cardiomyocyte and fibroblast clusters. Each point represents a transcription factor. TFs above the diagonal have higher centrality in the cardiomyocyte cluster; TFs below the diagonal have higher centrality in the fibroblast cluster. TFs closer to the diagonal are similarly connected across the two cell clusters.

As shown, many TFs fall off the diagonal, indicating broad rewiring of the transcriptional regulatory network between cardiomyocytes and fibroblasts. Egr1 and Mef2c are strongly enriched in the cardiomyocyte network across both metrics. Several TFs including Atf3, and Klf6 show concordant centrality in both networks, suggesting shared regulatory roles across cell states. The overall extent of off-diagonal scatter is consistent with a fundamental shift in regulatory control underlying the change in cell identity.

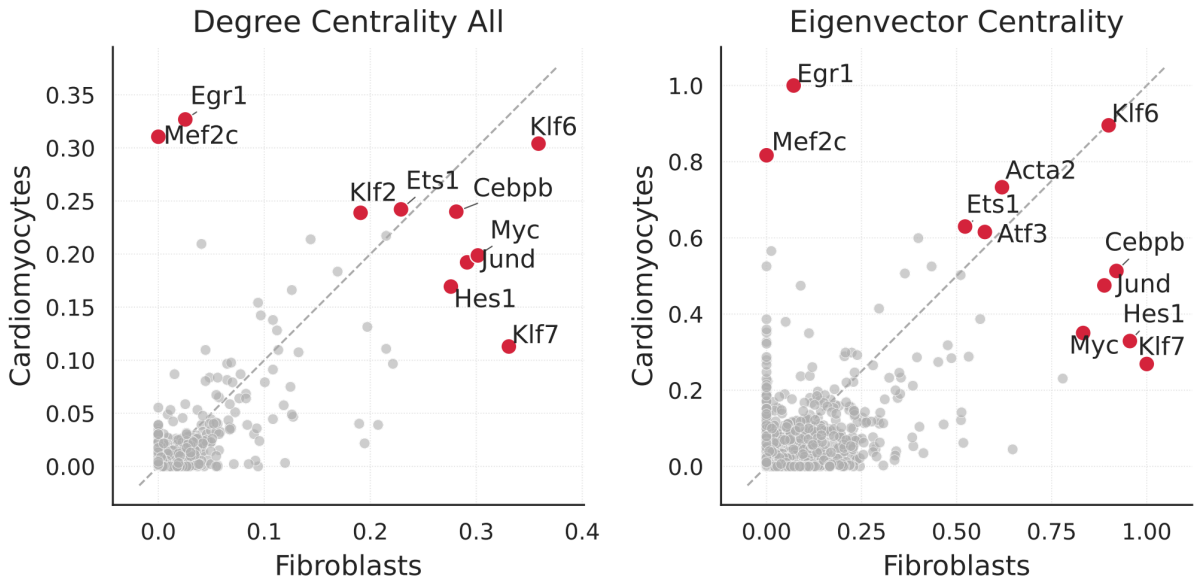

Figure 4. Centrality scatterplots comparing Cardiomyocytes to Fibroblasts.
Degree Centrality (Left) and Eigenvector Centrality (Right).

Cardiomyocyte vs Off-Target Cluster

The scatterplots in Figure 5 compare TF centrality between cardiomyocyte and off-target clusters. Each point represents a transcription factor. TFs above the diagonal have higher centrality in the cardiomyocyte cluster; TFs below the diagonal have higher centrality in the off-target cluster. TFs closer to the diagonal are similarly connected across the two cell clusters.

Comparing cardiomyocyte and off-target GRNs reveals a clear separation in transcriptional regulatory architecture. Egr1 and Mef2c again emerge as the most central cardiomyocyte regulators across both metrics, while the off-target network is anchored by a distinct set of highly central TFs including Lyl1, Fos, and Klf2 with minimal overlap at the core hierarchy level.

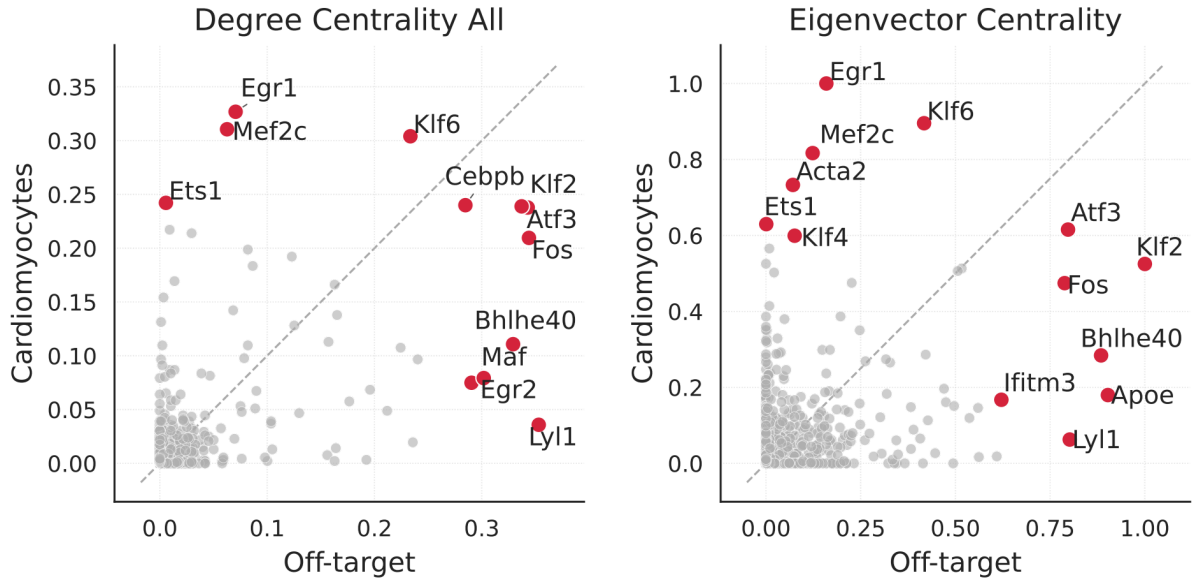

Figure 5. Centrality scatterplots comparing Cardiomyocytes to Off-target cells.
Degree Centrality (Left) and Eigenvector Centrality (Right).

Off-Target vs Fibroblast Cluster

The scatterplots in Figure 6 compare TF centrality between off-Target and Fibroblast clusters. Each point represents a transcription factor. TFs above the diagonal have higher centrality in the off-Target cluster; TFs below the diagonal have higher centrality in the Fibroblast cluster. TFs closer to the diagonal are similarly connected across the two cell clusters.

Comparing off-target and fibroblast GRNs reveals a more mixed regulatory landscape than the other pairwise comparisons. A substantial number of TFs fall near the diagonal, indicating shared regulatory roles consistent with the off-target population retaining elements of fibroblast identity. However, Egr1 and Mef2c show notably higher centrality in the off-target network than in fibroblasts, suggesting partial acquisition of cardiac transcriptional programs. Together, these patterns indicate that the off-target population occupies an intermediate regulatory space, having partially exited the fibroblast identity without fully committing to a cardiac program.

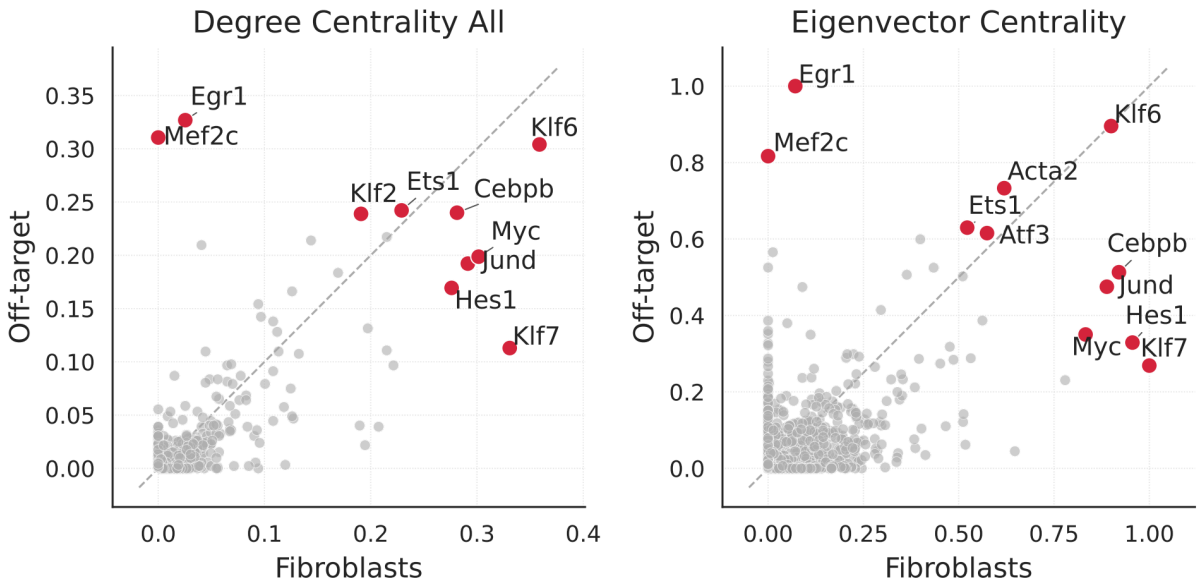

Figure 6. Centrality scatterplots comparing Off-Target to Fibroblast cells.
Degree Centrality (Left) and Eigenvector Centrality (Right).

In Silico TF Knockout

To identify transcription factors whose loss impacts cardiomyocyte differentiation, we simulated the knockout of the top 20 most variable TFs across clusters. For each TF, CellOracle removes its expression and models how this change propagates through the gene regulatory network, predicting how cells would shift in transcriptional space. These predicted shifts are then scored against the reference cardiomyocyte differentiation trajectory to produce a perturbation score. Throughout this report, "target" refers to the cardiomyocyte fate. A positive (pro-target) score indicates that removing the TF pushes cells toward the cardiomyocyte fate, suggesting it acts as a repressor of cardiac differentiation. A negative (anti-target) score indicates that its loss disrupts the cardiac program, suggesting it is required for cardiomyocyte differentiation.

Figure 7 shows the perturbation scores for all 20 simulated TFs, ranked by pro-target score. TFs on the left with high positive scores are predicted repressors of cardiac fate whose removal may improve reprogramming efficiency. TFs on the right with strong negative scores are predicted to be essential for the cardiac program. Some TFs perturbations show both pro- and anti-target components, suggesting a more context dependent role for them in iCM differentiation. Ets1 emerges as the strongest pro-cardiac knockout candidate, with Egr1 as a close second.

Figure 7. Cumulative perturbation scores for knockdown of the top 20 most variable TFs. Each bar represents one transcription factor. Purple bars indicate the cumulative pro-target perturbation score (predicted shift toward cardiomyocyte fate upon knockdown); orange bars indicate the cumulative anti-target score (predicted shift away from cardiomyocyte fate). TFs are sorted by pro-target score in descending order.

Knockout Perturbation Vector Fields

To visualize the predicted effect of knockout on the cell population, Figure 8 shows the predicted effect of knockout on the cell population for the four TFs with the strongest pro-target knockout scores. Each panel shows the predicted movement of cells across the embedding upon knockout, with arrows indicating direction and color indicating where cells accumulate (green) or deplete (red). TFs whose knockout consistently drives cells toward the cardiomyocyte cluster are strong candidates for experimental validation.

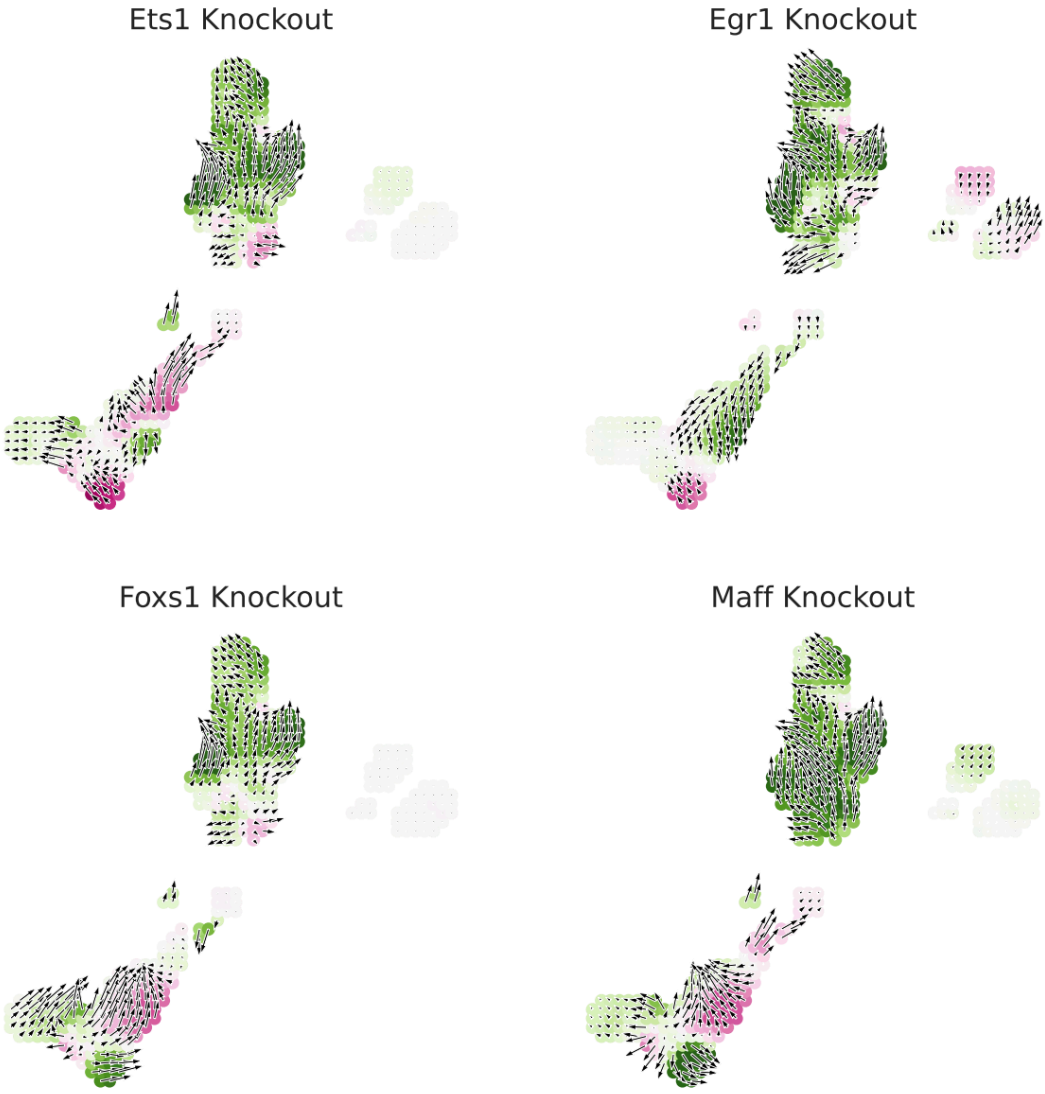

Figure 8. Perturbation vector fields for the four TFs with the strongest pro-target knockout scores. Each panel shows the predicted direction of cell movement upon knockout of the indicated TF. Green indicates regions of cell accumulation and red indicates depletion.

In Silico TF Overexpression

To identify transcription factors whose increased expression promotes cardiomyocyte differentiation, we simulated overexpression of the top 20 most variable TFs across clusters. For each TF, CellOracle increases its expression to the 98th percentile of all observed values and models how this change propagates through the gene regulatory network, predicting how cells would shift in transcriptional space. These predicted shifts are then scored against the reference cardiomyocyte differentiation trajectory in the same manner as the knockout analysis, with positive (pro-target) scores indicating a shift toward the cardiomyocyte fate and negative (anti-target) scores indicating opposition to the cardiac program.

Figure 9 shows the cumulative perturbation scores for all 20 simulated TFs, ranked by pro-target score. Klf5, Mef2c, and Chd2 emerge as the strongest pro-target overexpression candidates, predicted to drive cells toward the cardiomyocyte fate. Mef2c is a key TF part of the original reprograming cocktail used during iCM differentiation. Fos shows a mixed signal, with both pro- and anti-target components contributing to its overall score, indicating context-dependent effects that are visualized in the next section.

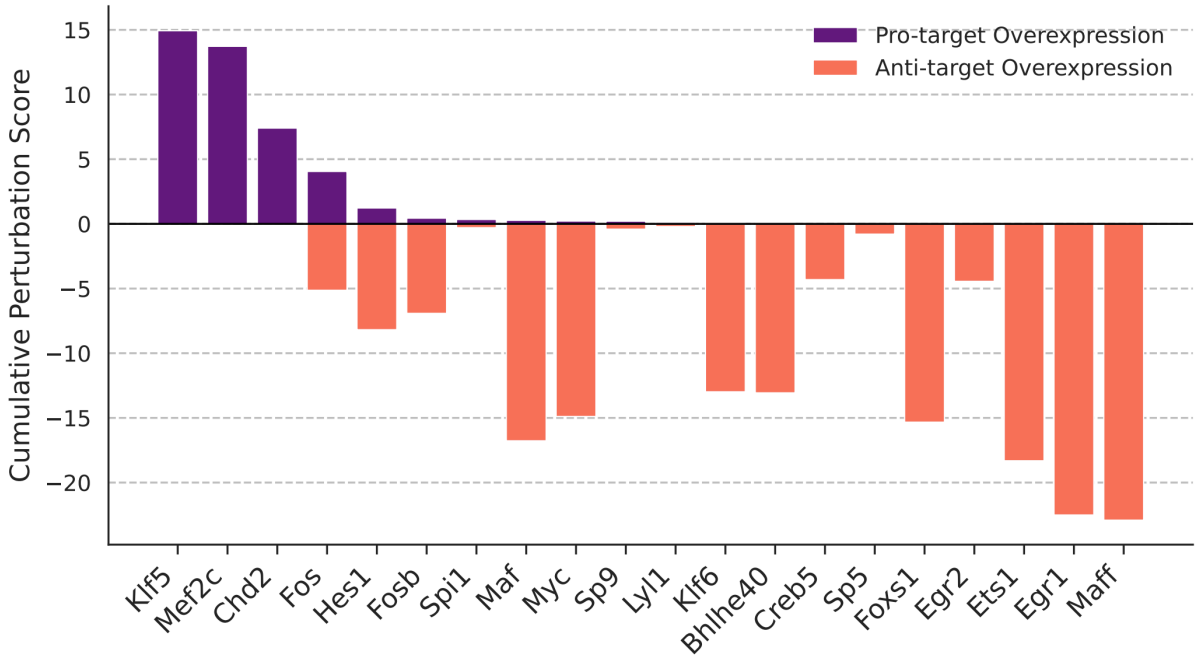

Figure 9. Cumulative perturbation scores for overexpression of the top 20 most variable TFs. Each bar represents one transcription factor. Purple bars indicate the cumulative pro-target perturbation score (predicted shift toward cardiomyocyte fate upon overexpression); orange bars indicate the cumulative anti-target score (predicted shift away from cardiomyocyte fate). TFs are sorted by pro-target score in descending order.

Overexpression Perturbation Vector Fields

To visualize the predicted effect of overexpression across the cell population, Figure 10 shows the predicted effect of overexpression for the four TFs with the strongest pro-target scores: Klf5, Mef2c, Chd2, and Fos. Each panel shows the predicted movement of cells across the embedding upon overexpression, with arrows indicating direction and color indicating where cells accumulate (green) or deplete (red). Klf5 and Mef2c show clean pro-cardiac effects, with consistent cell accumulation in the cardiomyocyte cluster and depletion in the fibroblast region. Chd2 shows a similar but more moderate effect, with some off-target depletion visible. Fos displays a mixed pattern, with both accumulation and depletion within the cardiomyocyte region, consistent with its ambiguous perturbation score and suggesting it plays a complex regulatory role during reprogramming.

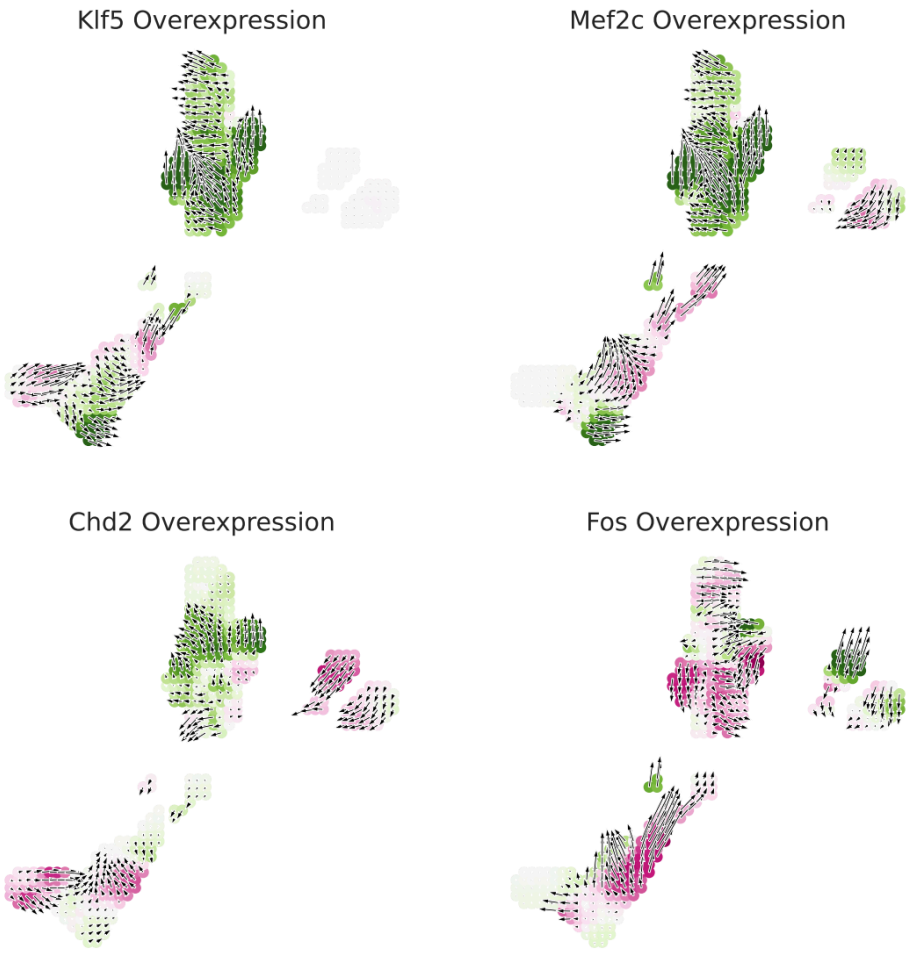

Figure 10. Perturbation vector fields for the four TFs with the strongest pro-target overexpression scores. Each panel shows the predicted direction of cell movement upon overexpression of the indicated TF. Green indicates regions of cell accumulation and red indicates depletion.

Conclusions

Gene regulatory network analysis reveals substantial rewiring of transcriptional regulatory programs across cell states. Network centrality analysis identifies Egr1 and Mef2c as the most central regulators in the cardiomyocyte network across both degree and eigenvector metrics. The off-target population occupies an intermediate and divergent regulatory space, retaining elements of fibroblast identity while partially acquiring cardiac transcriptional programs, consistent with incomplete reprogramming of a subset of cells.

In silico knockout simulations identify Ets1, Egr1, Foxs1, and Maff as the strongest candidate repressors of cardiac fate, whose removal is predicted to drive cells toward the cardiomyocyte trajectory. Notably, Egr1 ranks highly as a pro-target knockout despite its high centrality in the cardiomyocyte network, suggesting a central repressive regulatory role within the cardiac GRN. In silico overexpression simulations identify Klf5 and Mef2c as the strongest pro-cardiac overexpression candidates, while most of the simulated TFs show anti-target effects upon overexpression. Mef2c is a part of the TF cocktail used in the original differentiation process. Fos displays a mixed perturbation profile under both conditions, indicating a context-dependent regulatory role that warrants careful consideration before experimental follow-up. All predictions are based on the inferred GRN and should be validated experimentally.

Recommended Next Steps

As follow up to the results of this Cell Differentiation analysis, we recommend the following:

- Experimentally validate the top knockout candidates (Ets1, Foxs1, Maff) and over-expression candidate (Klf5) in follow-up reprogramming experiments to assess whether their suppression improves cardiomyocyte yield.
- Consider combinatorial perturbations of the top TF candidates to assess synergistic effects.
- Consider running CapyBio's Capybara Benchmarking Analysis alongside perturbation experiments to quantify changes in cell identity composition and confirm on-target reprogramming improvements.
- Consider re-running CapyBio's CellOracle Differentiation Optimization Analysis with the off-target population as the reference trajectory, to identify transcription factor perturbations that could reduce off-target reprogramming and improve the purity of the final cardiomyocyte product.

Methods

Gene regulatory network inference

Cell state-specific gene regulatory networks (GRNs) were inferred using CellOracle, CapyBio's computational platform for regulatory network analysis and in silico perturbation. Unlike conventional transcriptomic analyses that describe what genes are expressed, CellOracle models how transcription factors regulate gene expression within each cell state, producing a directed regulatory map that captures the logic underlying cell identity. A CapyBio-curated base network of known transcription factor-target gene relationships was used as the regulatory scaffold, then refined using machine learning to estimate the strength of each regulatory connection within each cluster. A total of 160 transcription factors and 2,394 target genes with sufficient regulatory evidence were retained for downstream analysis.

Network centrality analysis

The regulatory importance of each transcription factor within each cluster-specific GRN was quantified using two complementary metrics. Degree centrality was computed as the number of non-zero regulatory edges for each transcription factor node. Eigenvector centrality was computed using the leading eigenvector of the GRN adjacency matrix, capturing not just the number of connections but the importance of those connections within the regulatory hierarchy. Both metrics were computed per cluster and compared across cluster pairs.

In silico perturbation simulation

For each transcription factor, CellOracle simulates the effect of genetic perturbation by modifying its expression in the input matrix and propagating the resulting changes through the inferred GRN. Overexpression was simulated by setting TF expression to the 98th percentile of all observed values; knockout was simulated by setting expression to zero. Changes were propagated through the GRN for three iterations, producing cell-level predicted expression shifts that were converted to velocity vectors in the embedding space. This approach enables systematic, hypothesis-free identification of candidate regulators without requiring prior experimental perturbation data, a key advantage for prioritizing targets before committing to costly experimental validation.

Perturbation scoring

Perturbation scores were computed as the inner product between the simulated perturbation flow and the reference cardiomyocyte differentiation trajectory at each grid point, weighted by local cell density. Positive scores indicate alignment with the cardiomyocyte trajectory; negative scores indicate opposition. Scores were computed separately for positive (pro-target) and negative (anti-target) contributions to capture the full directionality of each perturbation.

Report generated by CapyBio Inc
Analysis powered by CellOracle
Questions: ***info@capybio.com***