

Glioblastoma DRUG-seq: gene set cut to six by an all-drugs filter

LN-229 cells, 6 h, 11 active neuroactive drugs (NAD) vs 6 oncology drugs (ONCD) and 3 negative references (Lee et al. 2024). Task: report the genes that most clearly support a shared early NAD response, plus three mean shifts over those genes. Hidden key: **19 genes**, Jaccard ≥ 0.8 . The prompt gives no cutoff or set size. GPT-6 Astra computed every shift correctly, then fixed the set with two rules of its own.

GPT-6 Astra (Codex CLI)

FAIL · 1/10

What the key required

19 GENES · J ≥ 0.8

BEFORE · LINE 21 · ITEM_9 · TOOL OUTPUT
Genome-wide screen ranks all 12,738 genes

```
TOP contrast
IER3 NAD=1.463 ONCD=0.079 ... contrast=1.383
  nad_positive=10.000 nad_min=-0.482
IL11 NAD=0.917 ONCD=-0.454 ... contrast=1.371
  nad_positive=9.000 nad_min=-0.751
CXCL8 NAD=1.569 ONCD=0.256 ... contrast=1.313
  nad_positive=11.000
```

Shifts are correct. The top two genes by contrast, both key genes, are below zero in one or two NAD drugs.

DURING · LINES 22-26 · ITEM_10-12 · MESSAGE, CODE, OUTPUT

Requires all 11 drugs, then keeps the top six

Several genes increase across all 11 active neuroactive drugs, with smaller average shifts in the comparison treatments.

```
eligible=[g for g,s in summary.items()
  if s['nad_positive']==len(nad_drugs)]
selected=sorted(sorted(eligible,key=lambda g:
  summary[g]['contrast'],reverse=True)[:6])
{'shared_genes': ['ATF3','CXCL8','EGR1','FOS','JUN',
  'NR4A1'], 'nad_mean_shift': 1.5232, ...}
```

Error: the all-11 filter leaves 42 genes and drops 9 key genes, IER3 among them (sertindole -0.48). The top-6 cut then drops 4 more that ranked 7th-11th. Code: submission/analyze.py lines 31-32 (item_11).

AFTER · LINE 28 · REASONING.MD · GRADER LINE 17 · SCORE.PY

Calls the size arbitrary, and the set scores 0.32

```
"six is a compact reporting choice, not an inferred
biological boundary or a significance threshold."
shared_genes score 0.3157894736842105 pass false
missing: btg1 dnajb9 fosb gem herpud1 hspa1a hspa5
ier3 il11 inhba ptgs2 sat1 tgif1 extra: []
```

All six reported genes are key genes. The set is too small.

AT THIS STEP · ANSWER_KEY.JSON · GRADER/README.MD

Key genes pass six threshold rules

Each rule: NAD shift above 0.5-0.75, NAD-ONCD above 0.4-0.75, and shift above 0.25 in at least **8 or 9 of 11** NAD drugs. Genes passing all six (21, less 2 removed in review) form the core; 65 genes passing some rules are ignored.

```
ATF3 BTG1 CXCL8 DNAJB9 EGR1 FOS FOSB GEM HERPUD1
HSPA1A HSPA5 IER3 IL11 INHBA JUN NR4A1 PTGS2 SAT1
TGIF1
```

THE 13 MISSING KEY GENES · SOLVER'S OWN
GENOME_SUMMARY.TSV

Where each was lost

All-11 filter (positive in 9-10 of 11)	IER3 IL11 INHBA HSPA1A SAT1 DNAJB9 FOSB HERPUD1 TGIF1
Top-6 cut (rank among 42)	PTGS2 #7 · GEM #9 · HSPA5 #10 · BTG1 #11

Required: a rule tolerating one or two non-responding drugs, with no fixed count. The top 11 of the solver's 42 would still hold only 10 key genes (Jaccard 0.53).

AFTER · JUDGMENTS.JSON · GRADER LINE 26

One set decision fails eight criteria

C-002-C-004 shifts	1.5232 / 0.3913 / -0.0369 vs key 1.2403 / 0.2114 / -0.0061; tol 0.2553 / 0.1261 / 0.0282
C-005-C-008	fail on the C-001 prerequisite
C-009 (independent)	fail: no pooled vs single-drug contrast (grader's example BTG2 0.1418 vs vortioxetine 0.7686)

Grader: the six-gene arithmetic "is correct ... These are scorer failures, not arithmetic errors."

WHAT HAPPENED

GPT-6 Astra built a correct table of per-drug shifts, then fixed the set with two untested rules: a positive mean in all 11 drugs and a stop at six genes, which its reasoning calls a reporting choice. C-002-C-004 compare against fixed 19-gene values, so its correct six-gene arithmetic failed too; four reasoning points failed with C-001.

TRAINING TARGET

When no set size is given, derive the set from a stated rule that tolerates a minority of non-responding conditions, and check how the set moves when the rule changes. Compare pooled and single-condition effects for a gene before calling it unresponsive.