Reversing Tumor–Normal Expression Differences with L1000 Drug Signatures

Georgios Kousis Tsampazis

November 11, 2025

Abstract :

Paired tumor–normal expression differences can reveal indication-specific biology and nominate compounds that reverse disease programs. Using curated GEO microarray cohorts, We formed per-patient signatures (normal-tumor) on L1000 landmark genes and trained a multiclass random forest to predict cancer type. The model achieved high cross-validated performance and its top features mapped to canonical hallmarks (cell cycle, immune, EMT). We then scored L1000 perturbational profiles with the calibrated classifier and aggregated per compound. Shortlists recovered plausible mechanisms (e.g., HDAC, MEK, HSP90; 5-FU) alongside context-mismatched tool compounds, highlighting both promise and caveats of signature reversal.

Index Terms: Cancer-omics, Drug repurposing, L1000.

1 Introduction

Transcriptional signature reversal has emerged as a strategy for therapeutic discovery. Perturbations that oppose disease-associated expression programs can nominate actionable pathways, targets, and candidate compounds (Lamb et al., 2006). The L1000 expansion of the Connectivity Map provides millions of perturbational profiles across diverse cell lines, doses, and time points, enabling systematic disease–perturbation matching at scale (Subramanian et al., 2017). Prior multi-cohort analyses demonstrate that expression-based matching can yield clinically relevant hypotheses and, in some cases, opportunities for drug repurposing (Sirota et al., 2011; Dudley et al., 2011). Public repositories such as GEO supply paired tumor–normal cohorts that ground discovery in human tissue context (Edgar et al., 2002). Given heterogeneity within and across cancers, methodologies that capture non-linear structure complement linear similarity measures, facilitating target nomination and compound prioritization from patient-derived tumor signatures queried against L1000 perturbations.

Objectives

- Construct paired normal-tumor gene-expression signatures from GEO microarray cohorts, restricted to a subset of L1000 landmark genes present in all samples .
- Train and evaluate a nonlinear *multiclass* classifier (random forest) to assign signatures to cancer types, using stratified cross-cohort validation and balanced accuracy.
- Estimate feature importance and map high-importance genes to canonical cancer pathways to contextualize model-derived targets.
- Classify L1000 drug-perturbation signatures with the trained model and aggregate predicted labels and probabilities across conditions to produce compound-level scores.
- Establish best practices for analyzing tumor–normal gene-expression signatures.

2 Data

Paired tumor–normal microarray cohorts spanning multiple cancer types were curated from the NCBI Gene Expression Omnibus (Edgar et al., 2002): esophageal squamous cell carcinoma (GSE38129), gastric adenocarcinoma (GSE65801), lung adenocarcinoma (GSE19804; GSE10072), clear-cell renal cell carcinoma (GSE53757), papillary thyroid carcinoma (GSE33630), head and neck squamous cell carcinoma (GSE6631), colon adenocarcinoma (GSE74602; GSE44076), hepatocellular carcinoma (GSE57957), breast cancer (GSE15852), and a mixed lung cancer cohort (GSE33356). Where a cancer type was represented by multiple cohorts, after assessment of cohort-specificities, the groups were combined and treated as one.

Table 1
Cancer-type counts for curated GEO cohorts.

Cancer type	Count
BRCA	43
COAD	128
ESCC	30
HNSC	22
KIRC	60
LIHC	37
LC/LUAD	153
STAD	32
THCA	44
Total	**549**

BRCA: breast carcinoma; COAD: colon adenocarcinoma; ESCC: esophageal squamous cell carcinoma; HNSC: head and neck squamous cell carcinoma; KIRC: kidney renal clear cell carcinoma; LIHC: liver hepatocellular carcinoma; LC/LUAD: lung cancer/lung adenocarcinoma; STAD: stomach adenocarcinoma; THCA: thyroid carcinoma.

3 Methods

3.1 Preprocessing, gene mapping, and harmonization

For each series, expression matrices and sample annotations were downloaded as provided by GEO. When raw-scale values were detected, data were log_2-transformed as this is considered standard in GEO-processed matrices analysis. Platform probes were mapped to HGNC symbols using the corresponding GPL annotations; when multiple probes mapped to the same gene, a single rep-

resentative was retained using a variance-based rule to preserve the most informative probe within each cohort *(as this unsupervised step does not reference labels, any potential leakage is expected to be negligible)*. To facilitate integration with perturbational profiles, the analysis was restricted to LINCS L1000 landmark genes, further requiring genes to be present in all samples of cohorts. Future imputation attempts may be suited to imrpove methodology.

3.2 Paired tumor–normal signatures

For each matched pair, a per-gene delta was computed as $\Delta_g = x_g^{\text{normal}} - x_g^{\text{tumor}}$, yielding the tumor–normal expression signature $\boldsymbol{\Delta} = \mathbf{x}^{\text{normal}} - \mathbf{x}^{\text{tumor}}$ restricted to landmark genes; signatures from all cohorts were concatenated into a single design matrix with accompanying cancer-type labels.

3.3 Paired normal–tumor signatures are structured by cancer type, but naive normalization distorts deltas

Scaling is significant in signature extraction. Percentile-based scaling (e.g., clipping to the 2nd–98th percentiles) is helpful for visualization but can still distort perceived signatures: by compressing the tails, it downweights high-variance, high-amplitude genes that often carry tumor–normal contrast and pathway signal. Likewise, applying gene-wise z-scoring *before* constructing deltas can change both magnitude and, in some cases, the sign of paired differences. Because the paired design already normalizes many nuisance effects, all inference and modeling use normal–tumor deltas computed on a common normal–tumor deltas.

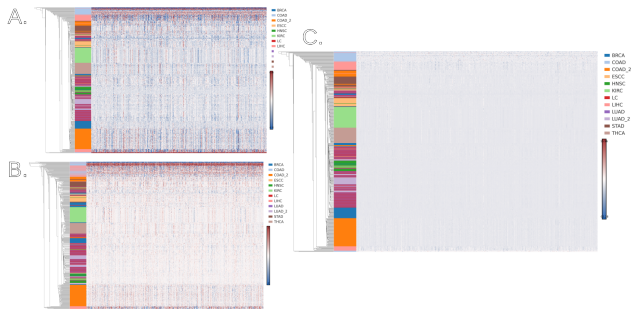


Figure 1. Preprocessing and display choices reshape how paired signatures appear. (**A**) Percentile-scaled heatmap of normal–tumor deltas (color limits set by percentiles) showing visually "tightened" contrasts due to tail compression. (**B**) Gene-wise z-scoring applied *before* delta computation alters dynamic range and can modify relative patterns across pairs, risking distortion of paired signals. (**C**) Baseline $\log_2$-scale tumor–normal deltas without additional scaling, preserving the native magnitude and direction of differences. Percentile and z-scored views are for visualization only; all analyses use baseline $\log_2$ deltas to avoid masking high-variance genes important for signature extraction.

3.4 Classifier training and evaluation

A non-linear multiclass Random Forest classifier (Pedregosa et al., 2011) was trained to assign normal–tumor signatures to cancer type. Model development followed a nested, stratified cross-validation scheme in which the inner loop performed hyperparameter tuning (including tree number, depth, and split/leaf constraints) and the outer loop provided an unbiased estimate of generalization across studies. Class labels were encoded prior to modeling and class imbalance was handled with built-in weighting. To obtain well-calibrated probabilities, the best inner-loop model on each outer training split was calibrated on the training partition only (to avoid leakage) using isotonic regression, with a logistic fallback when isotonic fitting was not feasible; all reported predictions and probabilities on the outer test folds came from these calibrated models. Performance was summarized primarily by balanced accuracy (macro-averaged recall), with macro-F1 and per-class confusion matrices as secondary summaries; we additionally reported a multiclass Brier score and an expected calibration error computed using equal-width binning over the maximum predicted class probability. Finally, a model was refit on the full dataset and then calibrated in the same manner to produce the probabilistic classifier used for downstream analyses.

3.5 Gene importance and pathway context

Gene-level contributions were summarized from the trained forest using model-based importance scores. High-importance genes were mapped to pathway knowledge bases using over-representation analysis (Enrichr via GSEApy; Kuleshov et al., 2016) against MSigDB Hallmark and related curated collections, with Benjamini–Hochberg false discovery rate control. Enriched pathways were used to contextualize model-derived targets in terms of canonical cancer biology.

3.6 L1000 perturbational predictions and compound summarization

Drug perturbation signatures were obtained from the Connectivity Map/LINCS L1000 resource (Lamb et al., 2006; Subramanian et al., 2017) and subset to the same landmark genes used in training. Each L1000 signature s (compound–cell line–dose–time) was classified by the trained multiclass random forest to obtain a probability vector $\mathbf{p}_s = (p_{s,1}, \ldots, p_{s,K})$. For each compound, all of its signatures were pooled and summarized with a best-shot rule: for every cancer type, the highest probability observed across the compound's signatures was taken as its score for that cancer (The per–cancer-type score was the highest probability observed for that cancer across the compound's signatures); a global score was the highest of these (the global score was the maximum of the per-cancer scores). Compounds were ranked by these scores (higher = more consistent with the corresponding tumor class).

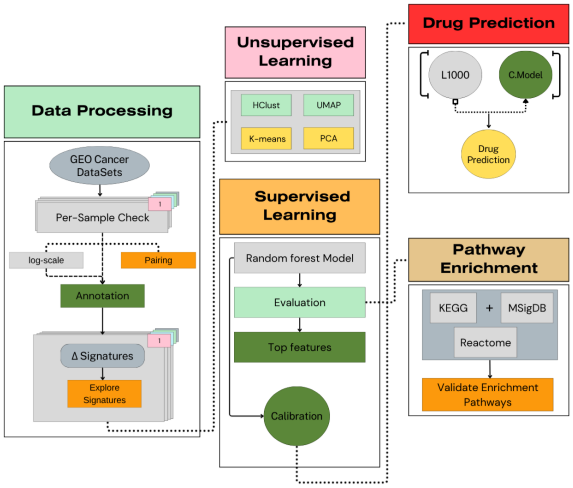


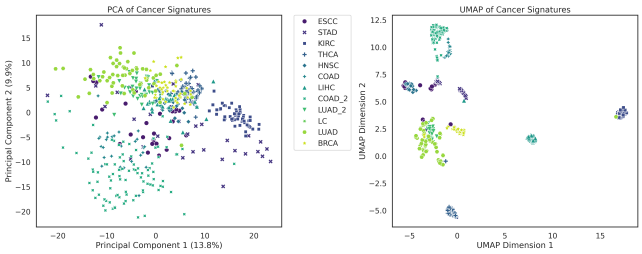
Figure 2. End-to-end analysis pipeline. (**A**) *Data processing:* GEO cancer datasets undergo per-sample checks, log scaling, and normal–tumor pairing with annotation. Paired samples are converted to Δ (normal–tumor) signatures for downstream use. (**B**) *Unsupervised learning:* PCA, UMAP, k-means, and hierarchical clustering (Ward) provide exploratory structure and quality control. (**C**) *Supervised learning:* A random-forest indication classifier is trained on Δ signatures, evaluated, and calibrated; top features are extracted. (**D**) *Pathway enrichment:* Top features are tested against KEGG, MSigDB Hallmark, and Reactome to confirm recovery of canonical cancer pathways, providing orthogonal biological validation of the model. (**E**) *Drug prediction:* L1000 perturbational signatures are scored by the trained classifier and aggregated to rank compounds.
Dotted arrows denote information flow and feedback between modules (e.g., features → enrichment; calibrated model → drug scoring).

4 Results

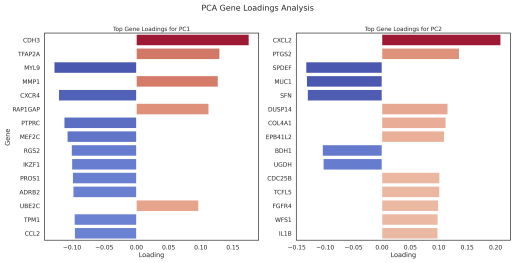
4.1 Low-dimensional structure captures indication-level separation with expected overlap

PCA and UMAP offer complementary views of the paired delta matrix. The first two PCs explain 13.8% and 9.9% of the variance, respectively, and yield only partial separation with many samples collapsing toward the center (Fig. 3A, left). By contrast, UMAP forms clearer neighborhoods (Fig. 3A, right): Lung cohorts (LUAD, LUAD-2, LC) form a tight, shared cluster. COAD and COAD-2 cluster together higher on the map, while LIHC is a separate island. ESCC mostly co-clusters with HNSC (and near STAD), with only a few ESCC drifting toward the lung group, consistent with a squamous/upper-GI axis. KIRC is a separate island (far right). THCA sits just below the lung cluster, and BRCA lies adjacent to the lung group along its right edge.

PCA gene loadings. PC1 is driven by epithelial/proliferative and ECM-remodeling genes (e.g., *CDH3*, *TFAP2A*, *MMP1*, *UBE2C*) and is anticorrelated with immune/mesenchymal markers (e.g., *PTPRC*, *CCL2*, *MYL9*, *IKZF1*). PC2 emphasizes immediate-early/inflammatory and stress–metabolic programs (e.g., *CXCL2*, *IL1B*, *PTGS2*, *COL4A1*) opposed to a secretory/epithelial axis (e.g., *SPDEF*, *MUC1*). Because these cross-cutting programs vary within multiple cancers, tumors with similar pathway activity overlap along the first linear components. UMAP, which preserves local structure, separates samples that share broad programs but differ in indication-specific details, yielding clearer grouping by indication and motivating supervised classifiers downstream.



(a) PCA and UMAP.



(b) Top gene loadings for PC1/PC2.

Figure 3. Low-dimensional views. Paired signatures separate by indication in PCA/UMAP, with loadings and feature–axis correlations implicating gene modules.

4.2 Unsupervised clustering shows only modest cancer-type structure

Distance-based clustering of the Δ matrix using k-means (Euclidean and Manhattan) and hierarchical agglomerative clustering (Ward) showed only modest agreement with indication labels and weak cohesion (ARI 0.42–0.63; silhouette 0.11–0.12; Fig. 4). Ward yielded the best ARI (0.63), but silhouette scores were uniformly low, indicating diffuse/unstable cluster geometry. Overall, some indication signal is present, but global distance–based unsupervised methods capture it poorly; again this motivates using supervised models to learn clearer decision boundaries.

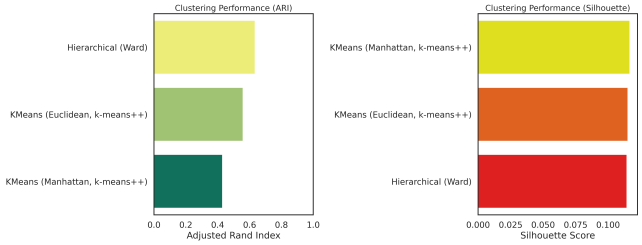


Figure 4. Unsupervised clustering performance. ARI and silhouette scores for alternative algorithms indicate modest alignment between clusters and cancer types.

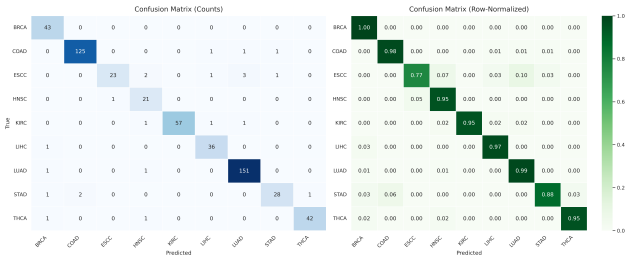
4.3 A multiclass random forest yields stable cross-validated performance

The multiclass random forest achieved high and stable performance under nested cross-validation: overall accuracy = 0.96, weighted F1 = 0.958, macro F1 = 0.937, and mean balanced ac-

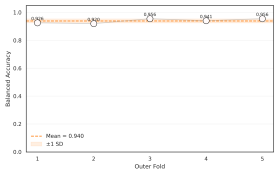
curacy = 0.94 across outer folds (range 0.92–0.956; Fig. 5). Confusion matrices show a strong diagonal with near-perfect recall for BRCA, LC, COAD (≈ 1.00), and high recalls for LIHC (≈ 0.97), HNSC/KIRC/THCA (≈ 0.95), and STAD (≈ 0.88). The main errors occur for ESCC (≈ 0.77 recall), which are chiefly misassigned to LC and, for STAD, partially to COAD, patterns consistent with shared epithelial/gastrointestinal programs. Per-class F1 scores are uniformly high (most ≥ 0.90), with the lowest values for ESCC and HNSC and the highest for COAD and LC. These results indicate that paired normal-tumor signatures carry strong discriminative signal and justify using the trained classifier as a scoring engine for perturbational profiles.

Table 2
Overall classification report (nested stratified CV).

Metric	Precision	Recall	F1	Support
macro avg	0.941	0.937	0.937	549
weighted avg	0.960	0.958	0.958	549



(a) Confusion matrices (counts and row-normalized).



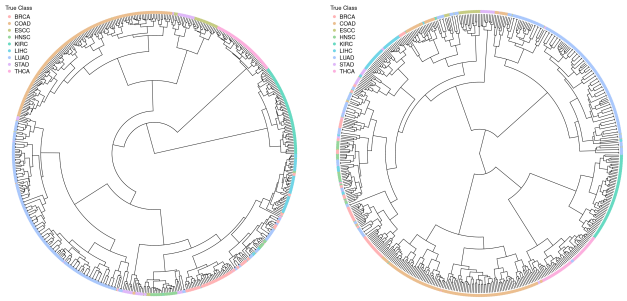
(b) Outer-fold balanced accuracy (nested CV).

(c) Per-class F1 scores.

Figure 5. Supervised performance. Stable cross-validated accuracy and well-structured confusion patterns indicate that tumor–normal signatures are predictive of cancer type.

Table 3
Aggregate performance and calibration metrics.

Metric	Value
Balanced Accuracy	0.9396 ± 0.0147
Brier score (↓)	0.0618
Expected Calibration Error (ECE; ↓)	0.0270



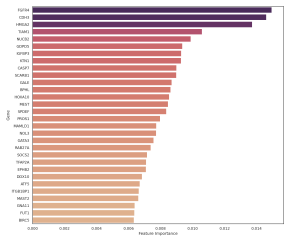
(a) Hierarchical clustering with model-selected features.

(b) Hierarchical clustering with 30 random genes.

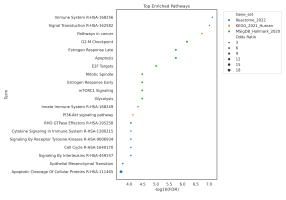
Figure 6. Top features sharpen class structure. Circular dendrograms from hierarchical agglomerative clustering (Ward linkage; Euclidean distance) applied to /delta signatures. The outer color ring encodes the true indication (BRCA, COAD, ESCC, HNSC, KIRC, LIHC, LUAD, STAD, THCA). (**A**) Using the classifier's top-importance genes yields long, contiguous color arcs, indicating purer, indication-specific clusters. (**B**) Using 30 random genes produces fragmented arcs and mixed branches, showing weaker separation. This contrast demonstrates that the model's selected features capture discriminative biology that organizes samples by indication, whereas random genes do not.

4.4 Important genes map to canonical cancer pathways

Enrichment of the highest-importance genes recovered canonical cancer programs. The most significant terms were Reactome *Immune System* and *Signal Transduction*, with *KEGG: Pathways in cancer* among the top hits. Hallmark modules capturing proliferation and hormonal/metabolic control were also enriched (*G2M Checkpoint*, *E2F Targets*, *Estrogen Response* early/late, *mTORC1 Signaling*, *Glycolysis*), alongside *Epithelial Mesenchymal Transition*, *PI3K–Akt*, *Signaling by receptor tyrosine kinases*, *Signaling by interleukins*, *Cytokine signaling in immune system*, *RHO GTPase Effectors*, *Cell Cycle*, and *Apoptotic Cleavage of Cellular Proteins*. These results indicate that the classifier's highest-weight features reflect core tumor biology rather than arbitrary gene sets.



(a) A Top 30 important genes from the random forest (Gini index).



(b) B Pathway enrichment of top features (MSigDB/KEGG/Reactome).

Figure 7. Model-derived biology. (**A**) Highest-importance genes concentrate in signaling, cell-cycle/proliferation, apoptosis/immune, ECM/adhesion, and metabolic modules. (**B**) Bubble plot shows enrichment for Reactome *Immune System* and *Signal Transduction*, *KEGG: Pathways in cancer*, Hallmark *G2M Checkpoint*, and additional Reactome/KEGG terms including *EMT*, *PI3K–Akt*, *Signaling by receptor tyrosine kinases*, *Signaling by interleukins*, and *Cytokine signaling*. Point size encodes odds ratio; color indicates gene-set source.

4.5 Classifier-based scoring of L1000 perturbations

We applied the trained indication classifier to L1000 perturbational signatures and ranked compounds with a best-shot rule (maximum score across signatures per compound), yielding compact, indication-specific lists (Fig. 8). The approach recovered several mechanistically plausible hits: HDAC inhibition (e.g., vorinostat; hydroxamic acids) for ESCC/HNSC; repeated appearance of Aurora-kinase inhibitors (MLN-8054, MK-5108, AZD-1152/Barasertib, GSK-1070916, tozasertib, ENMD-2076) across proliferative indications; MEK pathway targeting (selumetinib) for LUAD/STAD; HSP90 inhibition (geldanamycin) and anthracycline DNA damage (doxorubicin) in COAD; and 5-fluorouracil in ESCC/LIHC. These align with established mechanisms and, for several agents, clinical use in GI and other cancers (e.g., 5-FU in esophageal and colorectal cancer). (Subramanian et al., 2017; Eckschlager et al., 2017; Cicènas et al., 2016; Hedayat et al., 2023; Abdullah et al., 2024; Gmeiner et al., 2023; Li et al., 2017). At the same time, several outputs are questionable or context-mismatched: hormonal agents in non–hormone-driven settings (estradiol in HNSC; estrone in THCA), flutamide in LIHC, high scores for melatonin in LIHC (a compound with largely preclinical or adjunctive evidence rather than standard oncologic use), frequent appearance of broad tool compounds/pan-kinase agents (e.g., staurosporine), and HRAS tool compounds. Such patterns can arise from cell-line/context mismatches in L1000 and from best-shot aggregation favoring strong but nonspecific stress/transcriptional responses. (Karaman et al., 2008; Lim et al., 2021; Fernández-Palanca et al., 2021; Zhou et al., 2019).

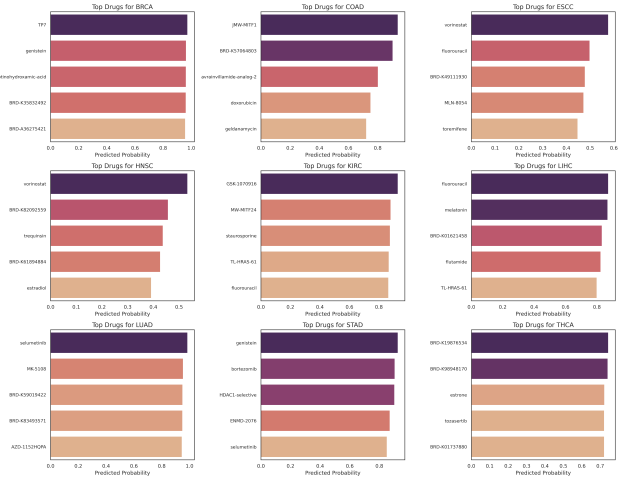


Figure 8. Top predicted compounds per cancer type. Top-5 compounds per indication ranked by the calibrated probability that the compound's L1000 response matches the normal→tumor **Δ** signature for that cancer. Bars show class probabilities from the multiclass model (higher = stronger **Δ**-concordance).

Overall The lists contain both correct (mechanistically and clinically coherent) and questionable (context-incongruent/tool-compound) signals. Those compounds should be treated more like hypothesis-generating, more complex pipeline and analysis should be applied, replicate-aware aggregation (median/trimmed mean), and MoA-level filtering before prioritization. (Subramanian et al., 2017; Lim et al., 2021).

5 Conclusion

This analysis shows that indication signal exists in the **Δ** expression space. UMAP reveals cancer type clustering structure that PCA does not. Unsupervised, global distance–based clustering achieves only modest ARI/silhouette, whereas a supervised classifier learns clearer boundaries and its top features map to canonical cancer pathways. When projected onto L1000, the classifier recovers mechanistically plausible drugs (e.g., HDAC, MEK, HSP90, 5-FU) alongside questionable, context-mismatched hits (tool compounds, broad cytotoxics, hormonal agents in non-hormone tumors).

Supplementary

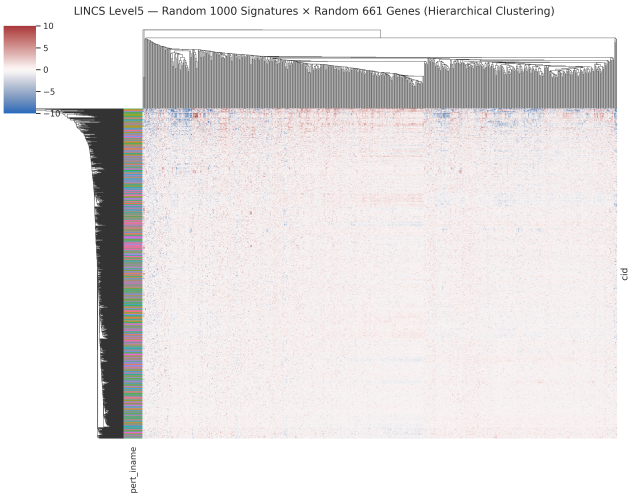


Figure 9. Heatmap of L1000 Level-5 signatures. Hierarchically clustered heatmap of 1,000 randomly selected LINCS L1000 Level-5 signatures by 661 randomly selected genes. Colors denote signed expression changes (Level-5 z-scores; red = up, blue = down). Row/column dendrograms and side annotations (perturbagen name and condition ID) highlight the heterogeneity of transcriptional responses across compounds, cells, and doses. This panel is for visualization only; downstream drug scoring uses all available signatures and genes.

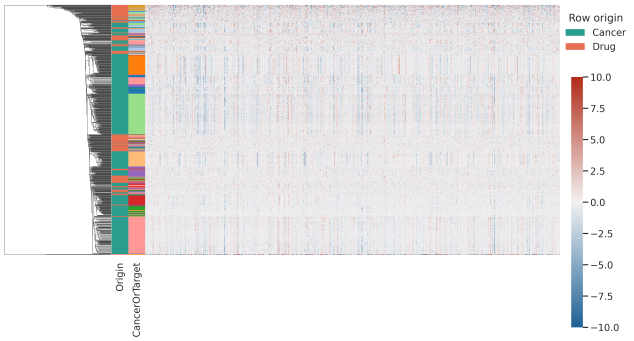


Figure 10. Cancer and predicted-drug signatures co-cluster. Hierarchical clustering of the combined matrix of /delta tumor–normal signatures (*Cancer*) and the top-scoring L1000 Level-5 signatures selected by the classifier (*Drug*). Values are Level-5 z-scores (red = up, blue = down). Row side bars mark the sample origin (teal = Cancer, salmon = Drug) and the cancer/target label. Drug-derived rows are interleaved with cancer rows rather than forming a single "drug-only" block, indicating that predicted perturbational signatures occupy the same expression space as our dataset and do not trivially cluster apart. This sanity check supports compatibility between L1000 and /delta signatures.

Acknowledgements

A project for BC203 is intended as an academic exercise, not a formal research paper. I did this alone but writing with "I" is awkward, so when I say "We" It's like "I"...So when We say:)

Code Availability

All code used in this study is available on GitHub as a Jupyter notebook to facilitate reuse and reproducibility. Nearly all plots were generated directly from this notebook, allowing readers to inspect and rerun the exact code that produced them: `Drugs_repurposing.ipynb`.

References

Abdullah, Omeima and Ziad Omran (Oct. 2024). "Geldanamycins: Potent Hsp90 inhibitors with significant potential in cancer therapy". en. In: *Int. J. Mol. Sci.* 25.20, p. 11293.

Cicènas, Saulius et al. (Dec. 2016). "Maintenance erlotinib versus erlotinib at disease progression in patients with advanced non-small-cell lung cancer who have not progressed following platinum-based chemotherapy (IUNO study)". In: *Lung Cancer* 102, pp. 30–37.

Dudley, Joel T, Tarangini Deshpande, and Atul J Butte (July 2011). "Exploiting drug-disease relationships for computational drug repositioning". en. In: *Brief. Bioinform.* 12.4, pp. 303–311.

Eckschlager, Tomas et al. (July 2017). "Histone deacetylase inhibitors as anticancer drugs". en. In: *Int. J. Mol. Sci.* 18.7, p. 1414.

Edgar, Ron, Michael Domrachev, and Alex E Lash (Jan. 2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". en. In: *Nucleic Acids Res.* 30.1, pp. 207–210.

Fernández-Palanca, Paula et al. (Jan. 2021). "Melatonin as an antitumor agent against liver cancer: An updated systematic review". en. In: *Antioxidants (Basel)* 10.1, p. 103.

Gmeiner, William H and Charles Chidi Okechukwu (Apr. 2023). "Review of 5-FU resistance mechanisms in colorectal cancer: clinical significance of attenuated on-target effects". en. In: *Canc. Drug Resist.* 6.2, pp. 257–272.

Hedayat, Mohaddeseh, Reza Jafari, and Naime Majidi Zolbanin (June 2023). "Selumetinib: a selective MEK1 inhibitor for solid tumor treatment". en. In: *Clin. Exp. Med.* 23.2, pp. 229–244.

Karaman, Mazen W et al. (Jan. 2008). "A quantitative analysis of kinase inhibitor selectivity". en. In: *Nat. Biotechnol.* 26.1, pp. 127–132.

Kuleshov, Maxim V et al. (July 2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". In: *Nucleic Acids Res.* 44.W1, W90–W97.

Lamb, Justin et al. (Sept. 2006). "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease". en. In: *Science* 313.5795, pp. 1929–1935.

Li, Zengyun et al. (May 2017). "Cisplatin-based chemoradiotherapy with 5-fluorouracil or pemetrexed in patients with locally advanced, unresectable esophageal squamous cell carcinoma: A retrospective analysis". en. In: *Mol. Clin. Oncol.* 6.5, pp. 743–747.

Lim, Nathaniel and Paul Pavlidis (Sept. 2021). "Evaluation of connectivity map shows limited reproducibility in drug repositioning". en. In: *Sci. Rep.* 11.1, p. 17624.

Pedregosa, Fabian et al. (Nov. 2011). "Scikit-learn: Machine Learning in Python". In: *J. Mach. Learn. Res.* 12.null, pp. 2825–2830. ISSN: 1532-4435.

Sirota, Marina et al. (Aug. 2011). "Discovery and preclinical validation of drug indications using compendia of public gene expression data". en. In: *Sci. Transl. Med.* 3.96, 96ra77.

Subramanian, Aravind et al. (Nov. 2017). "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles". en. In: *Cell* 171.6, 1437–1452.e17.

Zhou, Bei et al. (July 2019). "Melatonin increases the sensitivity of hepatocellular carcinoma to sorafenib through the PERK-ATF4-Beclin1 pathway". en. In: *Int. J. Biol. Sci.* 15.9, pp. 1905–1920.