

PRISM: potent compounds removed by a curve-shape filter

PRISM 19Q4 secondary screen, 1,448 compounds, 489 cell lines, 8 doses. Name the non-oncology compounds with the clearest dose-supported low-concentration activity; compute two group means. Required set (answer_key.json, Jaccard ≥ 0.85): **19 compounds** whose low-dose killing persists at high dose in the same cell lines. GPT-6 Astra got both means exact, then filtered per-condition percentages with its own prevalence and no-reversal cutoffs, which removed six of the 19.

GPT-6 Astra (Codex CLI) **FAIL** · 2/7 · C-001 critical

BEFORE · LINES 20-24 · ITEMS 9-11 · MESSAGE, TOOL OUTPUT

All six missed compounds are on its screen

I'm checking the strongest low-dose hits across their full dose series; some have sharp reversals that weaken the evidence for a consistent concentration response.

```
('colchicine', ...) [0.0, 55.02, 0.0, 30.52, 80.4, 66.06, ...]
('EVP4593', ...) [0.21, 0.0, 0.0, 1.26, 12.18, 52.73, 43.28, ...]
('CHIR-124', ...) [0.0, 0.22, 2.63, 0.0, 11.84, 36.64, ...]
```

Percent of cell lines sensitive at each dose from 0.0006 μM upward: one number per condition, without cell-line identity.

DURING · LINES 36-42 · ITEMS 17-20 · RULE IN SUBMISSION/ANALYZE.PY:56-67

Selects on prevalence and curve shape

```
# >=50% sensitive at a measured dose <1 uM; measured higher-dose support;
any(float(r['dose'])<1 and
r['sensitive_percent']>=50 ...)
all(a<=b+1 for a,b in zip(full,full[1:]))
```

Error: both cutoffs are its own ("analyst-defined" in reasoning.md). The majority bar drops potent compounds that kill a minority of lines; the shape test rejects a compound for one noisy condition. Neither asks whether the same lines stay sensitive.

AFTER · LINES 42-45 · ITEMS 20-22 · OUTPUT, MESSAGE

Submits 16 names

```
SELECTED ['BMS-387032', 'ONX-0914', 'PP-121', ...
'thiram']
```

Both means exact: 2.9145 and 37.2194 (C-002, C-003 pass).

What the key required

19-name core · 45 ignored

BEFORE · ANSWER_KEY.JSON, VARIANT-SUPPORT.JSON

Same-cell sustained sensitivity

A cell line counts when sensitive (viability < 0.3) at ≥ 2 doses $\leq 0.625 \mu\text{M}$ and at a top dose. A compound qualifies when ≥ 1 -3% of ≥ 200 -300 measured lines count. Core = intersection of seven cutoff variants (grader/README.md).

DURING · SIX MISSED CORE COMPOUNDS · HTS002

Why the solver dropped them, and the key's evidence

Compound	Agent's exclusion	V1 sustained / eligible
colchicine	55.02% $\rightarrow$ 0% reversal	278/403 (69.0%)
oligomycin-a	repeated reversals	183/483 (37.9%)
EVP4593	52.73% $\rightarrow$ 43.28%	51/476 (10.7%)
mefexamide	max 47.83% $< 1 \mu\text{M}$	90/471 (19.1%)
sirolimus	max 24.17% $< 1 \mu\text{M}$	83/480 (17.3%)
CHIR-124	max 36.64% $< 1 \mu\text{M}$	50/456 (11.0%)

Required: follow cell lines across doses. Each passes all seven variants (variant-support.json, V1 shown). Exclusion values: solver output, line 24.

AFTER · GRADER LINE 13 (SCORE.PY) · JUDGMENTS.JSON

Set fails; three criteria fall with it

```
"score": 0.6842105263157895, "pass": false,
"detail": {"matched": 13, "expected": 19, ...
"ignored": ["harringtonine", "onx-0914",
"pyrithione-zinc"]}
```

Threshold 0.85. C-004 to C-006 require C-001. C-007 failed on its own terms.

WHAT HAPPENED

GPT-6 Astra reproduced the group statistics exactly and printed the dose profiles of all six compounds it later omitted. It reduced each compound to one sensitive percentage per dose and applied two cutoffs it chose: a majority of lines sensitive below $1 \mu\text{M}$, and no decrease from $0.01 \mu\text{M}$ upward. The key instead tracks whether killing persists in the same cell lines, which separates a low-dose effect from one noisy condition. The rule was documented and applied consistently; its set matched 13 of 19 names.

TRAINING TARGET

When asked for dose-supported activity in a cell-line panel, test whether the same cell lines respond across doses before judging the curve shape of aggregate percentages. Treat a single reversed condition as noise to check against neighbouring doses, and do not set a prevalence bar that excludes potent compounds active in a minority of lines.