

Evolving a Gene Panel for the Mouse Embryonic Heart

An LLM-agent evolutionary search on `panel-selection-bench`

July 21, 2026

Abstract

We apply an LLM-agent evolutionary search to design a 500-gene panel for a mouse embryonic-heart (E10.5–E14.5) scRNA reference. The genome is the gene list itself; a full-capability coding agent edits it, reasons from per-dimension feedback and a dataset brief, researches candidate genes (web search + expression analysis), and submits—while a *hidden* “biological prior” dimension (dim7) rewards covering the study’s key biology. Evolution swapped 165 redundant DE markers for canonical cardiac developmental genes, lifting the hidden prior coverage from 0.520 to 0.625 while holding the other dimensions. On the `panel-selection-bench` fit-for-purpose composite (FFP; dim7 up-weighted $\times 3$, evaluated on a held-out mouse-level test split), the evolved panel scores 0.778—**above every benchmark method**, including the DE-marker leader (0.738)—with the highest prior coverage of any method (dim7 0.635 vs. next-best 0.537). It also recovers the most congenital-heart-disease (CHD) genes (54 of 247) and, though the perturbation was never part of the objective, enriches for genes responding to a Runx1t1 knockout (37 of 185).

1 Setup

The dataset is a mouse embryonic-heart scRNA reference (Feng et al. 2022, E10.5–E14.5; 22,530 cells, 8 cardiac cell types, 18,204 detectable genes) with a mouse-level train/test split. Each panel is scored on seven dimensions relative to the whole-transcriptome ceiling; the headline is the `panel-selection-bench` *fit-for-purpose composite* (FFP), a weighted mean of the dimensions that up-weights the biological prior ($\times 3$) and down-weights feasibility ($\times 0.5$), with each relative score clipped to $[0, 2]$. All panels here are selected on the train split and scored on the held-out test cells. **dim7 (Prior) is hidden during evolution:** the agent is never told the prior (497 CHD genes + heart-development GO terms $\rightarrow$ 449 mouse genes, 430 detectable), so it must *discover* the relevant biology itself. Because the prior (430 detectable) fits inside a 500-gene budget, dim7 has real headroom here, unlike large-prior datasets where it is capacity-bound.

2 Optimization process

The single agent edits `panel.txt`, verifies each change with a remote scorer, and submits its best; the evolution loop refines across generations. Figure 1 shows the best-so-far trajectory. The hidden dim7 climbs from 0.520 to 0.625 *while q134 (the dims 1/3/4 sub-score) is held flat*—the intended behaviour: lift the biological prior without breaking identifiability, structure, or reconstruction. Most of the gain lands early (the agent finds the obvious cardiac pathways), then plateaus below the dim7 ceiling.

3 What the agent added and removed

The evolved panel differs from the DE seed by 165 genes (Figure 2). It *removed* redundant/uninformative DE markers—ribosomal, housekeeping and mitochondrial genes (`Atp5*`, `Cox*`, `Gapdh`) and dupli-

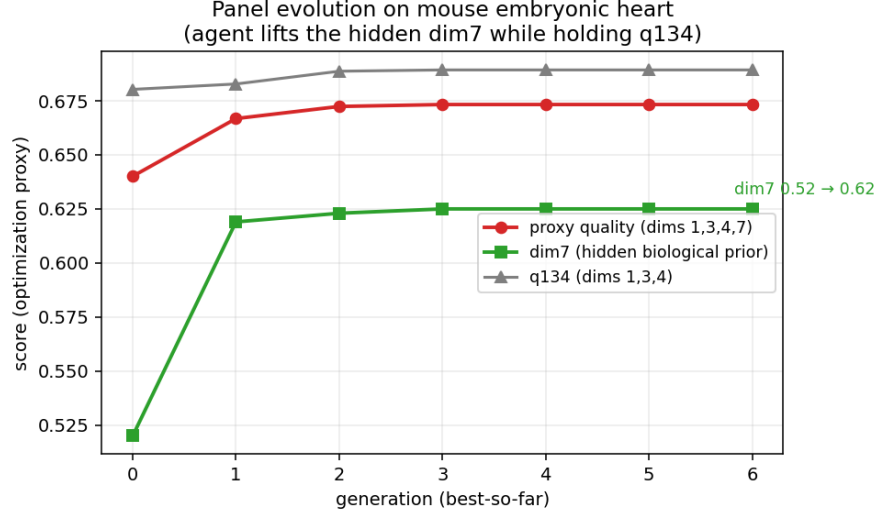


Figure 1: Best-so-far quality, hidden dim7, and q134 versus generation.

cated immune markers—and *added* canonical cardiac developmental genes spanning the major morphogenetic pathways: BMP/TGF β (Bmp2/4/10, Bmpr1a/2), Wnt (Ctnnb1, Axin2, Dkk1), Notch (D111/3/4, Hey1/2, Jag1), FGF (Fgf8/9), cardiac transcription factors (Gata4/6, Hand1/2, Isl1, Mef2c, Myocd), and conduction/ion-channel genes (Scn5a, Kcnq1, Cacna1c). Notably the agent *checked expression before committing*: it swapped back cardiac genes that are near-absent in this E10.5–E14.5 sample (Nodal, Tbx1, Zic3). These are precisely the biology of the hidden prior—rediscovered from web search and expression analysis, without ever seeing it.

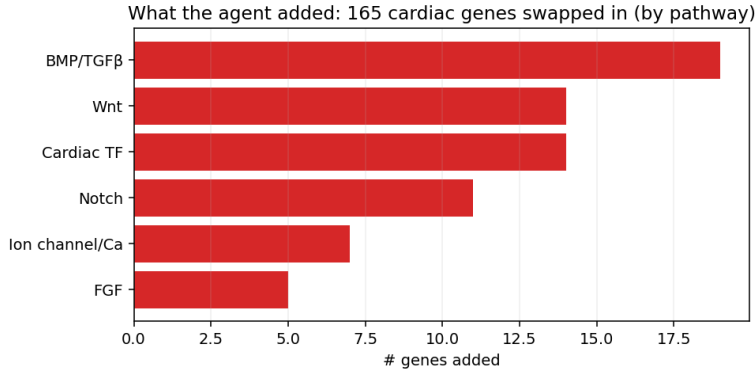


Figure 2: Cardiac developmental genes the agent added, grouped by pathway.

4 Comparison with DE and other panel-design methods

Table 1 and Figure 3 compare the evolved panel against the DE seed and every **panel-selection-bench** method at size 500, all on the same test-split FFP scale (our FFP values reproduce the published leaderboard for the benchmark methods to within 0.002). The evolved panel attains the highest composite (0.778 vs. the benchmark leader DE 0.738 and the DE seed 0.688), driven by the highest prior coverage of any method (dim7 0.635 vs. next-best 0.537)—while its identifiability, structure

and reconstruction stay in line with the strong baselines.

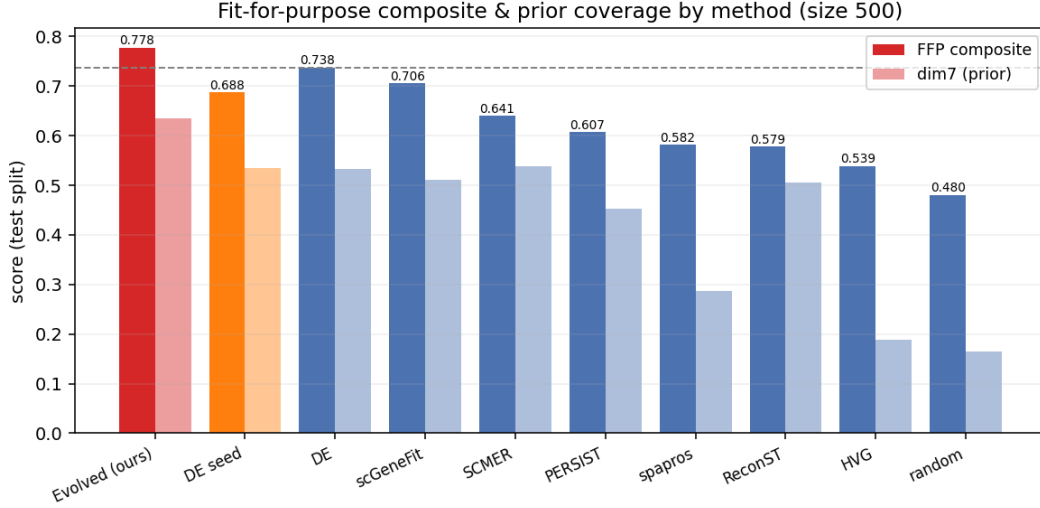


Figure 3: Quality and dim7 by method (mouse heart, size 500, full cells).

method	FFP	Identity	Struct	Recon	Prior	CHD $\cap$	KO $\cap$
Evolved (ours)	0.778	1.42	0.61	0.07	0.64	54	37
DE seed	0.688	1.35	0.64	0.09	0.54	30	30
DE	0.738	1.41	0.64	0.07	0.53	31	30
scGeneFit	0.706	1.35	0.61	0.11	0.51	23	16
SCMER	0.641	0.94	0.73	0.08	0.54	34	29
PERSIST	0.607	1.05	0.62	0.13	0.45	15	9
spapros	0.582	1.29	0.68	0.07	0.29	25	48
ReconST	0.579	0.85	0.65	0.07	0.51	26	8
HVG	0.539	1.37	0.59	0.07	0.19	6	31
random	0.480	1.05	0.59	0.10	0.16	4	3

Table 1: Per-method quality, dimensions, and biological-recovery overlaps (size 500). CHD $\cap$: genes shared with the 247 congenital-heart-disease genes. KO-DE $\cap$: genes shared with the 185 Runx1t1-knockout differentially-expressed genes ($p_{adj} < 0.05$).

5 Biological recovery: CHD genes, Runx1t1, and the Runx1t1-KO signature

Beyond the benchmark score we ask whether each panel captures disease-relevant biology (Figure 4). **CHD genes:** the evolved panel contains 54 of the 247 congenital-heart-disease genes, versus 30 for the DE seed—a substantial enrichment for the disease gene set. **Runx1t1:** the knock-out target gene *Runx1t1* is in the detectable universe but is *not* selected by the evolved panel nor the DE seed (see Table 1 for all methods)—it is not itself a strong cell-type marker. **Runx1t1-KO signature:** against the 185 genes differentially expressed in the Runx1t1 knockout (E14.5, $p_{adj} < 0.05$), the evolved panel recovers 37 versus 30 for the DE seed—so evolving toward the hidden cardiac-developmental prior also enriches for genes that respond to a cardiac-development perturbation,

even though the KO signature was never part of the objective. (One benchmark, spapros, recovers more of the KO set (48) despite a much weaker CHD overlap—its reconstruction-oriented, broadly-expressed selection happens to catch general KO-responsive transcriptional changes rather than the disease-gene core.)

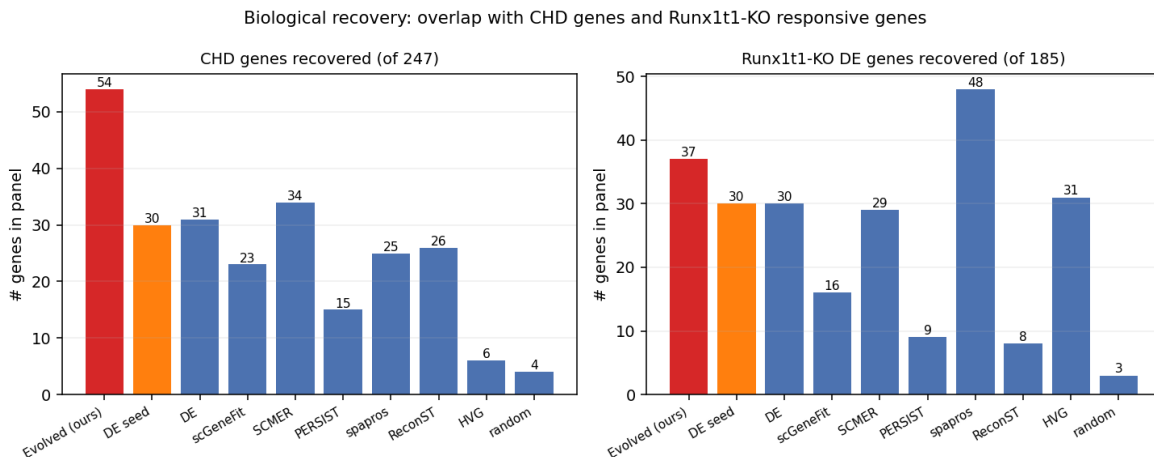


Figure 4: Overlap of each method’s panel with CHD genes (left) and Runx1t1-KO DE genes (right).

6 Discussion

The agent lifted the hidden biological dimension by ~ 0.10 while holding the other dimensions, recovering the CHD/heart-development biology of the (hidden) prior from first principles and incidentally enriching for a Runx1t1-KO perturbation signature. The main limitation is a plateau below the dim7 ceiling: once the obvious developmental pathways are in, further gains require either genes the agent cannot guess without the prior, or swaps that trade off against the other dimensions.