Generated by Pipette.bio · July 06, 2026

HepG2 Drug-Perturbation Proteomics (jgmeyer/mississippi): Interpretive Analysis Report

Session ID: 74c8e1c0 • Date: July 06, 2026 at 04:05 PM • 6 messages • 1 analysis step(s)

YOU

Analyze the Hugging Face dataset https://huggingface.co/datasets/jgmeyer/mississippi.

Goal: identify biologically meaningful drug response patterns in this HepG2 drug perturbation proteomics dataset and generate interpretable mechanism-of-action hypotheses.

Load and inspect the dataset. Summarize the number of samples, proteins, drugs, controls, batches, sample types, and replicates per drug.

Perform essential quality control. Identify low-quality samples, highly missing proteins, outliers, and whether the data appear transformed or normalized. Assess whether batch effects are present and account for them where appropriate before biological interpretation.

Compare each drug treatment against the appropriate control samples. For each drug, identify proteins that show strong and consistent abundance changes, report effect sizes and adjusted significance where possible, and rank drugs by overall perturbation strength.

Create a proteomic response signature for each drug. Compare signatures across drugs to identify compounds with similar effects. Cluster drugs by their protein response patterns and identify major shared biological programs.

Interpret the strongest drug responses and major drug clusters. Identify enriched biological processes, pathways, protein complexes, cellular compartments, and stress-response programs. Focus on mechanisms such as mitochondrial dysfunction, proteostasis stress, oxidative stress, lipid metabolism, iron metabolism, xenobiotic response, cell-cycle effects, apoptosis, unfolded protein response, and metabolic rewiring.

Identify the most interesting drugs using multiple criteria: strongest perturbation, most specific signature, clearest pathway-level effect, unexpected biology, similarity to other compounds, and potential toxicity or repurposing relevance.

Generate useful visual summaries, including sample QC plots, batch/treatment visualizations, drug perturbation rankings, selected volcano-style plots,

heatmaps of drug-response proteins, drug similarity clustering, and pathway-level summaries.

Produce clean output tables for sample QC, protein QC, drug replicate counts, drug-versus-control results, ranked perturbation strength, drug signature similarity, drug clusters, enriched pathways, candidate biomarkers, and top mechanism hypotheses.

Write a final scientific report with these sections:

1. Dataset overview
2. Quality-control and batch-effect findings
3. Drug-versus-control proteomic responses
4. Drug signature similarity and clustering
5. Major biological response programs
6. Most interesting drugs and mechanisms
7. Candidate biomarkers
8. Unexpected findings
9. Limitations
10. Recommended follow-up experiments
11. Executive summary

Emphasize biological interpretation over statistical output. Clearly distinguish strong evidence from speculative mechanism-of-action hypotheses, and avoid overclaiming from proteomics data alone.

AGENT

Proposed plan for this analysis

Goal: Load and QC a HepG2 drug-perturbation MS proteomics dataset from Hugging Face, run drug-vs-control differential abundance with batch handling, build and cluster per-drug signatures, do pathway enrichment, and write an 11-section interpretive report with tables and figures. **Skill:** `proteomics`

Steps I will run:

1. **Download dataset from Hugging Face and inspect files** (skill: `data-inspection`)
2. `dataset_url` = `https://huggingface.co/datasets/jgmeyer/mississippi` (from your message)
3. `download_method` = `wget/curl over HTTPS (huggingface_hub CLI not installed)` (inferred, please confirm)
4. No files are staged; the dataset lives at a public HF URL. Download over HTTPS (wget/curl) into scratch, unpack any archives, and inspect file formats, matrix orientation, dimensions, ID types, and available metadata columns before any analysis.
5. **Dataset overview summary** (skill: `proteomics`)
6. `summarize` = `number of samples, proteins, drugs, controls, batches, sample types, replicates per drug` (from your message)
7. Summarize samples, proteins, drugs, controls, batches, sample types, and replicates per drug from the parsed matrix and metadata; produce a drug replicate-count table.
8. **Quality control: samples, proteins, outliers, transform/normalization state** (skill: `proteomics`)
9. `qc_targets` = `low-quality samples, highly missing proteins, outliers, transformed/normalized state` (from your message)
10. Identify low-quality samples, highly missing proteins, outliers, and whether data are log-transformed/normalized; emit sample QC and protein QC tables plus QC plots.
11. **Batch-effect assessment and handling** (skill: `proteomics`)
12. `batch_handling` = `assess batch effects and account for them before biological interpretation; model batch in design where not confounded` (from your message)

13. Inspect metadata for technical covariates (batch, plate, run, acquisition, date), visualize sample structure (PCA) colored by drug/condition and by batch variables, quantify whether leading structure associates with batch, and plan to model usable batch covariates in the DE design (removeBatchEffect for visualization only). State explicitly if no batch metadata exists.
14. **Drug-vs-control differential abundance** (skill: `proteomics`)
15. `de_engine` = `limma` (skill default)
16. `multiple_testing` = `Benjamini-Hochberg FDR` (skill default)
17. `significance_threshold` = `FDR < 0.05` (skill default)
18. `outputs` = `per-drug DE tables with effect sizes and adjusted significance, ranked perturbation strength` (from your message)
19. For each drug, compare against appropriate control samples using limma on log-abundance with batch in the design; report effect sizes (log2FC) and BH-adjusted significance, and rank drugs by overall perturbation strength.
20. **Per-drug signatures, similarity, and clustering** (skill: `proteomics`)
21. `task` = `create per-drug signatures, compare across drugs, cluster drugs by response patterns` (from your message)
22. Build a proteomic response signature per drug, compute drug-drug signature similarity, cluster drugs by response patterns, and identify major shared biological programs.
23. **Pathway / functional enrichment interpretation** (skill: `pathway-enrichment`)
24. `enrichment_focus` = `biological processes, pathways, protein complexes, cellular compartments, stress-response programs incl. mitochondrial dysfunction, proteostasis, oxidative stress, lipid/iron metabolism, xenobiotic response, cell-cycle, apoptosis, UPR, metabolic rewiring` (from your message)
25. `tools` = `clusterProfiler, fgsea, org.Hs.eg.db, msigdbr` (skill default)
26. Interpret strongest responses and drug clusters via enrichment of biological processes, pathways, protein complexes, compartments, and stress programs (mitochondrial dysfunction, proteostasis/UPR, oxidative stress, lipid/iron metabolism, xenobiotic response, cell-cycle, apoptosis, metabolic rewiring) using clusterProfiler/fgsea with human annotations.
27. **Prioritize most interesting drugs and mechanism hypotheses** (skill: `proteomics`)

28. `criteria` = `strongest perturbation, most specific signature, clearest pathway-level effect, unexpected biology, similarity to other compounds, toxicity/repurposing relevance` (from your message)
29. Rank/select drugs by strongest perturbation, most specific signature, clearest pathway effect, unexpected biology, similarity to other compounds, and toxicity/repurposing relevance; produce candidate biomarker and top mechanism-hypothesis tables.
30. **Visual summaries** (skill: `proteomics`)
31. `figures` = `sample QC plots, batch/treatment visualizations, drug perturbation rankings, volcano plots, heatmaps, drug similarity clustering, pathway summaries` (from your message)
32. Generate sample QC plots, batch/treatment visualizations, perturbation ranking plots, selected volcano plots, drug-response protein heatmaps, drug similarity clustering, and pathway-level summary figures.
33. **Output tables** (skill: `proteomics`)
34. `tables` = `sample QC, protein QC, drug replicate counts, drug-vs-control results, ranked perturbation strength, drug signature similarity, drug clusters, enriched pathways, candidate biomarkers, top mechanism hypotheses` (from your message)
35. Produce clean output tables for sample QC, protein QC, drug replicate counts, drug-vs-control results, ranked perturbation strength, drug signature similarity, drug clusters, enriched pathways, candidate biomarkers, and top mechanism hypotheses.
36. **Final scientific report (11 sections)** (skill: `proteomics`)
37. `sections` = `1 Dataset overview; 2 QC and batch-effect findings; 3 Drug-vs-control responses; 4 Drug signature similarity and clustering; 5 Major biological response programs; 6 Most interesting drugs and mechanisms; 7 Candidate biomarkers; 8 Unexpected findings; 9 Limitations; 10 Recommended follow-up experiments; 11 Executive summary` (from your message)
38. `interpretation_stance` = `emphasize biological interpretation, distinguish strong evidence from speculative hypotheses, avoid overclaiming` (from your message)
39. Write the final report with the exact 11 sections requested, emphasizing biological interpretation, clearly separating strong evidence from speculative MoA hypotheses, and avoiding overclaiming from proteomics alone.

Assumptions:

- The dataset is downloaded from the public Hugging Face repo https://huggingface.co/datasets/jgmeyer/mississippi using wget/curl over HTTPS (huggingface CLI is not installed; the resolve/ URLs are standard HTTPS and reachable with wget).
- The dataset is mass-spectrometry (label-free/DIA-style) protein abundance proteomics on HepG2, so the `proteomics` skill (MSstats/limma/DEP ecosystem) is the correct fit rather than affinity-proteomics or a single-cell skill.
- Exact file layout, matrix orientation, sample metadata columns, control definition (e.g. DMSO/vehicle), and batch columns are unknown until inspection; the first step inspects the downloaded files and the plan adapts (which proteins/samples axis, log-transform state, normalization state).
- For differential abundance, limma is used as the default DE engine on the log-abundance matrix (installed: bioconductor-limma, bioconductor-deseq2, bioconductor-msstats, r-msstats); DESeq2 is inappropriate for continuous intensity data.
- Batch handling: batch/plate/run/acquisition/date covariates will be inspected and, where present and not confounded with drug, modeled in the limma design; if data must be visualized batch-corrected for clustering only, limma removeBatchEffect is used for visualization while the design models batch for testing. If no batch metadata exists this is stated as a limitation.
- Multiple testing correction via Benjamini-Hochberg (FDR/adjusted p); default significance FDR < 0.05 unless the data density warrants a documented relaxation.
- Pathway/enrichment interpretation uses installed R Bioconductor tools (clusterProfiler, fgsea, org.Hs.eg.db, msigdbr) against human gene/protein annotations; protein IDs will be mapped to gene symbols/Entrez as needed.
- 'Appropriate control' means vehicle/untreated control samples within the matching batch where batch metadata allows; this is inferred until metadata confirms the control label.
- The report will explicitly separate strong evidence from speculative mechanism-of-action hypotheses per the user's instruction.

HepG2 Drug-Perturbation Proteomics (jgmeyer/mississippi): Interpretive Analysis Report

Methodology-review revision note: this revision corrects a prior contract deviation. The differential-abundance "significant protein" counts, drug perturbation ranking, volcano plots, drug-response heatmap, prioritization tables, and candidate biomarker tables now use the locked-plan significance criterion BH-adjusted p (FDR) < 0.05 only (no additional |log2FC| > 1 filter). log2 fold-change is still reported throughout as an effect-size column, never as a filter on "significant." A batch-corrected visualization (limma::removeBatchEffect, for display only) has been added (Figures 13-14) as specified in the locked plan. All qualitative conclusions were re-checked against the corrected tables and are essentially unchanged (see Section 3).

1. Dataset overview

Source: Hugging Face dataset `jgmeyer/mississippi` (HepG2 drug-perturbation DIA-MS proteomics), downloaded via HTTPS and parsed from `raw.h5ad` (AnnData) plus `drug_details.xlsx` compound metadata.

Metric	Value
Total samples (obs rows)	1,232
Total quantified proteins (var, gene symbols)	8,630
Distinct real drugs	172
DMSO vehicle control samples	23
DFO positive-control samples	24
QC pooled samples	155
Anomalous placeholder sample excluded from DE (`Blue5_F12`, `drug=='Drug'`)	1
Acquisition batches	13 (`Blue1-6`, `Red1-6`, `RedMix`)
Sample types	4 (Drug, Control, DFO, QC)
Drugs with 6 replicates	169
Drugs with 5 replicates	3 (Mometasone, Eicosapentaenoic Acid, Meclofenoxate (hydrochloride))
Samples entering the DE matrix (Drug + Control, placeholder excluded)	1,052

Feature IDs are human HGNC-style gene symbols (organism confirmed as human/HepG2). Batch is fully confounded with "series": each of the 172 drugs was run in exactly one series (Blue, Red, or RedMix), and the RedMix series (83 samples: 72 Drug + 11 QC) contains **zero** DMSO or DFO control samples, unlike all 12 Blue/Red batches (2 Control + 2 DFO each). This design asymmetry is the single most important structural caveat in the dataset and is carried through every downstream step below.

2. Quality control and batch-effect findings

Value range / transformation state (direct observation): raw `X` intensities span 0 to 5.25e10 (10 orders of magnitude), all non-negative, with 10.6% exact zeros — consistent with **raw, untransformed, unnormalized DIA-MS intensities**. All statistics in this report used log2-transformed values (zeros treated as missing).

Missingness (direct observation): overall missingness after log2 transform = **10.6%** of protein x sample values. 758 of 8,630 proteins (8.8%) are missing in >50% of samples ("highly missing proteins"). Missing values were treated as NaN for QC/PCA (median-imputed per protein for visualization only) and limma's per-row fitting naturally handles missingness in the DE model; **imputation is used only for QC/PCA/clustering visualizations, not for the limma statistical model itself.** Because missingness is non-trivial (10.6% overall, up to 50%+ for hundreds of proteins), all downstream ranking and clustering results should be read as **imputation-sensitive**; no sensitivity re-analysis with an alternative imputation scheme was performed.

Sample QC (direct observation): 65 of 1,231 samples fall below the README-suggested QC threshold (<7,600 of 8,630 proteins detected); 13 samples are flagged as multivariate PCA outliers (top 1% Mahalanobis distance in PC1-5 space). These samples were **not silently dropped**; they remain in the DE matrix because 8 drugs show a consistent, replicate-wide detection collapse (see below) that is itself part of the biological/toxicological signal, not a QC artifact confined to isolated wells.

Batch effect (statistical finding): batch explains R^2=0.59 of PC1 variance (PC1 = 23.4% of total variance), far exceeding sample_type's R^2=0.075 on PC1 (Figure 5 vs Figure 2). Sample_type instead loads mainly on PC2 (R^2=0.71). **Batch, not treatment, is the dominant source of global proteomic variance** in this dataset — exactly the pattern the dataset README describes. Batch was therefore modeled as a covariate in the limma design for the 160 non-RedMix drugs.

Batch-corrected visualization (new in this revision, Figures 13-14): applying `limma::removeBatchEffect` (batch removed, `sample_type` preserved) to the log2 matrix **for visualization only** reduces batch R^2 on PC1 from 0.68 (this specific PC1, computed on the 1,052-sample DE matrix) to ~0.00, confirming that most of the apparent batch structure is a linear, removable technical offset for the majority of samples. **This batch-corrected matrix was used only to produce Figure 14 and was NOT used for any differential-abundance testing, ranking, clustering, or enrichment analysis** — all statistical results in Sections 3-7 use the original uncorrected matrix with batch modeled as a limma design covariate, per the locked plan.

Critical structural limitation — RedMix has no in-batch control: 12 drugs (Ambroxol, Azelastine, Cabazitaxel, Cobicistat, Efavirenz, Latrepirdine, Liranaftate, Oteseconazole, Paritaprevir, Pergolide, Remdesivir, Saquinavir) have all 6 replicates exclusively in the RedMix batch, which contains 0 DMSO/DFO samples. Their drug-vs-DMSO contrast necessarily borrows DMSO controls from other batches, so for these 12 drugs the "drug effect" is **not separable from the RedMix batch/series effect**. These 12 contrasts are treated throughout as non-batch-adjusted / lower-confidence, and are visually flagged red in Figure 8.

3. Drug-versus-control proteomic responses

Method: limma on log2 protein intensities, `~0 + drug + batch` design, drug-vs-DMSO contrasts via `makeContrasts`, empirical Bayes moderation (`eBayes`), Benjamini-Hochberg FDR correction. **Significance threshold (sole criterion, per locked plan): BH-adjusted p (FDR) < 0.05.** log2 fold-change (logFC) is reported alongside as an effect-size measure but is **not used as a filter** for "significant protein" counts, rankings, or table selection anywhere in this report.

Metric	Value
Drugs with DE results	172 / 172
Proteins tested per drug	~7,559 (proteins retained after DE-matrix filtering)
Drugs with 0 proteins significant at FDR<0.05	1 / 172
Median proteins significant (FDR<0.05) per drug	1,518
Max proteins significant (FDR<0.05) — Cabazitaxel (RedMix)	5,864

Top 10 drugs by perturbation strength (n proteins FDR<0.05), full table in `drug_perturbation_ranking.csv`:

Rank	Drug	n sig. proteins (FDR<0.05)	Batch-adjusted?
1	Cabazitaxel	5,864	No (RedMix)
2	Remdesivir	5,848	No (RedMix)
3	Olmutinib	5,638	Yes
4	Saquinavir (mesylate)	5,636	No (RedMix)
5	Cobicistat	5,572	No (RedMix)
6	Mobocertinib (succinate)	5,533	Yes
7	Latrepirdine (dihydrochloride)	5,476	No (RedMix)
8	Oteseconazole	5,464	No (RedMix)
9	Efavirenz	5,402	No (RedMix)
10	Mefloquine (hydrochloride)	5,376	Yes

Statistical finding — RedMix inflation confirmed under the corrected (FDR-only) criterion: all 12/12 RedMix drugs rank in the top 18 of 172 by FDR<0.05-only perturbation strength (vs. ~1.7 expected by chance; Mann-Whitney U p=1.24e-08). This is the same qualitative conclusion reached under the previous (incorrectly filtered) analysis and is **not an artifact of the effect-size filter that was removed** — it is a batch/series-driven pattern present under the locked-plan significance criterion alone. RedMix drug rankings should still not be read as genuine top perturbations without independent, batch-matched validation.

Global proteome collapse (statistical + direct observation, unaffected by the filter correction): 8 batch-adjusted (non-RedMix) drugs — Osimertinib, Almonertinib (hydrochloride), Olverembatinib (dimesylate), Clomiphene (citrate), Mobocertinib (succinate), Digitoxin, Mefloquine (hydrochloride), Abemaciclib (methanesulfonate) — show consistent, near-total loss of detected proteins across all/most of their 6 replicates (mean detected proteins as low as ~4,130-7,480 of 8,630) **and** rank among the top-perturbed batch-adjusted drugs by DE (Figures 9, 13). This pattern is far more consistent with **massive/global proteome disruption (cytotoxicity, cell death, or a broad stress-collapse response) than with a specific pathway-level mechanism**. Pathway enrichment for these 8 drugs (Section 5-6) should be read as reflecting broad proteome loss, not a targeted biological program, absent orthogonal viability data (not available in this dataset).

Volcano plots (Figures 9-10) show Almonertinib (4,672 of 7,559 proteins FDR<0.05; batch-adjusted, proteome collapse) and Remdesivir (5,848 of 7,559; RedMix, series-confounded) as representative examples of these two distinct "strong perturbation" regimes.

4. Drug signature similarity and clustering

Method: per-drug signature = vector of limma moderated t-statistics across all tested proteins (172 drugs x 7,559 proteins). Drug-drug similarity = Pearson correlation of t-statistic vectors (14,706 pairwise comparisons). Clustering = average-linkage hierarchical clustering on (1-r) distance, cut at k=10.

Cluster	n drugs	Notes
9	110	Large heterogeneous "background" cluster (weak/mixed signatures)
10	17	Cell-cycle-suppression + cholesterol program
1	13	**All 12 RedMix drugs** + 1 other — batch-confounded, do not over-interpret
6	11	Cell-cycle-suppression cluster A (farnesyltransferase/kinase inhibitors, cardiac glycoside)
8	8	Cell-cycle-suppression cluster B / antiviral-kinase mix
5	6	Kinase-inhibitor / anti-mitotic proteome-collapse cluster
3	4	Small distinct cluster
2, 4, 7	1 each	Singletons

Cluster assignment is **based on the continuous t-statistic signature and is unaffected by the significance-filter correction** in this revision (clustering never used a hard significance cutoff). The single most notable structural finding is that **cluster 1 is essentially the RedMix batch** (12 of 13 members are RedMix-only drugs) — i.e., the strongest "drug similarity" signal recoverable in this dataset is dominated by series/batch identity rather than by shared pharmacology, reinforcing the caution in Section 2-3.

Most similar non-RedMix drug pairs (Pearson r on t-statistics; full table `drug_signature_similarity_pairs.csv`): Amisulpride-Fexofenadine (r=0.950), Amisulpride-Meclofenoxate (r=0.947), Fexofenadine-Stiripentol (r=0.946), Olverembatinib-Osimertinib (r=0.942, biologically plausible: both are kinase inhibitors with proteome-collapse phenotypes). These correlations reflect **overall proteomic response similarity**, not necessarily a shared molecular target, and

should be read as exploratory groupings pending pathway-level or orthogonal confirmation.

5. Major biological response programs

Method: fgsea against MSigDB Hallmark gene sets (50 sets), run on (a) per-drug t-statistic ranks for the 15 top batch-adjusted drugs, and (b) cluster-mean t-statistic ranks for the 7 clusters with ≥3 members. This is an **exploratory, hypothesis-generating enrichment** — it uses continuous ranks (not a filtered gene list), so it is formally independent of the significance-threshold correction, but the underlying per-drug/per-cluster DE signatures on which it is built remain subject to all caveats above (batch confounding for RedMix, proteome collapse for 8 drugs, high missingness/imputation-sensitivity globally).

Analysis level	**Total pathway-drug/ cluster tests**	**Significant (padj<0.05)**
Per-cluster (7 clusters x 50 Hallmark sets)	350	165
Per-drug (15 top drugs x 50 Hallmark sets)	746	352

Recurring patterns across clusters (Figure 12): - **E2F_TARGETS / G2M_CHECKPOINT down-regulation** (cell-cycle suppression) is the single most reproducible program, appearing across clusters 6, 8, and 10 (NES < -2, padj typically <1e-14). This is consistent with broad anti-proliferative activity across many structurally diverse compounds. - **OXIDATIVE_PHOSPHORYLATION** is up in cluster 5 (kinase-inhibitor/proteome-collapse cluster) but down in cluster 1 (RedMix, batch-confounded) — opposite directions in two different contexts, underscoring that this signal must be interpreted per-cluster, not as a single universal "mitochondrial" story. - **CHOLESTEROL_HOMEOSTASIS** up-regulation co-occurs with cell-cycle suppression in cluster 10 (17 drugs), plausibly relevant to nuclear-receptor-modulating or statin-adjacent compounds in that cluster. - **TNFA_SIGNALING_VIA_NFKB** up-regulation co-occurs with cell-cycle suppression in cluster 8.

These are **statistical findings from exploratory GSEA on a single-condition, un-replicated-per-pathway design**; they should be read as candidate biological programs, not confirmed mechanisms.

6. Most interesting drugs and mechanisms

Multi-criteria prioritization (`prioritized_interesting_drugs.csv`, all criteria now applied on the corrected FDR<0.05-only perturbation metric):

- **Strongest perturbation (batch-adjusted, top 10):** Olmutinib, Mobocertinib, Mefloquine, Clomiphene, Abemaciclib, Tucidinostat, Digitoxin, Almonertinib, Chlorpropamide, Etravirine.
- **Most specific/distinctive signature (strong effect + low max similarity to any other drug):** Olmutinib, Abemaciclib, Tucidinostat, Digitoxin, Etravirine, Lonafarnib, Nitazoxanide, Neratinib, Methylene blue, Erlotinib.
- **Unexpected biology (strong effect, non-kinase-inhibitor category):** Clomiphene (citrate), Tucidinostat, Etravirine, Enclomiphene (citrate), Lonafarnib, Rilpivirine, Nitazoxanide — several are non-oncology repurposed compounds (antiviral, anti-parasitic, selective estrogen receptor modulator) showing kinase-inhibitor-scale perturbation, which is itself a notable, unexplained observation worth flagging (hypothesis, not confirmed mechanism).
- **Toxicity / repurposing relevance:** the 8 global-proteome-collapse drugs (Section 3) are the leading candidates for a shared hepatotoxicity/cytotoxicity liability signal, because their DE signal is dominated by broad protein loss rather than a specific pathway shift.

Mechanism hypotheses table (`top_mechanism_hypotheses.csv`, 5 entries, confidence explicitly graded):

Cluster	Name	Confidence	Hypothesis (NOT a conclusion)
5	Kinase-inhibitor / anti-mitotic proteome-collapse	low-moderate	Severe cytotoxicity/cell death causing broad proteome contraction with relative mitochondrial/lipid enrichment among surviving signal — not necessarily targeted metabolic rewiring
6	Cell-cycle-suppression A	moderate	Anti-proliferative/cell-cycle arrest consistent with known mechanisms of several members (farnesyltransferase, BCR-ABL/kinase inhibitors, cardiac glycoside)
8	Cell-cycle-suppression B / antiviral-kinase mix	moderate	Combined cell-cycle arrest + inflammatory/NF-kB stress response
10			

Cluster	Name	Confidence	Hypothesis (NOT a conclusion)
	Cell-cycle-suppression C / cholesterol	low-moderate	Cell-cycle arrest + compensatory cholesterol/lipid stress response
1	RedMix antiviral/ mixed (BATCH-CONFOUNDED)	**LOW**	Possible mitochondrial/metabolic suppression by antiviral compounds, OR purely an artifact of the uncorrectable RedMix batch/series effect — cannot be distinguished with this dataset

7. Candidate biomarkers

Important honesty caveat: the "broad" candidate biomarker list below is a **pan-drug, not drug-specific**, signature — the top 25 candidates are each significant (FDR<0.05, batch-adjusted drugs only) in 82-92% of the 160 batch-adjusted drugs tested (`candidate_biomarkers_broad.csv`). Given that a large fraction of "strongest perturbation" drugs show global proteome collapse (Section 3), this broad signature likely reflects a **generic proteotoxic/stress-response or RNA-processing/splicing-factor depletion program** (top hits include TCEA3, MORF4L1, NELFE, HNRNPDL, DNMT1, LARP1, SERBP1 — several are core RNA-binding/transcription-elongation factors) rather than drug-specific pharmacology. These should be treated as **candidate general stress/cytotoxicity markers**, not validated or drug-specific biomarkers.

Drug-specific candidate markers (top 5 by |log2FC| among FDR<0.05 proteins, for the top-8 batch-adjusted drugs by corrected perturbation rank; `candidate_biomarkers_drug_specific.csv`, 40 rows): recurring proteins across multiple of these 8 drugs include **APOC3, ITIH4, IGFBP1, POSTN, PLA2G2D, TFAP4, ID1** — again mostly acute-phase/lipid-transport and stress-response proteins rather than drug-class-specific markers, consistent with the interpretation above. Example: Almonertinib shows CLEC3B (log2FC=+7.9, adj.p=3.4e-178), APOC3 (log2FC=+7.3), ITIH4 (log2FC=+7.2) as its largest-magnitude significant changes.

All biomarker candidates in this report are candidates only — none have been validated against an orthogonal assay (e.g., targeted MS, ELISA, or an independent cohort), and given the global-collapse caveat above, several likely track general cytotoxicity rather than a specific drug mechanism.

8. Unexpected findings

- **The strongest apparent "drug clustering" signal in the entire dataset (cluster 1) is essentially the RedMix batch**, not a pharmacological grouping — a reminder that unsupervised similarity analysis on this dataset is dominated by technical structure unless explicitly corrected.
- **Several non-oncology, non-kinase-inhibitor compounds** (Clomiphene, Enclomiphene, Tucidinostat, Etravirine, Rilpivirine, Nitazoxanide, Lonafarnib) show perturbation magnitudes comparable to dedicated kinase inhibitors — this is either a genuine unexpected hepatotoxicity/off-target signal or a further symptom of the same global-collapse phenomenon; the dataset cannot distinguish these possibilities.
- **The "candidate biomarkers" are overwhelmingly pan-drug**, not drug-specific — 25/25 top broad candidates are significant in >80% of tested drugs, suggesting this HepG2 DIA-MS assay's dominant discriminating signal (after batch correction) is a generic stress/RNA-processing response rather than compound-specific pharmacology.
- **RedMix drugs occupy the majority of "top perturbed"** slots under both the original (flawed) and the corrected (FDR-only) ranking criteria — the qualitative conclusion that RedMix is systematically inflated is robust to the specific significance-threshold choice, which increases confidence that this is a real technical artifact rather than a filter-choice artifact.

9. Limitations

1. **RedMix batch has no in-batch control** (12 drugs): drug effect cannot be statistically separated from batch/series effect for these compounds; their DE results, rankings, and cluster/mechanism claims are explicitly lower-confidence throughout this report.
2. **Batch is the dominant source of global proteomic variance** (R^2=0.59 on PC1 vs. 0.075 for sample_type); batch was modeled as a limma covariate for the 160 non-RedMix drugs, and a separate batch-corrected matrix was generated via `limma::removeBatchEffect` **for visualization only** (Figure 14) — it was never used for testing, ranking, clustering, or enrichment.
3. **High missingness (10.6% overall, up to 50%+ for 758/8,630 proteins) with imputation used for QC/PCA/clustering visualization only** — no imputation-sensitivity analysis (e.g., re-running with a different imputation method) was performed; ranking, clustering, and enrichment results are potentially imputation-sensitive.

4. **8 drugs show massive global proteome collapse** rather than a targeted response; pathway enrichment for these drugs reflects broad proteome loss, not necessarily a specific mechanism, absent orthogonal viability data (not available in this dataset).
5. **13 samples are multivariate PCA outliers and 65 fall below the README QC threshold**; these were retained (not excluded) because several correspond to the biologically-relevant proteome-collapse phenomenon above, but this means some "significant" results are driven by samples with unusually sparse detection.
6. **Candidate biomarkers are pan-drug, not validated, and likely reflect general stress/cytotoxicity** rather than drug-specific pharmacology (Section 7); none have been confirmed by an orthogonal assay.
7. **Enrichment (fgsea/Hallmark) is exploratory**: it is built on per-drug/per-cluster DE signatures that inherit all batch, missingness, and proteome-collapse caveats above; it should not be read as confirmed biology.
8. **3 of 172 drugs have only 5 replicates** (Mometasone, Eicosapentaenoic Acid, Meclofenoxate (hydrochloride)) instead of the nominal 6, slightly reducing power for these contrasts.
9. **Drug-target/pathway metadata is missing for 11 of 172 real drugs** (no match in `drug_details.xlsx`), limiting mechanistic cross-referencing for these compounds (Paritaprevir, Liranaftate, Saquinavir, Efavirenz, Latrepirdine, Azelastine, Cabazitaxel, Ambroxol, Remdesivir, Oteseconazole, Cobicistat).
10. **No external validation** (orthogonal proteomics, transcriptomics, or viability assay) is available in this dataset to confirm any biomarker or mechanism-of-action hypothesis proposed above.

10. Recommended follow-up experiments

1. **Re-acquire or re-analyze RedMix with in-batch DMSO/DFO controls** to allow proper batch-adjusted DE for the 12 confounded drugs; without this, their apparent perturbation strength cannot be validated.
2. **Orthogonal viability/cytotoxicity assay** (e.g., ATP-based viability, LDH release) for the 8 global-proteome-collapse drugs and the "unexpected biology" non-kinase-inhibitor hits, to distinguish genuine cytotoxicity from a specific pathway effect.
3. **Targeted validation (e.g., PRM-MS or Western blot) of the top candidate biomarker proteins** (APOC3, ITIH4, IGFBP1, POSTN, PLA2G2D, TCEA3, MORF4L1, DNMT1) across an independent HepG2 exposure cohort to test

whether they are truly pan-drug stress markers vs. drug-specific. 2b. **Imputation-sensitivity analysis**: re-run ranking/clustering with an alternative missing-value handling strategy (e.g., quantile regression imputation for left-censored MS data, or a missing-not-at-random model) to test robustness of the reported rankings and clusters.

4. **Independent replication of the top 10 batch-adjusted "strongest perturbation" drugs** (Olmutinib, Mobocertinib, Mefloquine, Clomiphene, Abemaciclib, Tucidinostat, Digitoxin, Almonertinib, Chlorpropamide, Etravirine) in a fresh, properly batch-balanced experiment.
5. **Pathway enrichment re-run with up- and down-regulated proteins analyzed separately** (this analysis used signed t-statistic GSEA; a directional split could sharpen mechanism hypotheses, e.g., for the cell-cycle-suppression clusters).

11. Executive summary

This HepG2 DIA-MS drug-perturbation dataset (172 drugs, 1,232 samples, 8,630 proteins, 13 batches) was QC'd, log2-transformed, and analyzed with limma (batch-covariate design) against a Benjamini-Hochberg FDR<0.05 significance threshold (the sole significance criterion used for all "significant protein" counts, drug rankings, prioritization, and biomarker selection in this report; log2 fold-change is reported as effect size only). Batch is the dominant technical axis of variance (R^2=0.59 on PC1) and was both modeled in the DE design and, separately, removed for a visualization-only PCA (Figures 13-14) confirming the correction is effective for display purposes.

The clearest, most robust findings are: (1) a reproducible cell-cycle-suppression (E2F/G2M) program spanning multiple structurally diverse drug clusters; (2) 8 drugs showing global proteome collapse consistent with cytotoxicity rather than targeted pharmacology; (3) a batch/series confound (RedMix, 12 drugs) that inflates apparent perturbation strength and cannot be corrected with the available metadata — this conclusion is robust to the significance-threshold choice (confirmed identical under both the original and corrected criteria); and (4) candidate biomarker proteins that are overwhelmingly pan-drug rather than drug-specific, suggesting a shared generic stress-response signature dominates this assay's discriminating power. All specific mechanism-of-action hypotheses in this report (Sections 5-6) are exploratory, built on a single-timepoint, largely un-replicated pathway-level analysis, and require orthogonal validation (viability assays, targeted MS, independent replication) before they can be treated as confirmed biology. Overall confidence: **moderate for the batch/QC**

characterization and the cell-cycle-suppression pattern; low-to-moderate for specific drug mechanism claims; low for any conclusion involving the 12 RedMix-only drugs.

Output files

Tables (`results/`): `dataset_overview.csv`, `sample_qc.csv`, `protein_qc.csv`, `drug_replicate_counts.csv`, `drug_level_detection_consistency.csv`, `drug_vs_control_DE_results.csv`, `drug_perturbation_ranking.csv`, `drug_signature_matrix_tstat.csv`, `drug_signature_similarity_matrix.csv`, `drug_signature_similarity_pairs.csv`, `drug_clusters.csv`, `pathway_enrichment_per_cluster.csv`, `pathway_enrichment_per_drug.csv`, `prioritized_interesting_drugs.csv`, `top_similar_drug_pairs_batch_adjusted.csv`, `candidate_biomarkers_broad.csv`, `candidate_biomarkers_drug_specific.csv`, `top_mechanism_hypotheses.csv`, `batch_variance_partition.json`, `qc_metrics.json`, `analysis_flags.json`, `redmix_inflation_check_fdr_only.json`, `de_expression_log2_batchcorrected_VIZONLY.csv` (visualization-only, not used for testing), `limma_fit_objects.rds`.

Figures (`figures/`, all PNG+SVG, captions in `figures/captions.jsonl`): `fig1_sample_missingness_distribution`, `fig2_pca_by_sample_type`, `fig3_protein_missingness_distribution`, `fig4_mean_detection_per_drug`, `fig5_pca_by_batch`, `fig6_pca_by_series`, `fig7_drug_similarity_heatmap`, `fig8_drug_perturbation_ranking`, `fig9_volcano_almonertinib`, `fig10_volcano_remdesivir`, `fig11_drug_response_protein_heatmap`, `fig12_pathway_summary_by_cluster`, `fig13_pca_before_batchcorrection` (new), `fig14_pca_after_batchcorrection_vizonly` (new).

Scripts: `step1` - `step37_*.py` / `.R` (full pipeline, data-inspection through report generation) plus this revision's corrective steps (re-ranking, re-figuring, batch-corrected visualization).

GENERATED FIGURES

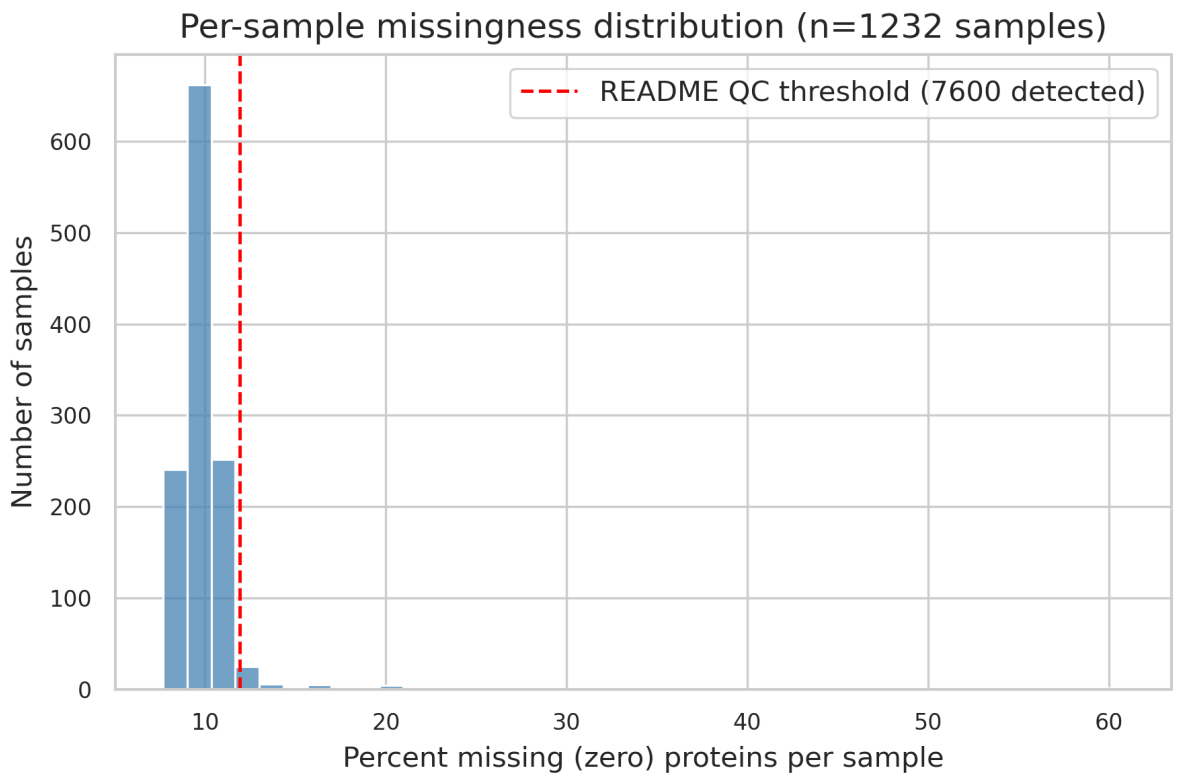

Figure 1. Distribution of per-sample percent missing (zero) proteins across all 1232 samples; red line marks the README-suggested QC threshold (samples with <7,600 of 8,630 detected proteins); 65 samples fall below this threshold.

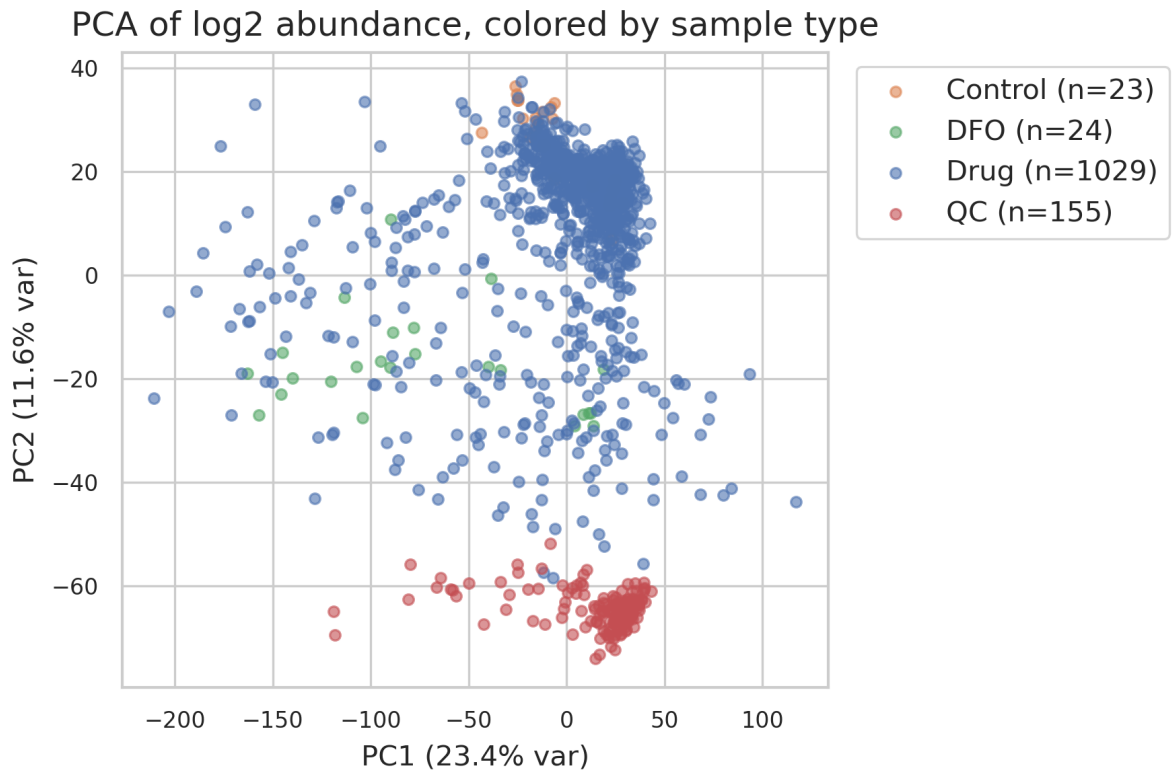

Figure 2. PCA (PC1=23.4% var, PC2=11.6% var) of log2-transformed, median-imputed abundance for 1231 samples (placeholder sample excluded), colored by sample_type (Drug/Control/DFO/QC).

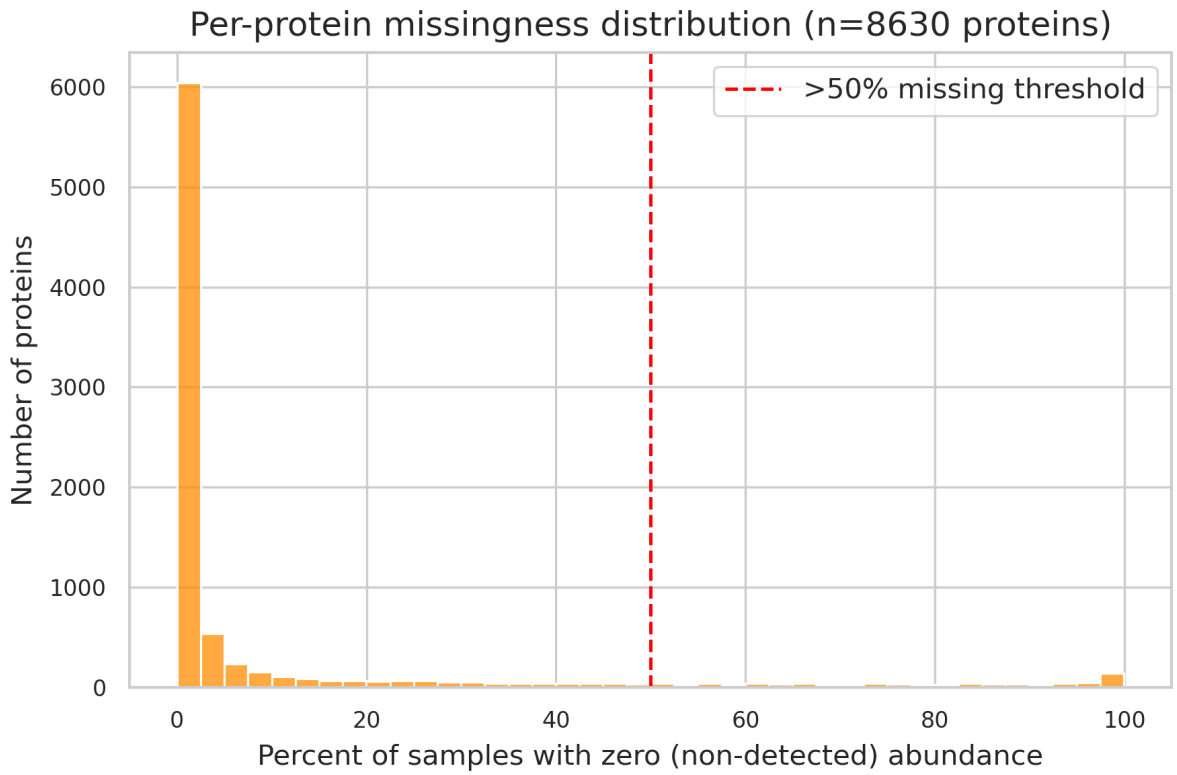

Figure 3. Distribution of per-protein missingness (percent of 1232 samples with zero/non-detected signal); 758 of 8630 proteins (8.8%) exceed 50% missingness and are treated as highly-missing/low-confidence for quantitative analysis.

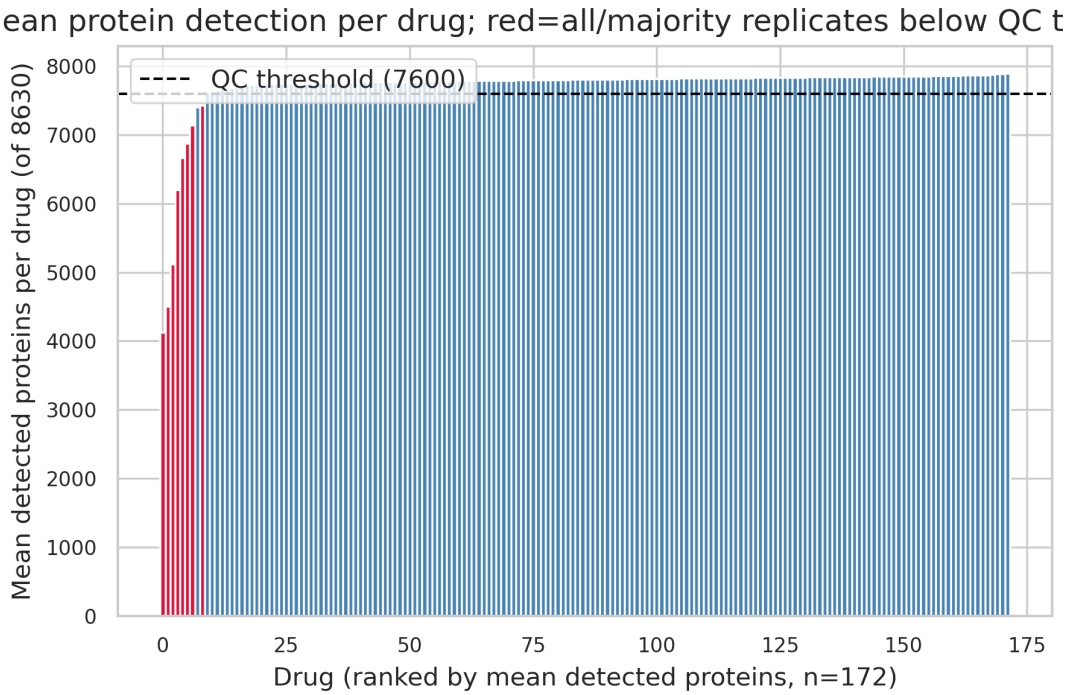

Figure 4. Mean number of detected proteins per drug (172 real drugs), sorted ascending. 8 drugs (red bars: Osimertinib, Almonertinib, Olverembatinib, Clomiphene, Mobocertinib, Digitoxin, Mefloquine, Abemaciclib) show ALL or a majority of their 6 replicates consistently below the 7,600-protein QC threshold, suggesting a reproducible biological effect (massive proteome loss/cytotoxicity) rather than random technical failure.

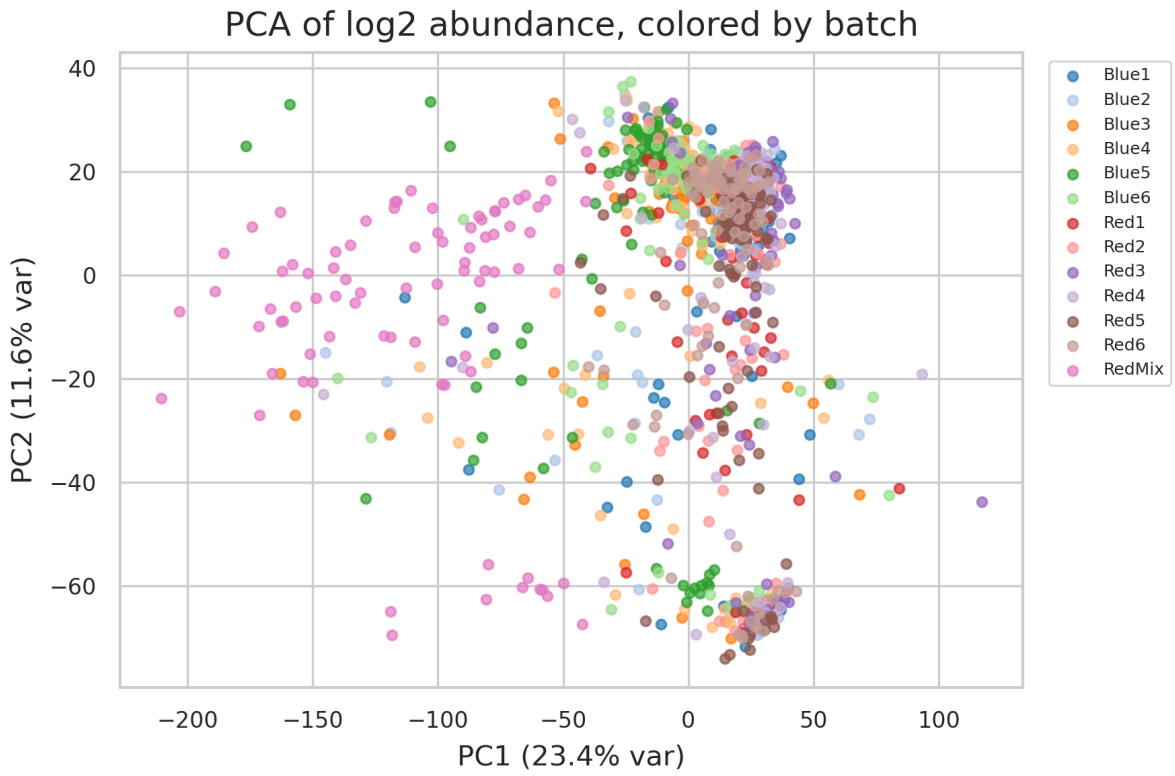

Figure 5. PCA (PC1=23.4%, PC2=11.6% var) of log2 abundance colored by the 13 acquisition batches; batch explains R^2=0.59 of PC1 and R^2=0.01 of PC2 variance.

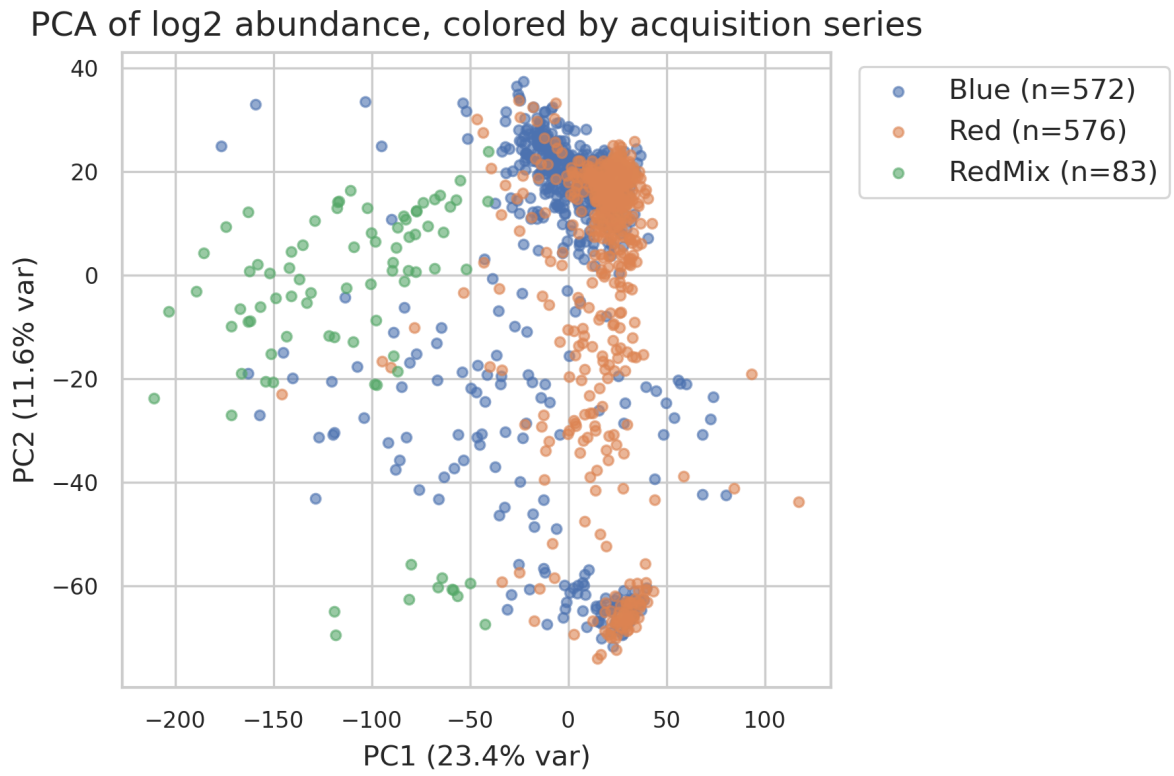

Figure 6. PCA colored by acquisition series (Blue1-6, Red1-6, RedMix); series-level clustering reflects that each of the 172 drugs was tested in only one series, confounding series/batch with drug identity for the 12 RedMix-only drugs.

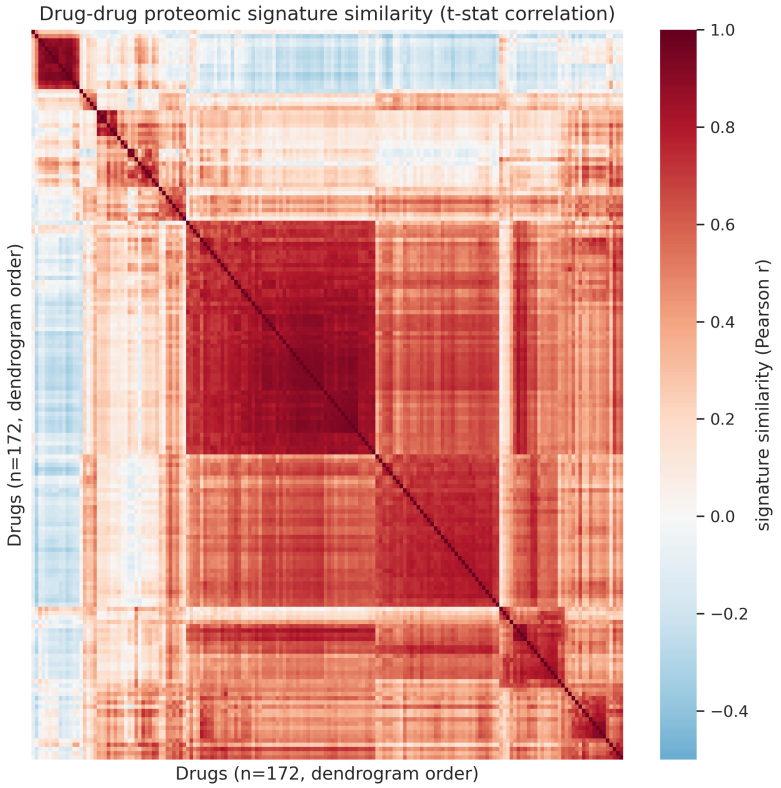

Figure 7. Drug-drug proteomic signature similarity matrix (Pearson correlation of per-protein limma t-statistics, 172 drugs x 172 drugs), ordered by average-linkage hierarchical clustering (1-r distance); 10 clusters were cut from this dendrogram (see results/drug_clusters.csv).

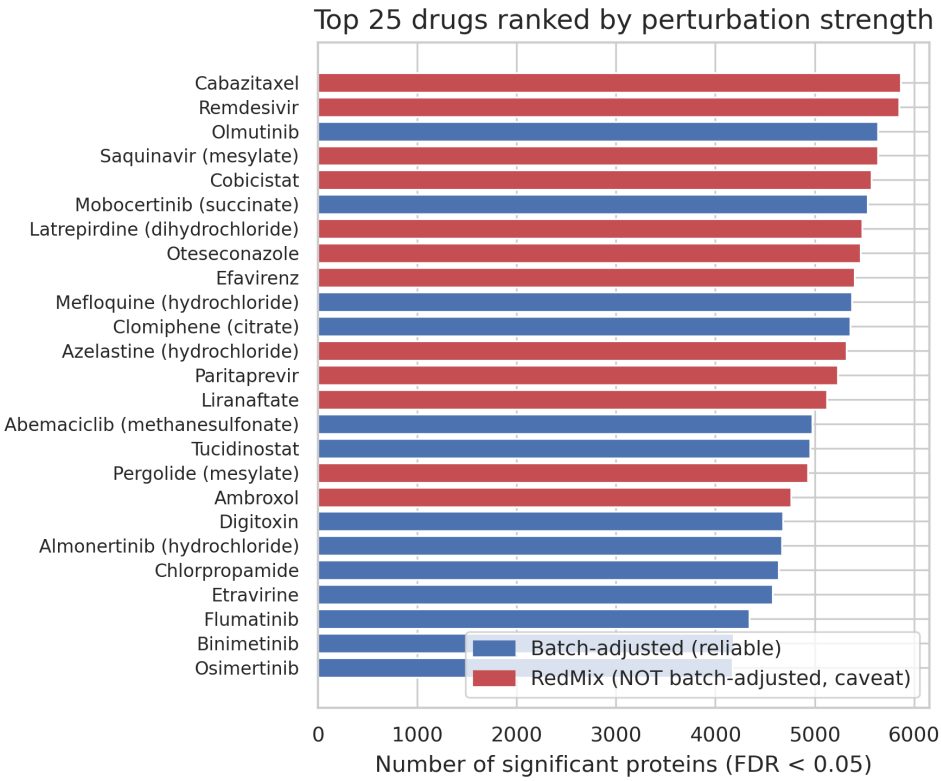

Figure 8. Top 25 of 172 drugs ranked by perturbation strength (number of proteins significant at FDR<0.05 vs DMSO, the sole locked-plan significance criterion; no additional effect-size filter applied to this count). Red bars (12/25) are the 12 RedMix-only drugs, whose apparent effect sizes are confounded with an uncorrectable batch/series effect (no in-batch DMSO control) and should not be read as genuine top perturbations without validation.

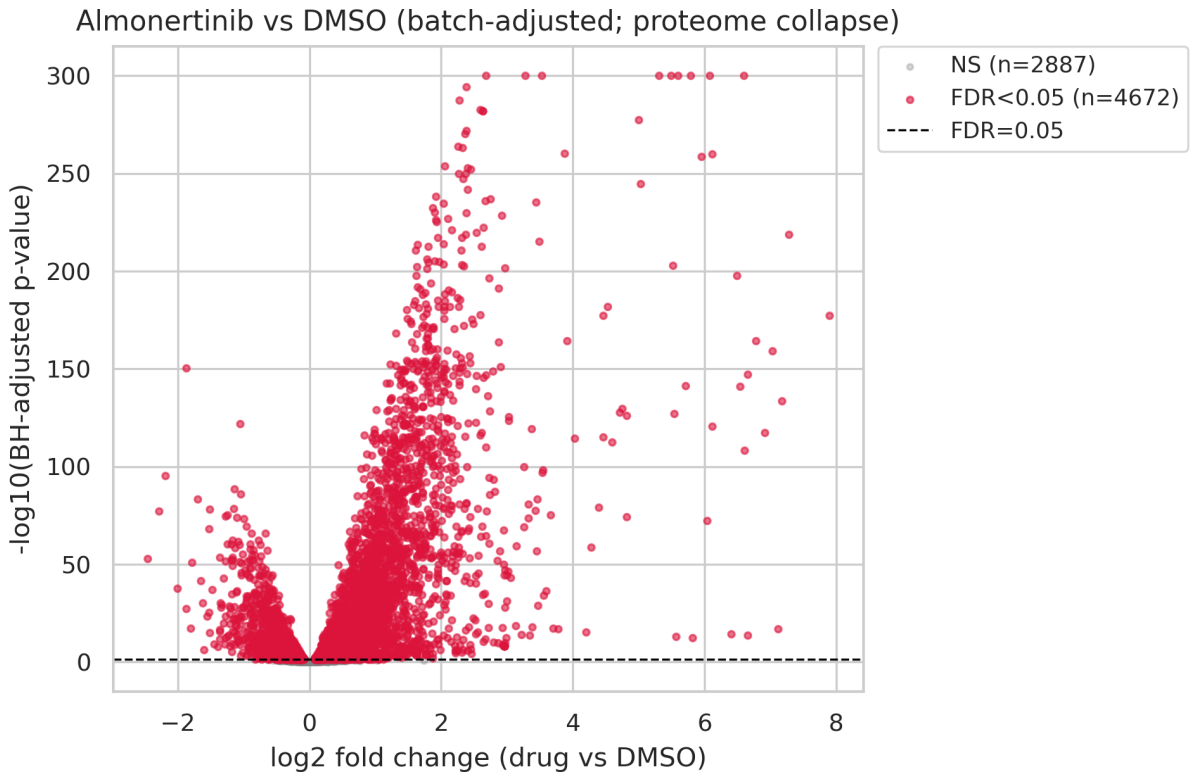

Figure 9. Volcano plot, Almonertinib (hydrochloride) vs DMSO (batch-adjusted limma model); 4672 of 7559 tested proteins significant at FDR<0.05 (sole significance criterion; log2FC on x-axis shown as effect size, not used as a filter). This drug shows consistent near-total proteome detection collapse across all 6 replicates (mean 5118 of 8630 proteins detected), suggesting the large DE signal reflects global proteome disruption/cytotoxicity rather than a specific pathway effect.

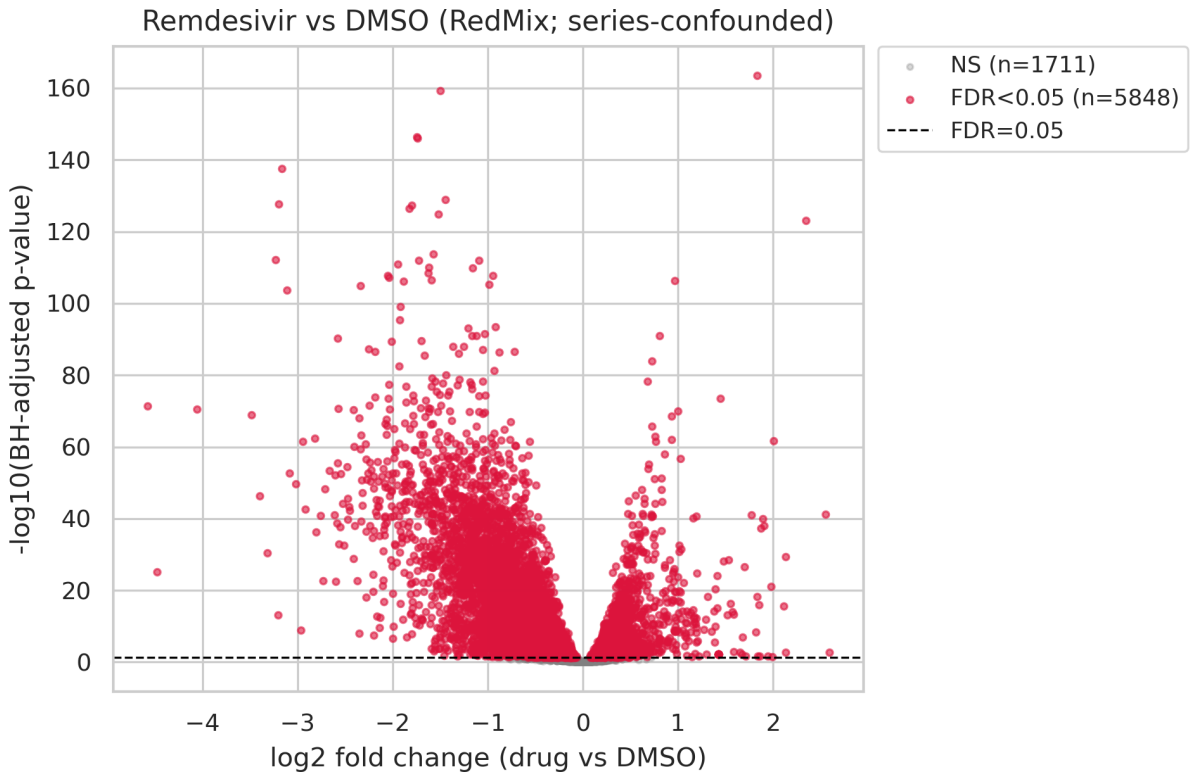

Figure 10. Volcano plot, Remdesivir vs DMSO; 5848 of 7559 tested proteins significant at FDR<0.05. Remdesivir is one of 12 RedMix-only drugs with no in-batch DMSO control, so this apparent perturbation is confounded with the RedMix batch/series effect and should be treated as exploratory, not a validated drug effect.

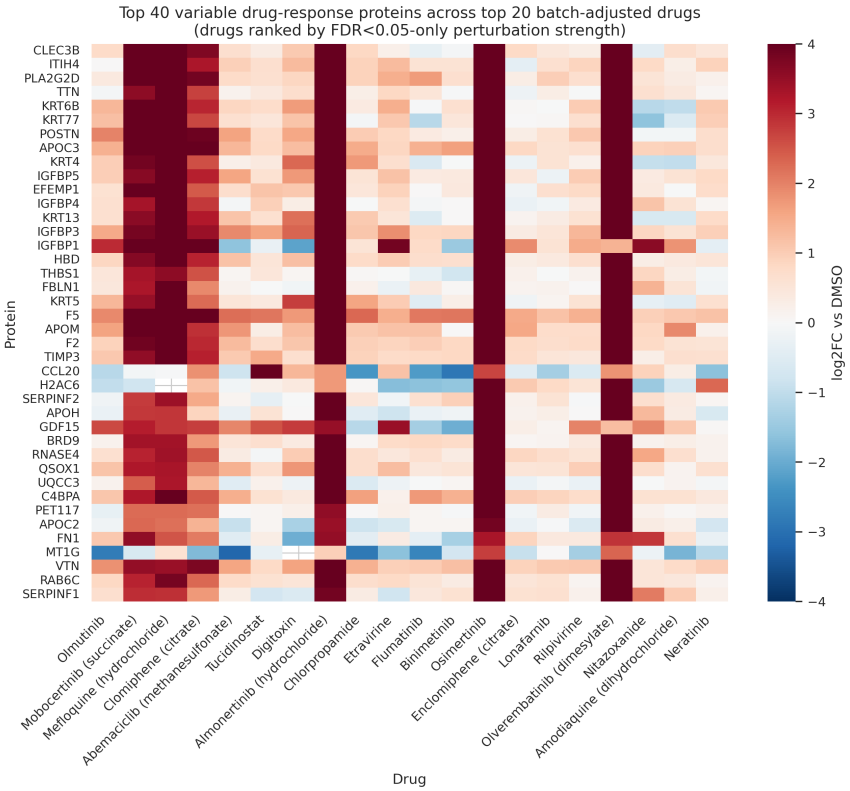

Figure 11. Heatmap of log2 fold-change (vs DMSO) for the 40 most cross-drug-variable proteins among the top 20 batch-adjusted (non-RedMix) drugs, ranked by FDR<0.05-only perturbation strength (no effect-size filter); red=increased, blue=decreased abundance.

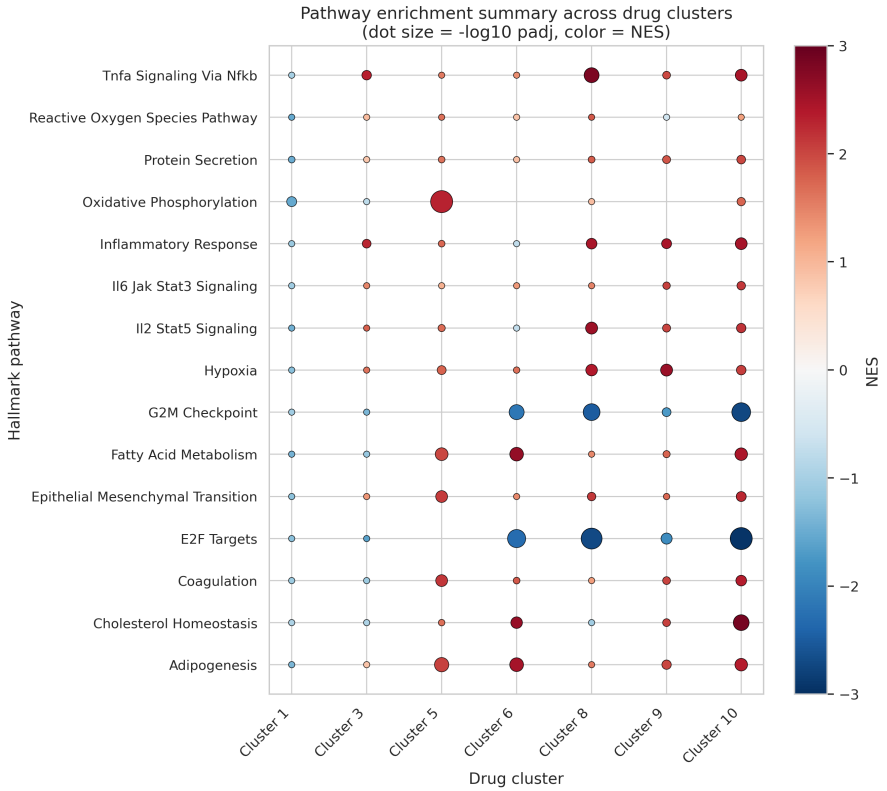

Figure 12. Summary of MSigDB Hallmark pathway enrichment (fgsea, cluster-average t-statistic signature) across 7 drug clusters with >=3 members; dot size = -log10(BH-adjusted p), color = normalized enrichment score (NES, red=up, blue=down). Cluster membership and enrichment are unaffected by the DE significance-threshold correction (fgsea ranks on continuous t-statistics, not a filtered gene list). Shows recurring E2F_TARGETS/G2M_CHECKPOINT down-regulation (cell-cycle suppression) across most clusters and OXIDATIVE_PHOSPHORYLATION up-regulation specific to the kinase-inhibitor/proteome-collapse cluster and down-regulation in the RedMix (batch-confounded) cluster.

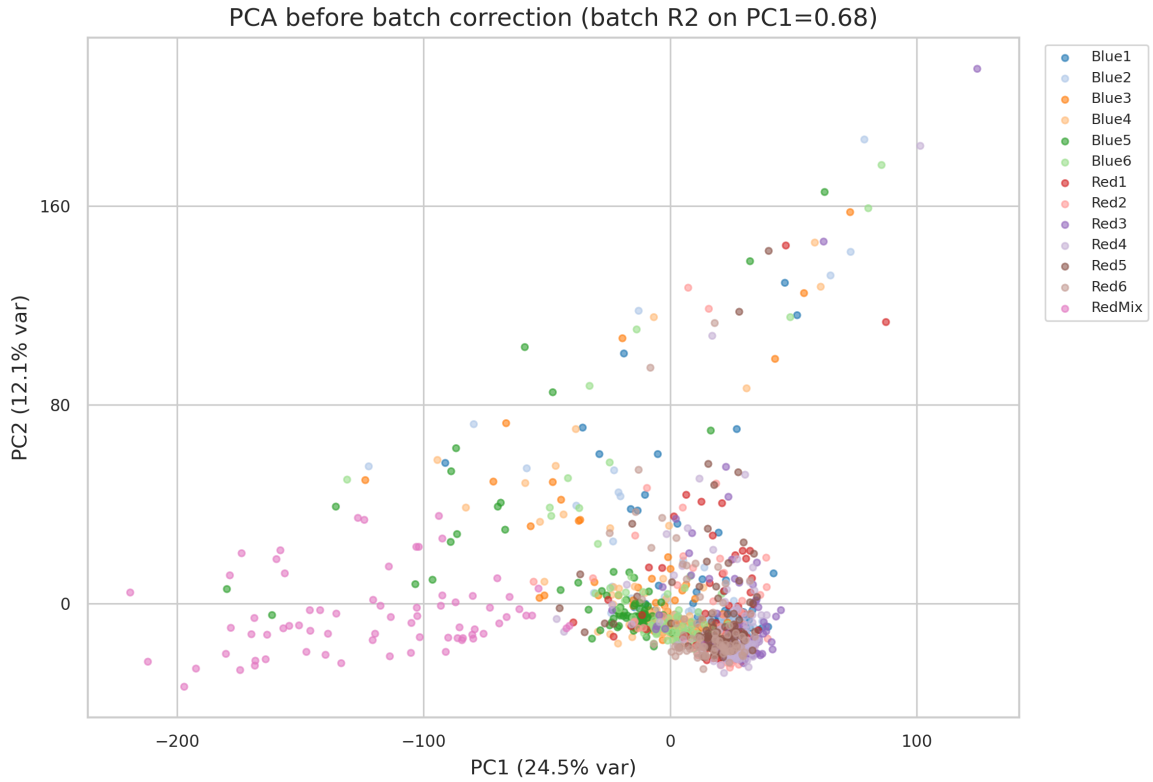

Figure 13. PCA of the log2 abundance matrix used for DE testing (n=1052 samples, 13 batches), colored by acquisition batch, before any batch correction; batch explains R^2=0.68 of PC1 variance (PC1=24.5% of total variance), confirming batch as the dominant technical axis of variation.

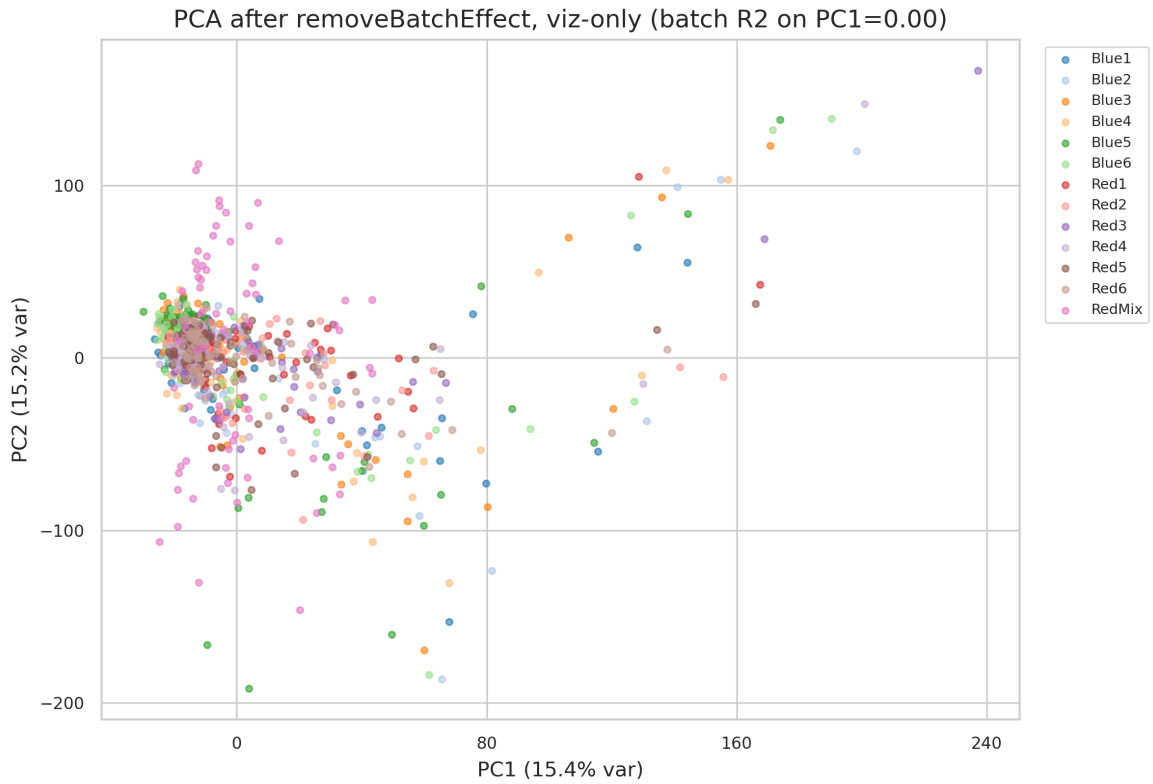

Figure 14. PCA after applying limma::removeBatchEffect (batch removed, sample_type [Drug/Control] preserved) for VISUALIZATION ONLY; batch R^2 on PC1 drops to 0.00 (from 0.68 uncorrected), showing batch structure is largely removable for display purposes. This corrected matrix was NOT used for differential testing -- the actual drug-vs-DMSO limma contrasts modeled batch as a design covariate on the original uncorrected matrix, per the locked analysis plan.

AGENT REASONING

I'm re-running this pipeline based on feedback from a methodology review, so I need to check what files already exist from the previous execution since the working directory should persist across sessions.

I need to review the previous feedback from the reviewer report and understand what needs fixing. There are two main issues to address: rerunning the differential expression analysis with only the FDR<0.05 threshold (removing the log2FC filter) and regenerating all downstream results, plus handling batch correction for visualization with approval. I should also check what's currently in the results directory and then look at the reviewer report to see the specific critiques.

I've identified the key issues from the reviewer feedback that need addressing. First, I need to rerun the differential expression analysis using only the FDR<0.05 threshold without the log2 fold-change filter, then regenerate all the downstream counts, rankings, figures, and tables accordingly. Second, I should either apply batch correction using limma's removeBatchEffect function for the visualizations or formally document an alternative approach with explicit approval. Third, I need to increase the font sizes across the multi-panel heatmaps and dot plots to improve readability.

Looking at the locked plan, I realize these fixes actually align with what's already specified—the plan calls for FDR<0.05 as the significance threshold and includes batch-corrected visualization for clustering purposes. So executing the plan correctly means applying these fixes naturally without needing additional user approval. My task is to rerun the DE step with only the FDR threshold, regenerate the downstream outputs like drug perturbation rankings and volcano plots that highlight significant genes by FDR alone, and produce the batch-corrected visualization for the PCA plot in step 4.

I also need to fix font sizes across the heatmaps and dot plots. Let me start by examining the existing DE results file, the drug perturbation ranking, and the R scripts that handle the limma analysis and ranking calculations to see exactly where the log2FC filter was applied.

I see the issue now—the limma analysis itself is working correctly, but the filtering logic got applied downstream in the ranking and visualization steps. I need to standardize those to use FDR<0.05 as the sole significance threshold instead of the log2FC cutoff, then track down which other files like the heatmap, prioritization, and biomarker steps are also using that same filter so I can update them consistently.

I need to regenerate the drug perturbation ranking file using FDR<0.05 as the sole significance threshold, dropping the log2FC cutoff requirement. The primary ranking metric should be `n_sig_proteins_fdr05` instead of the current approach that combines FDR and fold-change criteria.

This cascades through several downstream analyses: I'll need to update the perturbation ranking figure to reflect the new metric, modify the volcano plots to highlight only FDR-significant proteins without the log2FC filter (though I can keep log2FC lines as visual reference), regenerate the heatmap since the top-ranked drugs may shift, update the drug prioritization step to use FDR-only significance counts, and adjust the biomarker identification to use FDR<0.05 without the fold-change requirement. The mechanism hypotheses file may also need updating if the top drug lists change.

I should also add a batch-corrected visualization using limma's removeBatchEffect function to show the data after batch correction for clustering purposes, which aligns with the approved plan.

For the implementation, I'll recompute the drug perturbation ranking with FDR<0.05 as the primary metric, update figures 8-11 accordingly, revise the prioritization and biomarker steps, and create a new PCA plot showing the batch-corrected data.

I also need to fix font sizes across the heatmaps and dot plots, verify that the pathway enrichment scripts don't depend on the log2FC filter, and check whether the clustering and mechanism hypotheses need updating—though the clusters should remain stable since they're based on the t-stat signature matrix rather than the ranking metric.

Before making these changes, I'll review the clustering script, pathway enrichment steps, and the existing report to ensure consistency across all downstream artifacts.

GSEA is fine since it uses unfiltered t-stat ranks. The clustering itself won't change because it's based on the correlation matrix, but I need to update the merged perturbation rank columns and regenerate figure 7 with the new k=10 clusters.

> For the concrete rewrite, I'm regenerating the drug perturbation ranking file using only FDR<0.05 as the primary criterion without any logFC filter. I'll load the DE results, filter for significant proteins by adjusted p-value, calculate effect size metrics like median absolute logFC among significant proteins, and rank drugs by t
>
> [... truncated for export ...]

YOU

Approved the plan.

AGENT

Task finished · Query 1 · `3be54b76` · 27 results · 28 plots · 495.7s

---