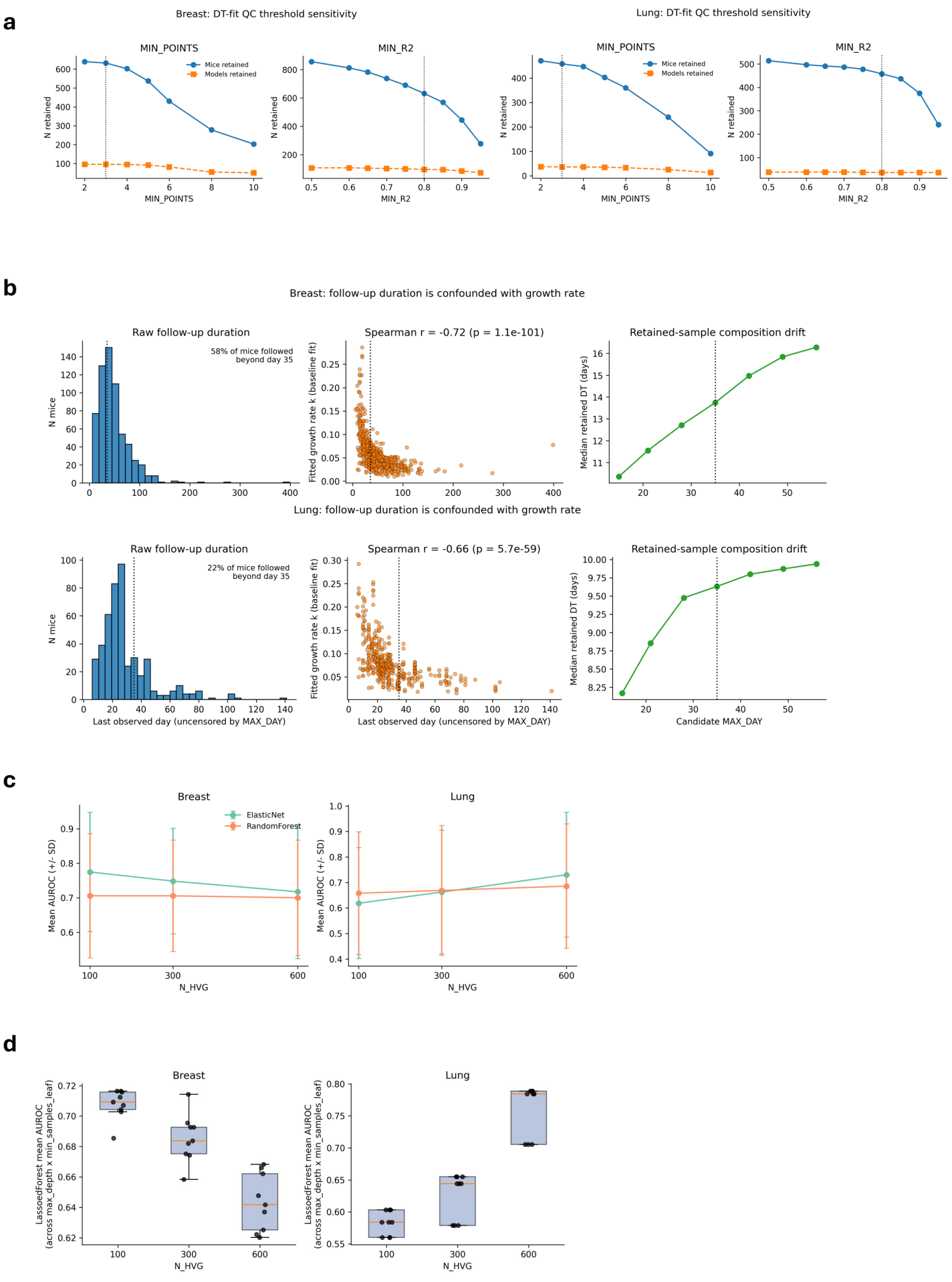
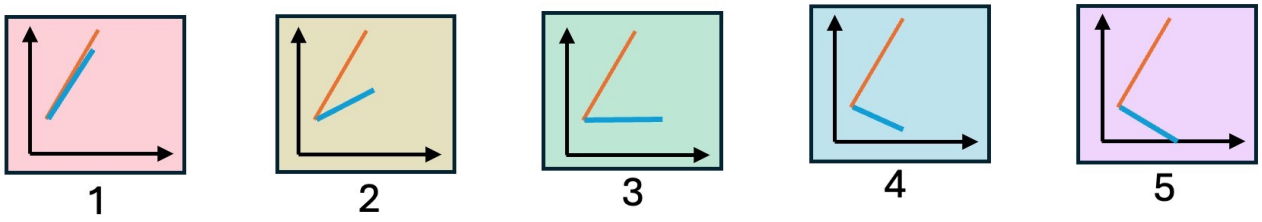

a
Breast: DT-fit QC threshold sensitivity
MIN_POINTS
MIN_R2
Mice retained
Models retained
N retained
MIN_POINTS
MIN_R2
Lung: DT-fit QC threshold sensitivity
MIN_POINTS
MIN_R2
Mice retained
Models retained
N retained
MIN_POINTS
MIN_R2
b
Breast: follow-up duration is confounded with growth rate
Raw follow-up duration
58% of mice followed beyond day 35
N mice
Spearman r = -0.72 (p = 1.1e-101)
Fitted growth rate k (baseline fit)
Retained-sample composition drift
Median retained DT (days)
Lung: follow-up duration is confounded with growth rate
Raw follow-up duration
22% of mice followed beyond day 35
N mice
Last observed day (uncensored by MAX_DAY)
Spearman r = -0.66 (p = 5.7e-59)
Fitted growth rate k (baseline fit)
Last observed day (uncensored by MAX_DAY)
Retained-sample composition drift
Median retained DT (days)
Candidate MAX_DAY
c
Breast
ElasticNet
RandomForest
Mean AUROC (+/- SD)
N_HVG
Lung
Mean AUROC (+/- SD)
N_HVG
d
Breast
LassoedForest mean AUROC (across max_depth x min_samples_leaf)
N_HVG
Lung
LassoedForest mean AUROC (across max_depth x min_samples_leaf)
N_HVG

Supplementary Figure 1. Sensitivity of the growth-kinetics analysis (case study 1) to quality-control thresholds, follow-up window, and highly-variable-gene count.
(a) Number of mice and PDX models retained across a grid of candidate MIN_POINTS (minimum tumour-volume measurements per mouse) and MIN_R2 (minimum goodness-of-fit) thresholds, for the breast (left) and lung (right) cohorts; dotted lines indicate the thresholds used in the primary analysis (MIN_POINTS = 3, MIN_R2 = 0.8). (b) For the breast (top) and lung (bottom) cohorts: distribution of raw, uncensored follow-up duration per mouse (left); correlation between fitted growth-rate coefficient k and follow-up duration (two-sided Spearman r annotated; middle); and drift in median retained doubling time as a function of the candidate follow-up-window cutoff, MAX_DAY (right). Dotted line, the 35-day cutoff used in the primary analysis. (c) Cross-validated AUROC (mean ± SD) for ElasticNet and Random Forest across a grid of candidate highly-variable-gene (HVG) counts, breast (left) and lung (right). (d) Distribution of Lassoed Forest mean AUROC across the max_depth × min_samples_leaf hyperparameter grid, across the same HVG-count grid, breast (left) and lung (right).

a Drug sensitivity categorization criteria provided to panel of experts



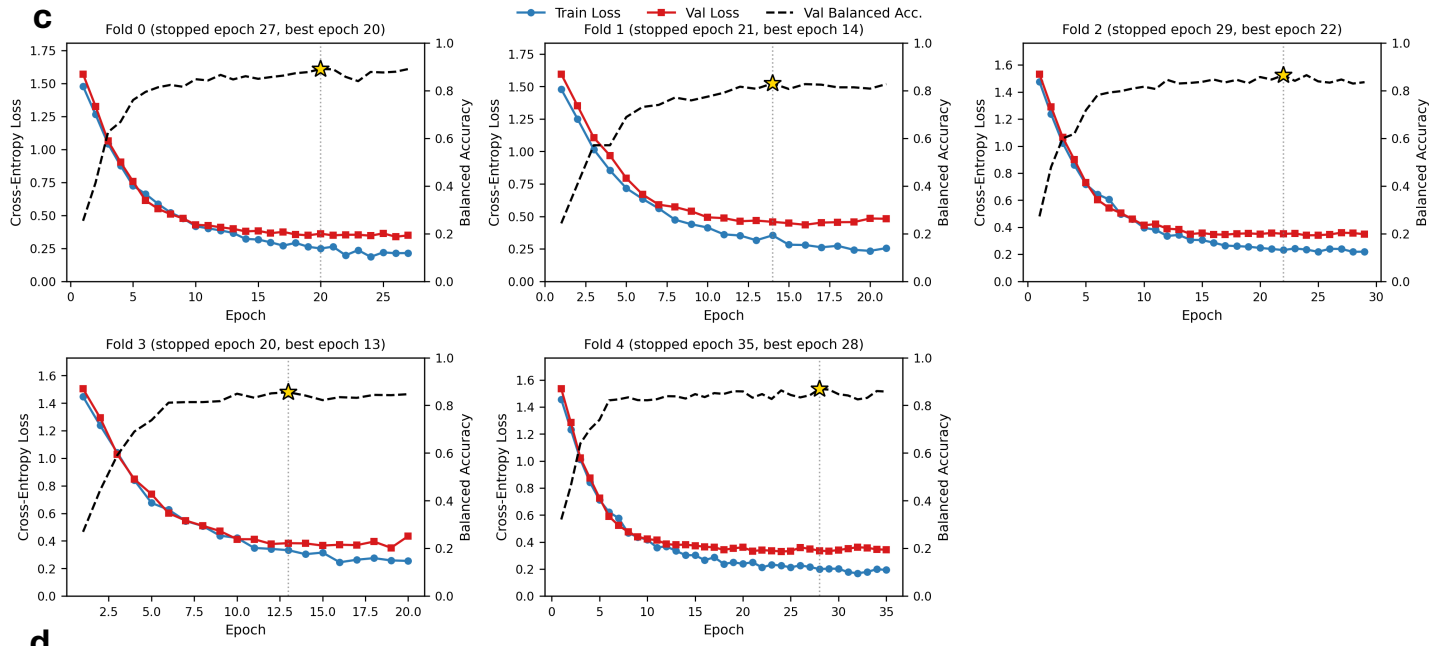
Category 1: tumour growth **similar to control**.
Category 2: tumour growth is **inhibited compared to control**, but tumour is still **enlarging** compared to time 0.
Category 3: tumour growth is **inhibited compared to control**, and tumour volume stays **stable**.
Category 4: tumour growth is **inhibited compared to control**, and tumour volume is **shrinking**.
Category 5: tumour growth is **inhibited compared to control**, and tumour **disappeared** during treatment.

b

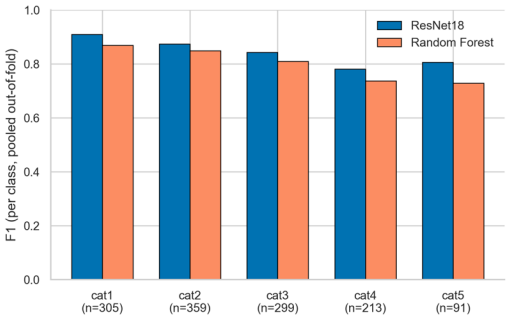


	Cat_1	Cat_2	Cat_3	Cat_4	Cat_5	Total
Expert 1	233	435	282	250	100	1300
Expert 2	255	423	362	233	24	1300
Expert 3	387	289	300	174	139	1300
Consensus	305	359	299	213	91	1267

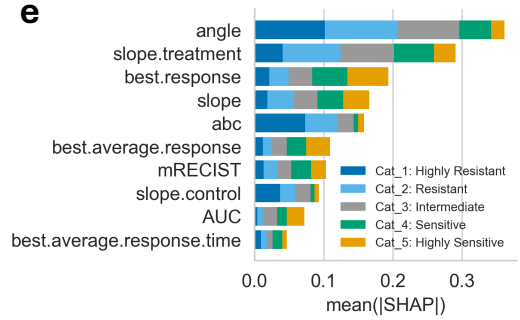
c



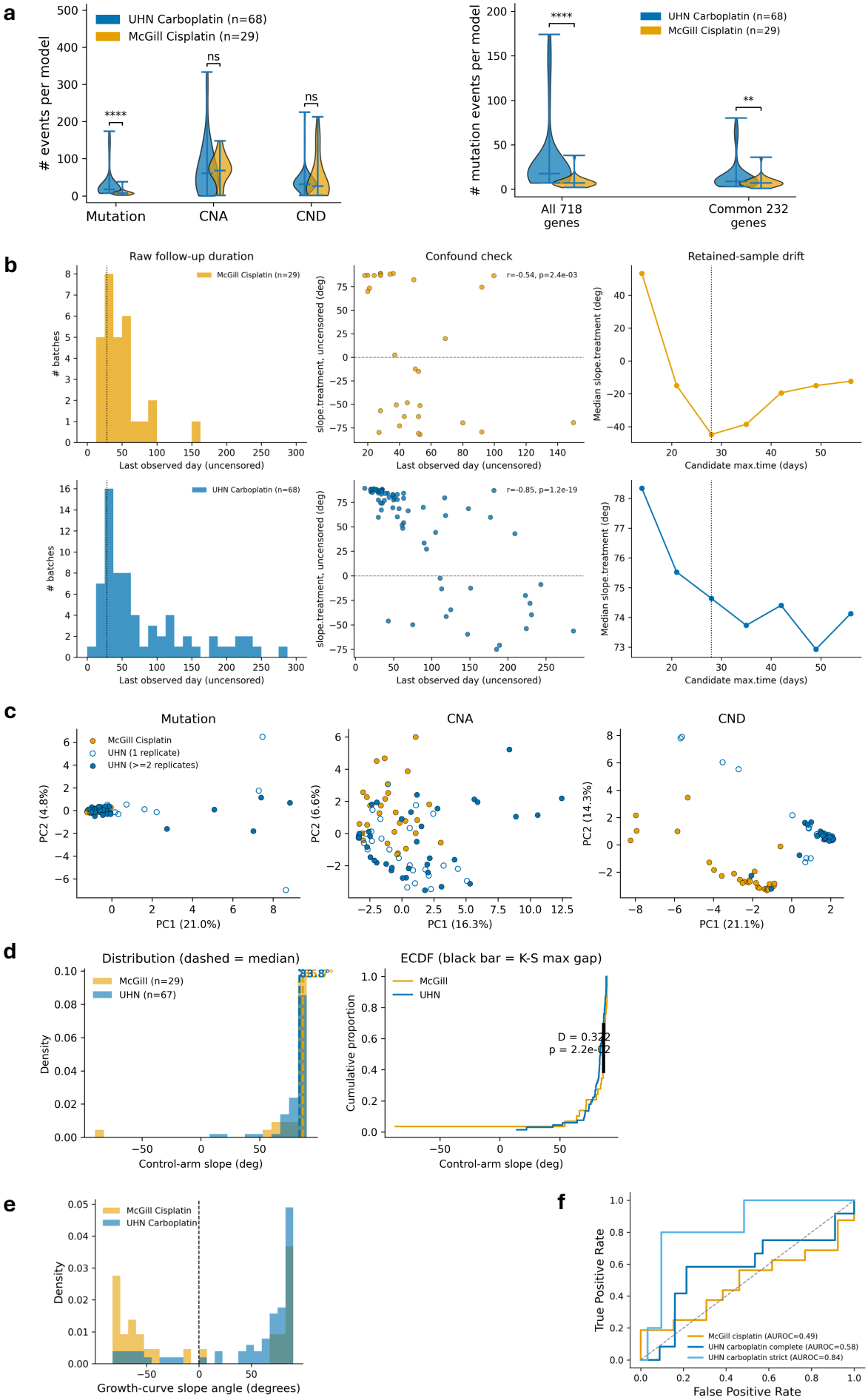
d



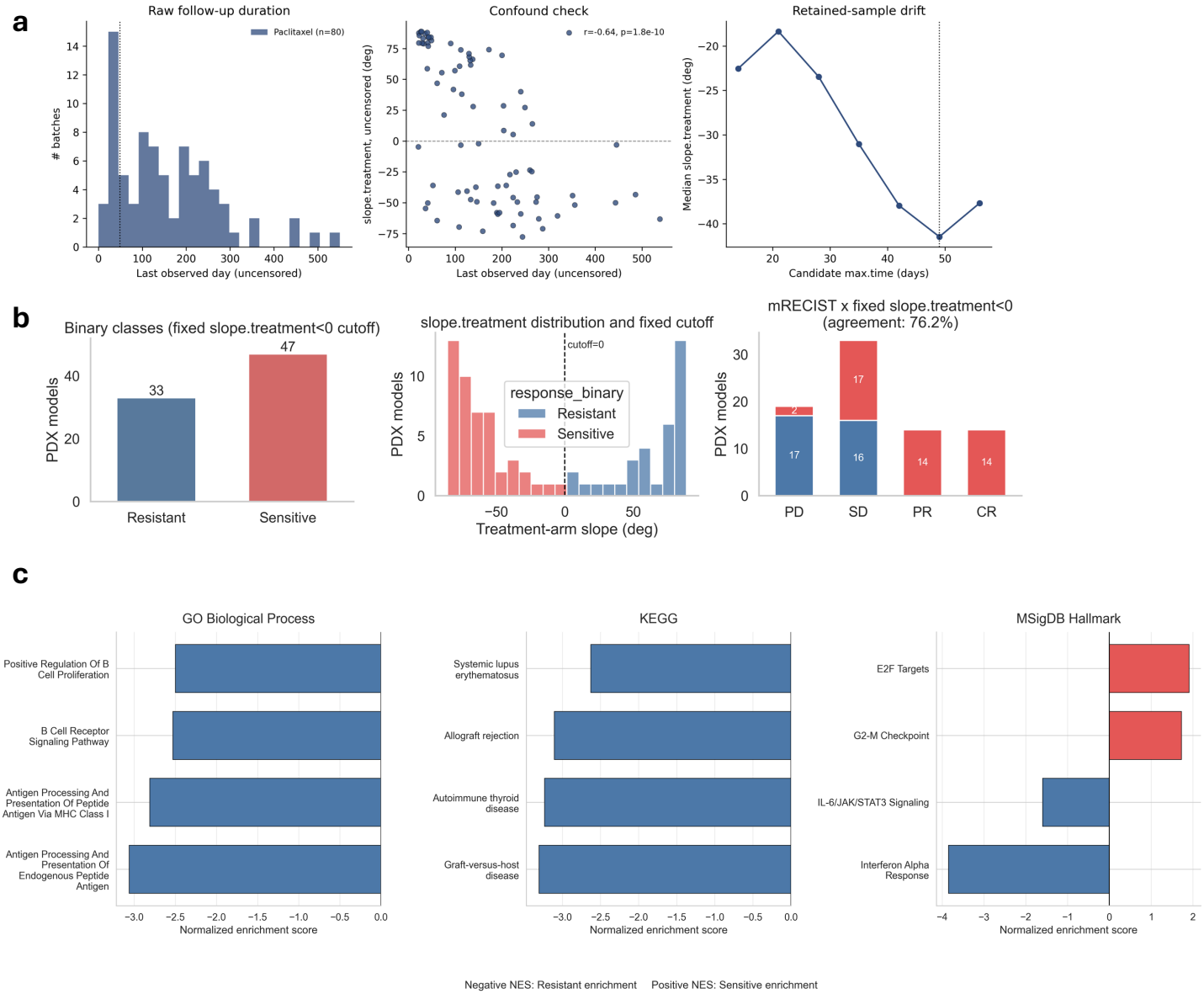
e



Supplementary Figure 2. ResNet18 training curves across cross-validation folds (case study 2).
(a) Drug sensitivity categorization criteria provided to panel of experts. (b) Rating from expert 1-3, and the consensus (majority voting) for each drug category. (c) Training/validation cross-entropy loss (left axis) and validation balanced accuracy (right axis) across epochs for each of the five outer cross-validation folds (Fold 0–4) used in the image-based drug-response classification; the checkpoint with the best validation balanced accuracy (star) in each fold was reloaded and used for evaluation on that fold's held-out test partition. (d) F1 score per category comparing ResNet18 and Random Forest models. (e) Random Forest feature importance (mean absolute SHAP value, stacked by response category) for the top 10 of thirteen engineered response features, ranked by total mean absolute SHAP across categories.


a
UHN Carboplatin (n=68)
McGill Cisplatin (n=29)
events per model
****
ns
ns
Mutation
CNA
CND
mutation events per model
****
**
All 718 genes
Common 232 genes
b
Raw follow-up duration
Confound check
Retained-sample drift
batches
Last observed day (uncensored)
slope.treatment, uncensored (deg)
r=-0.54, p=2.4e-03
Median slope.treatment (deg)
Candidate max.time (days)
r=-0.85, p=1.2e-19
c
Mutation
CNA
CND
McGill Cisplatin
UHN (1 replicate)
UHN (>=2 replicates)
PC1 (21.0%)
PC2 (4.8%)
PC1 (16.3%)
PC2 (6.6%)
PC1 (21.1%)
PC2 (14.3%)
d
Distribution (dashed = median)
McGill (n=29)
UHN (n=67)
Density
Control-arm slope (deg)
ECDF (black bar = K-S max gap)
McGill
UHN
Cumulative proportion
D = 0.322
p = 2.2e-02
e
McGill Cisplatin
UHN Carboplatin
Density
Growth-curve slope angle (degrees)
f
True Positive Rate
False Positive Rate
McGill cisplatin (AUROC=0.49)
UHN carboplatin complete (AUROC=0.58)
UHN carboplatin strict (AUROC=0.84)

Supplementary Figure 3. Cohort-comparability and robustness analyses for NeST-VNN cross-cohort validation (case study 3).
(a) Per-model event burden across the full 718-gene NeST-VNN panel: number of mutation, copy-number-amplification (CNA), and copy-number-deletion (CND) events per model, UHN carboplatin vs. McGill cisplatin (left; two-sided Mann–Whitney U test; ****, $p < 0.0001$; ns, not significant). Right, sensitivity check for the Mutation comparison: the same test repeated on only the 232 genes with recorded mutation data in both cohorts, alongside the original all-718-gene comparison (****, $p = 4.02\times10^{-11}$; **, $p = 5.09\times10^{-3}$) — the cohort gene-coverage gap (Fig. 4c) inflates the difference but does not fully explain it. (b) For the McGill cisplatin (top) and UHN carboplatin (bottom) cohorts: distribution of raw, uncensored treatment-arm follow-up duration per batch (left); correlation between the uncensored treatment-arm slope and follow-up duration (two-sided Spearman r annotated; middle); and drift in median treatment-arm slope as a function of the candidate response-window cutoff, max.time (right). Dotted line, the 28-day cutoff used in the primary analysis. (c) Principal component analysis (PC1 vs. PC2) of binarized mutation, copy-number-amplification (CNA), and copy-number-deletion (CND) feature matrices, McGill cisplatin vs. UHN carboplatin (open circles, one replicate; filled circles, ≥2 replicates). (d) Control-arm growth-curve slope distribution (left) and empirical CDF with Kolmogorov–Smirnov statistic (right), McGill vs. UHN. (e) Treatment-arm growth-curve slope distribution, McGill cisplatin vs. UHN carboplatin; dashed line, the 0° growth/regression threshold used to define response in the primary analysis. (f) ROC curves for the pretrained 5-fold NeST-VNN ensemble (fold-averaged prediction) against sensitive/resistant status (slope.treatment < 0), per cohort; AUROC annotated per curve. Related to Fig. 4a–e.



Supplemental Figure 4. Follow-up window selection and full pathway-enrichment detail for the paclitaxel case study (case study 4).

(a) Justification for the max.time = 49-day follow-up window used to compute slope.treatment. Left, distribution of raw (uncensored) last-observed day across the 80 paclitaxel batches; dotted line marks the chosen max.time = 49-day cutoff. Middle, uncensored slope.treatment vs. last-observed day, showing a significant negative correlation ($r = -0.64$, $p = 1.8 \times 10^{-10}$) — batches followed longer trend toward more negative (regressing) slopes, so an uncapped follow-up window confounds duration with apparent response. Right, median slope.treatment among the retained sample as a function of candidate max.time cutoffs (14–56 days, the same sweep range used in case study 3); the deepest point (most negative median) occurs at day 49 (dotted line), the value chosen for the analysis. (b) Definition of the binary response endpoint used for classification. Left, number of PDX models called Resistant vs. Sensitive under the fixed slope.treatment < 0° cutoff (n = 33 vs. 47). Middle, distribution of treatment-arm slope colored by the resulting binary call; dashed line, the 0° cutoff — the two groups separate by construction, since the label is defined directly from this threshold. Right, cross-tabulation against the categorical mRECIST call (PD/SD/PR/CR; PD and SD collapsed to Resistant, PR and CR to Sensitive for comparison); the two response definitions agree for 76.2% of models (61/80). (c) GSEA results against GO Biological Process, KEGG, and MSigDB Hallmark gene sets, ranked by limma moderated t-statistics (Sensitive vs. Resistant, fixed slope.treatment<0 cutoff, full rna_all transcriptome). Positive NES: enriched in Sensitive models; negative NES: enriched in Resistant models (FDR q < 0.05) .