

Cell Painting: low-dose hits cut by an added per-pair filter

A549 Cell Painting, 24 plate tables, 447 eligible samples. Required answer: the samples whose morphology is repeatable at 0.04 μ M but not at 10 μ M, each judged against a dose- and position-matched background. The key holds **13 sample IDs**; score.py passes at Jaccard ≥ 0.80 . GPT-6 Astra returned 3 IDs, all correct. Its final filter step removed the other 10.

GPT-6 Astra (Codex CLI) **FAIL** · 0/6 · Jaccard 0.231

What the data and rubric required

KEY · 13 IDs

BEFORE · LINES 15-37 · ITEMS 6-18 · OUTPUT, MESSAGE

Computes the defined score; matches backgrounds by column

```
rows/features/eligible 3267 1745 447
rates 0.05 [126, 279]
```

Some apparent low-dose exceptions cluster by well column, so I'm checking whether they remain distinctive against treatments in the same columns and plate layouts.

Uses the median of three replicate-pair Spearman correlations, as the task defines. Its backgrounds are pooled, same-layout, same-column within a layout, and DMSO. It never compares a sample with other samples at the same well position, and never compares DMSO wells by position.

DURING · LINES 48-52 · ITEMS 24-26 · MESSAGE, CODE, OUTPUT

Adds per-pair and all-background filters

Several other apparent exceptions retain a strongly correlated high-dose pair, making their low-dose specificity less convincing.

```
low_strong=all(r['p1_'+m]<=.01 for m in
['pooled','layout','column']) and r['p1_dms0']<=.025
consistent=min(r['pairs1'])>low_floor and
max(r['pairs6'])<high_ceiling
['BRD-A02710418-003-11-8', 'BRD-A31811760-001-06-7',
'BRD-K33852358-001-09-9']
```

Error: the task scores the median pair; finalize.py also requires every pair to clear a threshold. That screen removes 9 of the 10 missed key samples, and is the only reason for 3 of them. Column p = 0.0606 removes trimethobenzamide. (Filters re-applied by Gamow to the solver's own results_sensitivity.json.)

AFTER · LINE 56 · GRADER LINE 9 · ANSWER, SCORE.PY

Three correct IDs, ten missing

```
"low_dose_specific_samples": [3 IDs above]
Jaccard similarity 0.230769230769; threshold 0.8;
... missing [10 IDs]; extra [].
```

C-001 fails, so C-002 and C-003 fail on their prerequisite. The grader fails C-004 to C-006 for the missing DMSO position contrast and the column-only background.

BEFORE · SOLVER-TASK.JSON · DERIVATION.JSON

Repeated well position is a signal in the controls

profile repeatability means the median Spearman correlation across features between distinct replicate pairs at a given dose

DMSO pairs across plates	mean r	pairs
same well position	0.309	6,624
different position	-0.011	152,352

C-004 asks for this contrast. C-002 and C-005 ask for a background that accounts for dose, repeated well position and replicate count.

DURING · REFERENCE/TREATMENT_STATISTICS.CSV

Each sample against its own same-well background

sample	well	low	q95	high	q95
zaprinst	F18	0.719	0.630	0.081	0.274
trimethobenzamide	D24	0.745	0.667	-0.021	0.387
felbamate	B12	0.644	0.483	0.321	0.466

Required comparison: other samples at the same dose and well, from other layouts (945 triplets each). All three are low-dose-specific in all seven key analyses. The solver dropped zaprinst for one high-dose pair of 0.508 and felbamate for one of 0.531 (its own pairs6), and trimethobenzamide for its column p.

AFTER · ANSWER_KEY.JSON · DERIVATION.JSON

13 stable IDs; 34 boundary IDs ignored

```
core 13 · dont_care 34 · jaccard_threshold 0.8
submitted 3, all in core → 3/13 = 0.231
95% rule: 102/447 low dose, 249/447 high dose
```

11 of the 13 with no extras would pass (0.846). The three returned IDs were all correct; the recall was too low.

WHAT HAPPENED

GPT-6 Astra computed the specified score correctly and found real low-dose-specific samples. It treated repeated well position as a column effect inside one layout, so it never built the same-position background across layouts or tested DMSO wells by position. At the last step it combined four conservative filters, including a per-pair screen the task's median definition does not call for. That step removed 10 of the 13 key samples.

TRAINING TARGET

Keep the decision rule the task defines. When adding a robustness screen, report what it removes instead of letting it set the answer. Match backgrounds on the confounder the controls reveal, here exact well position, and show that control contrast before relying on it.