Supplementary Information

Network-guided discovery and virtual perturbation identify clinically associated gene modules in acute myeloid leukemia

Supplementary Methods

A1 netLDA network modeling and age–gene reweighting

This Supplementary Information provides detailed methods, supporting results, and sensitivity analyses for the AML KPRM study. The original frozen BEAT2 model is the biological reference; patient-grouped refits provide a separate robustness assessment. Numbered references correspond to the manuscript bibliography.

A1.1 Marginal effects and construction of local gene networks

Let X denote expression of the 660 candidate genes from the PI3K–Akt, JAK–STAT, MAPK, NF-κB, and Wnt KEGG pathways. Under the one-year survival classification definition, Y = 0 denotes death at or before day 365 and Y = 1 denotes survival beyond 365 days; samples censored at or before day 365 were excluded. Expression was first centered and scaled separately within each cohort before endpoint filtering. Within netLDA fitting, these inputs were further centered and divided by the sample standard deviation estimated in the BEAT2 fitting partition; constant predictors were excluded. For gene j, the marginal effect δ_j and class midpoint m_j were defined from the standardized class means:

$$\delta_j = \mu_{0j} - \mu_{1j}, \qquad m_j = (\mu_{0j} + \mu_{1j})/2.$$

The implementation divides the class-mean difference by the overall standard deviation of the standardized gene, which equals one up to numerical precision. Available genes in the union of the literature-informed and ML-selected hub sets were ranked by $|\delta_j|$ in the full BEAT2 training dataset, with gene name resolving ties. Their column indices were supplied as strong.X.set.

Cross-validation within BEAT2 selected τ, ν, and α. The final model used up to 20 ranked hub genes (τ = 20) and up to 200 selected genes in total, including hub genes and network-recruited genes (ν = 200). A correlation threshold of 0.50 (α = 0.50) defined the connection topology. Neighborhood searches considered up to the 10 strongest connected neighbors (nb = 10) and extended through three connection layers (d = 3), corresponding to paths of up to three edges from the target gene. The neighborhood-search parameters d and nb were fixed at their default values; varying them in sensitivity checks did not change the prediction results. Covariance blocks and generalized inverses within these neighborhoods produced the gene inverse-covariance operator Ω.

The selected-gene graph was reconstructed directly from the returned network.Omega matrix, preserving gene order and checking agreement with the saved discriminant coefficients. For $j \neq k$, the normalized inverse weight was $w_{jk} = -\Omega_{jk}/\sqrt{(\Omega_{jj}\Omega_{kk})}$. An undirected edge was retained when $|w_{jk}| > 10^{-8}$, a numerical tolerance. The resulting edges represent fitted inverse-covariance associations.

A1.2 Network-adjusted predictive effects and hidden co-regulators

The gene discriminant coefficient vector is $\beta=\Omega\delta$. Its decomposition separates a diagonal contribution from an off-diagonal network contribution:

$$\beta_j = \Omega_{jj}\delta_j + \sum_{k\neq j} \Omega_{jk}\delta_k.$$

A gene with a small marginal effect can acquire a substantial final coefficient through its dependence on other predictive genes. The network term can also oppose the diagonal term, reducing or reversing the final coefficient. Hub and network-recruited labels distinguish how genes entered the model. Hidden co-regulator candidates were characterized using the marginal and network criteria below.

Welch two-sample tests compared the survival classes for every usable candidate gene, with Benjamini–Hochberg correction across the complete candidate universe. A selected gene was annotated as a network-supported weak-signal candidate when its marginal FDR was ≥ 0.05, its graph degree was positive, and its absolute off-diagonal contribution exceeded 10^{-8}. Network-recruited genes meeting these criteria were termed hidden co-regulator candidates.

A1.3 Discriminant rule and final model specification

Age was standardized using the BEAT2 training mean c_a and sample standard deviation s_a, so that $a = (\text{age} - c_a)/s_a$. The age coefficient γ and class midpoint m_a were computed as follows:

$$\gamma = [E(a|Y=0) - E(a|Y=1)]/\text{Var}(a),$$

$$m_a = [E(a|Y=0) + E(a|Y=1)]/2.$$

Var(a) denotes the overall training-sample variance of standardized age. For standardized selected-gene expression x, the linear discriminant score was

$$D(x,a) = \sum_j \beta_j(x_j - m_j) + \gamma(a - m_a).$$

The mLDA.pair call used pair = c(0, 1): $D \geq 0$ assigned class 0 and $D < 0$ assigned class 1. The continuous score $-D$ was therefore used to calculate class-1 AUC. This centered Fisher discriminant rule contains no empirical log class-prior odds term. Reconstructed scores were checked against the returned FisherDR before the model was frozen.

The final base model used the available ranked union hubs as strong.X.set, with $\tau = 20$, $\nu = 200$, $\alpha = 0.50$, $d = 3$, and nb = 10. Standardized age was supplied as Z and Z.new. Saved outputs included gene and age coefficients, class midpoints, training centers and scales, selected genes, retained hub genes, fitting arguments, and the full library result. After the cohort-specific expression preprocessing described in Section A4, external prediction used these frozen BEAT2 transformations, coefficients, and decision threshold.

Age entered as a separate additive component without age–gene interactions; network.Omega represents the expression-based gene network.

A1.4 Definition and construction of predictive regulatory modules

A predictive regulatory module was defined as a connected component containing at least two selected genes in the graph reconstructed from Ω. These components were termed KPRMs because the candidate genes originated from KEGG pathways. Components were identified from the fitted graph, and genes without retained edges were reported as singletons. KPRMs have disjoint gene membership, although a module can span several KEGG pathways.

For each KPRM, we summarized gene membership, hub-gene and network-recruited-gene composition, edges, normalized inverse weights, and KEGG annotations. The representative anchor was the hub gene with the largest absolute marginal effect within the component; for a component without a hub gene, the selected gene with the largest absolute marginal effect would be used. In pathway-composition summaries, a gene annotated to k pathways contributed 1/k to each pathway, giving a total contribution of one per gene.

The final graph contained 200 genes: 20 hub genes and 180 network-recruited genes. Seven multi-gene components contained 64, 51, 44, 17, 12, 3, and 2 genes, respectively, with seven additional singletons. Their anchors were EIF4B, RAC3, GNB1, EDAR, COL6A1, PTK2, and CACNB2. At the numerical edge tolerance, each multi-gene component was internally complete, reflecting the dense local inverse-covariance blocks. STRING functional and physical networks were analyzed separately.

A1.5 Age–gene reweighting

The two additive components of the age-adjusted classifier were reweighted while retaining the base model's gene coefficients, selected genes, and network. With scores oriented toward class 1, the components were

$$S_a = -\gamma(a - m_a), \qquad S_G = -\sum_j \beta_j(x_j - m_j),$$

$$S_a + S_G = -D.$$

Reweighting changed the relative contributions of age and gene expression to prediction.

BEAT2 specimens were allocated to five stratified folds using random seed 20261006. In each fold, netLDA was refitted on the other four folds using the saved τ, ν, α, d and nb settings, then applied to the held-out specimens. Within-fold fitting estimated gene and age transformations, class means, coefficients and selected features. The full-BEAT2 outcome-based union-hub ranking was retained across folds. These predictions condition on the full-data ranking; the nested assessment is reported in Section A13.

After confirming complete out-of-fold coverage, the age and gene components were standardized using their respective means and sample standard deviations across the BEAT2 out-of-fold predictions. With $u_a = (S_a - c_{oa})/s_{oa}$ and $u_G = (S_G - c_{oG})/s_{oG}$, the logistic combination was

$$\eta = b_0 + b_a u_a + b_G u_G, \qquad \Pr(Y=1) = 1/[1+\exp(-\eta)].$$

The intercept and component weights were estimated by minimizing the summed logistic negative log-likelihood with a fixed ridge penalty λ=1:

$$L(b) = \sum_i \{\log[1+\exp(\eta_i)] - Y_i\eta_i\} + (\lambda/2)(b_a^2 + b_G^2).$$

The intercept was unpenalized, and component weights were unconstrained. BFGS optimization used a maximum of 2,000 iterations and a relative tolerance of 10^{-10}. The penalty λ was fixed without consulting external outcomes. Fitting required successful folds, complete out-of-fold coverage, nonconstant components, and optimizer convergence.

For final prediction, the full-BEAT2 base model generated S_a and S_G. Saved out-of-fold centers and scales transformed these components before application of the frozen logistic weights. A probability ≥0.5 assigned class 1. The same rule was applied in BEAT2 and all seven external cohorts.

Within each cohort, AUCs were calculated for the base score, reweighted score, age-only component, and gene-only component. Mean external AUC was the unweighted mean across all seven non-BEAT2 cohorts, requiring valid results in

every cohort. Full-model BEAT2 AUCs were apparent training estimates. The logistic-model AUC on its out-of-fold input scores describes meta-model fit.

Saved outputs included weights and transformations, fold assignments, out-of-fold scores, optimization results, sample-level predictions, cohort AUCs, and the frozen base-plus-meta model. Reweighting preserved KPRM membership and edge weights.

A separate patient-grouped nested analysis repeated hub ranking within fitting partitions and evaluated reweighting on outer held-out patients. It preserved the original frozen model and is reported in Section A13.

A2 Earlier exploratory configuration

An earlier configuration used $\tau = 15$, $\nu = 40$, $\alpha = 0.80$, $d = 3$ and $nb = 10$ and retained 15 hub genes without additional network-recruited genes. Supplementary Table S1 reports its cohort AUCs. The final biological analyses and seven KPRMs were based on the subsequent 200-gene model.

Supplementary Table S1 Archived AUCs for the earlier 15-hub configuration

Cohort	N specimens	Base netLDA	Reweighted	Age score	Gene score
BEAT2	568	0.775	0.785	0.697	0.742
TCGA-LAML	135	0.683	0.688	0.664	0.648
GSE146173	241	0.726	0.740	0.665	0.694
GSE37642/GPL570	134	0.745	0.755	0.648	0.724
GSE37642/GPL96	409	0.636	0.682	0.687	0.582
GSE12417/GPL96	152	0.674	0.688	0.630	0.649
GSE12417/GPL570	77	0.611	0.636	0.623	0.601
GSE106291	245	0.733	0.746	0.671	0.708

A3 Clinical evaluation of the TP53-associated predictive module

A3.1 Supplementary methods

BEAT2 cohort and molecular annotation. One initial-diagnosis specimen was selected per patient before filtering on response or survival availability, using the earliest recorded collection time relative to study inclusion when multiple diagnosis specimens were available. Adult patients with valid positive overall-survival times and event indicators entered the survival analysis. Response eligibility required a known CR/CRi or refractory outcome, recorded standard chemotherapy, and complete required scores and covariates. Valid survival follow-up was not required for response eligibility. The primary response analysis comprised 298 unique patients sampled on study-inclusion day zero (205 CR/CRi and 93 refractory). The broader sensitivity analysis comprised 306 unique patients (211 CR/CRi and 95

refractory), adding eight patients sampled 1–49 days after inclusion. Study-inclusion timing was recorded separately from confirmation of pretreatment status.

TP53 alteration adjustment used the dedicated molecular annotation among patients meeting the sequencing-availability and annotation-consistency criteria. A summary-inclusive annotation was retained for sensitivity analyses.

Score construction and survival sensitivity analyses. TP53 belonged to KPRM 3, comprising TP53 and 43 network neighbors. Weighted scores for the complete module, the 43-gene neighbor component, and the ATM/PTEN neighbor subset were calculated by summing model-transformed expression values multiplied by their frozen netLDA coefficients. The ATM/PTEN subset was defined by KEGG p53 signaling (hsa04115; cached annotation accessed October 17, 2026), retrieved separately from the five candidate pathways. TP53 expression was evaluated separately without coefficient weighting. Before final score standardization, the complete module score equaled the coefficient-weighted TP53 contribution plus the neighbor score. Each predictor was then standardized using its mean and sample standard deviation in the BEAT2 diagnosis-reference cohort. Age retained the original BEAT2 model standardization. Complete-module and component scores were evaluated in separate models; joint models included TP53 expression and the neighbor score.

Survival sensitivity analyses examined proportional-hazards assumptions, time-varying associations, and alternative covariate specifications. Piecewise models estimated score effects during days 0–90 and after day 90, using separate interval baseline hazards and patient-clustered robust covariance. Primary piecewise models used a common age effect; secondary models allowed the age effect to vary between intervals. Early–later coefficient contrasts incorporated their full covariance.

External response endpoints and accession overlap. GSE146173 supplied the primary external comparison of CR/CRi against refractory disease. Two sensitivity endpoints additionally classified deaths with residual disease, or all recorded induction-death categories, as nonresponse. GSE106291 sensitive/resistant labels were retained as reported. Shared sample aliases and demographic information were used to audit overlap between accessions.

Matched-accession analyses used identical patients, GSE146173 response labels, and identical age and sex covariates, substituting only the accession-specific expression scores. Score agreement was summarized using Pearson and Spearman correlations and root mean square differences. Gene contributions were reconstructed from the original expression files using the saved normalization parameters and frozen coefficients. Patient influence was assessed using Cook’s distance,

leverage, DFBETAS, and leave-one-patient-out refitting of the four primary GSE146173 score models. All patients remained in the primary analyses.

A3.2 Supplementary results

Survival sensitivity analyses. In GSE37642/GPL96, the neighbor and module associations were strongest during the first 90 days. Early hazard ratios were 1.240 (95% CI, 1.068–1.438) and 1.238 (95% CI, 1.068–1.435), respectively, compared with later hazard ratios of 0.973 and 0.972. Direct early–later comparisons supported attenuation for both scores (neighbor: P = 0.0106, q = 0.0160; module: P = 0.00939, q = 0.0160). In TCGA-LAML, early and later point estimates were similar, with no significant temporal difference. Allowing the age effect to vary between intervals changed the estimates little (Supplementary Figure S3A).

Treatment-response sensitivity analyses. Additional adjustment for TP53 alteration status in 284 primary BEAT2 patients retained the neighbor association (OR, 0.680; 95% CI, 0.512–0.904; q = 0.0320) and module association (OR, 0.717; 95% CI, 0.542–0.949; q = 0.0398). Results were similar in the broader sensitivity cohort of 306 unique patients, including 211 CR/CRi and 95 refractory cases; 291 had annotations available for alteration-adjusted models.

In GSE146173, adding residual-disease deaths to the nonresponse group preserved the module association (n = 229; OR, 0.747; 95% CI, 0.573–0.975; q = 0.0423). Including all induction deaths weakened the association (n = 246; OR, 0.848; 95% CI, 0.678–1.060; P = 0.147). In GSE106291, the module and neighbor scores were not associated with the reported sensitive/resistant endpoint (Supplementary Figure S3B).

Cross-accession diagnostics and patient influence (Supplementary Figure S4). GSE146173 and GSE106291 shared 237 sample aliases with concordant age and sex. Among 208 patients with comparable primary response labels, the module odds ratio was 0.692 (95% CI, 0.505–0.949) using GSE146173 scores and 0.902 (95% CI, 0.747–1.089) using GSE106291 scores. Thus, the different estimates persisted after harmonizing patients, outcomes, and covariates.

Across shared samples, TP53-expression scores showed strong agreement (Pearson r = 0.904), whereas neighbor and module scores showed moderate agreement (r = 0.558 and 0.575). DAXX (death domain associated protein), a network-recruited gene in KPRM 3, showed the largest discrepancy in its weighted expression contribution between GSE146173 and GSE106291 (Pearson r = 0.124; root mean square difference = 1.008). Reconstruction reproduced the saved module scores to numerical precision in both accessions.

All 856 leave-one-patient-out fits completed successfully, and no score association changed direction. Module odds ratios ranged from 0.662 to 0.717, and neighbor odds ratios ranged from 0.716 to 0.773. The estimates remained stable under individual-patient omission, while score agreement varied between accessions (Supplementary Figure S4).

Additional Supplementary Methods

A4 Cohort preparation and analysis eligibility

Bulk expression and clinical records were assembled for BEAT2, TCGA-LAML, GSE146173, GSE37642/GPL570, GSE37642/GPL96, GSE12417/GPL96, GSE12417/GPL570, and GSE106291. BEAT2 normalized expression and TCGA-LAML UCSC Xena Toil gene expected-count values were imported as supplied. GSE146173 gene counts were transformed as log2(count + 1). Processed GEO series-matrix expression values were used for GSE37642 and GSE12417; GSE106291 expression was imported from the RNASeq worksheet of its supplied workbook. Ensembl version suffixes were removed before mapping to gene symbols with org.Hs.eg.db. Affymetrix probes were mapped using hgu133a.db (GPL96) or hgu133plus2.db (GPL570). Unmapped features were excluded. When multiple features mapped to the same gene, the feature with the largest across-sample variance was retained. Matrices were organized as samples by genes.

For one-year classification, death at or before day 365 defined short survival, and follow-up beyond day 365 established long survival. Samples censored at or before day 365 were excluded. These criteria yielded 571 BEAT2 specimens and 135, 241, 134, 409, 152, 77, and 245 specimens in the seven external cohorts, respectively. Three BEAT2 specimens lacked age, leaving 568 age-available classification specimens representing 513 distinct patient identifiers. Two recorded zero ages and other ages below 18 were retained in this classification population, as documented in manuscript Table 1. Adult-only diagnosis-stage TP53 analyses used separate eligibility criteria.

Clinical and expression records were linked using cohort-specific identifiers. GSE146173 age was recovered from the GEO phenotype field age_(years):ch1 by matching the sample identifier preceding the colon in each GEO title. TCGA survival identifiers were harmonized to reconcile patient identifiers with sample-type suffixes. Matching audits recorded unmatched records, missing covariates, and inclusion status. Overlapping records between GSE146173 and GSE106291 were tracked explicitly in cross-accession analyses.

The 69-gene biological/ML hub union included 57 available candidates. Cohort-wide gene preprocessing preceded patient-grouped validation. Additional fitting-partition scaling and the age transformation were retained for prediction.

The candidate universe comprised PI3K–Akt (hsa04151), JAK–STAT (hsa04630), MAPK (hsa04010), NF-κB (hsa04064), and Wnt (hsa04310) signaling genes. Annotations were retrieved on October 1, 2026, using KEGGREST 1.46.0, with Entrez identifiers mapped through org.Hs.eg.db 3.20.0. The 865-gene union was restricted to genes measured in all eight cohorts; genes with nonfinite values or zero variance in BEAT2 were removed, yielding 660 candidate genes in every

cohort. Expression was centered and scaled separately within each cohort before one-year endpoint filtering. Any column producing nonfinite standardized values was set to zero in that cohort. netLDA then estimated its fitting-partition gene transformations in BEAT2 and retained them for prediction; age used the BEAT2 fitting-partition mean and sample standard deviation. Genome-wide matrices were used for GSVA, ESTIMATE, and transcriptomic rank scoring.

A5 Comparator signatures and survival evaluation

Survival evaluation used separate age-adjusted Cox models for the netLDA gene signature and each comparator signature within each cohort. The Cox analysis used genes selected by the gene-only netLDA fit. Complete required expression, age, and survival measurements defined eligibility. Age was centered and scaled within the eligible samples of each cohort-specific refit. When the number of expression predictors exceeded half the eligible sample size, the minimum number of principal components explaining at least 95% of standardized expression variance replaced the gene predictors. Age remained a separate covariate. Models used Efron's approximation for tied event times, and global Schoenfeld-residual tests assessed proportional hazards. Harrell's C-index summarized apparent within-cohort discrimination.

Kaplan–Meier curves summarized overall survival by predicted short- and long-survival groups from the frozen age-adjusted netLDA classifier and a 20-hub-gene-only LDA classifier. ITGB6 and CCND3 expression groups used the corresponding BEAT2 medians, with values at or below the median classified as low and values above it classified as high. Training rules and cutoffs were retained externally. Log-rank tests compared survival distributions, and curves displayed pointwise 95% confidence bands and censoring marks. The earlier 15-hub-gene configuration in Section A2 remains a separate exploratory analysis.

In the retained comparator panel, netLDA had a mean evaluation-cohort AUC of 0.643, compared with 0.638 for GBM, 0.636 for Enet[$\alpha = 0.2$], and 0.635 for LogitBoost+RF (Supplementary Figure S6A). netLDA AUC was 0.803 in BEAT2 and 0.589–0.702 across the seven evaluation cohorts. The panel was restricted after inspection of evaluation-cohort AUCs; these comparisons therefore describe the retained methods.

Age-adjusted cohort-specific Cox refits of the netLDA gene set yielded a mean apparent C-index of 0.803 across the seven non-BEAT2 cohorts; RF+Enet[$\alpha = 0.5$] yielded 0.712 in the retained comparator panel (Supplementary Figure S6B). netLDA-set C-indices were 0.814 in BEAT2, 0.830 in TCGA-LAML, 0.750 in GSE146173, 0.807 in GSE37642/GPL570, 0.810 in GSE37642/GPL96, 0.793 in GSE12417/GPL96, 0.855 in GSE12417/GPL570, and 0.776 in GSE106291. These apparent refit estimates compare signatures with different dimensionality and PCA representations.

The age-adjusted netLDA classifier identified groups with significantly different overall survival in BEAT2 and six of the seven external cohorts. In BEAT2, median survival was 249 days in the predicted short-survival group and 934 days in the predicted long-survival group (log-rank $P = 8.19 \times 10^{-29}$). Applying the frozen BEAT2 classification rule externally yielded median survivals of 305.5 versus 792 days in TCGA-LAML ($P = 6.29 \times 10^{-4}$), 407.5 versus 938 days in GSE146173 ($P = 0.0439$), 239.5 versus 789 days in GSE37642/GPL570 ($P = 3.31 \times 10^{-5}$), 242 versus 463 days in GSE37642/GPL96 ($P = 2.71 \times 10^{-5}$), 238 versus 427 days in GSE12417/GPL96 ($P = 0.0194$), and 399.5 versus 1,177 days in GSE106291 ($P = 0.00475$). In GSE12417/GPL570, median survival was also longer in the predicted long-survival group, although the difference was not significant (365 versus 624 days; $P = 0.233$).

The 20-hub-gene-only LDA classifier separated survival in BEAT2 (246 versus 761 days; $P = 3.83 \times 10^{-26}$) and the same six external cohorts. External log-rank P-values ranged from 7.37×10^{-7} in GSE37642/GPL570 to 0.0221 in GSE106291. In GSE12417/GPL570, median survival was 286 versus 624 days ($P = 0.0638$). Both multigene classifiers therefore showed survival separation across most evaluated cohorts. Survival curves for selected cohorts are shown in Supplementary Figure S6C–G.

Individual-gene associations varied across cohorts. High ITGB6 expression was associated with longer median survival in BEAT2 (582 versus 329.5 days; $P = 1.15 \times 10^{-5}$) and TCGA-LAML (792 versus 365.5 days; $P = 0.00855$), with no significant separation in the remaining six cohorts. High CCND3 expression was associated with shorter survival in BEAT2 (358.5 versus 555 days; $P = 0.00882$) and four external cohorts: TCGA-LAML (304 versus 822 days; $P = 1.93 \times 10^{-5}$), GSE146173 (378 versus 1,382 days; $P = 2.41 \times 10^{-4}$), GSE37642/GPL96 (301 versus 363.5 days; $P = 0.0273$), and GSE106291 (387 versus 1,352 days; $P = 1.53 \times 10^{-4}$). All reported log-rank P-values were unadjusted for multiple comparisons.

A6 Functional characterization and supporting statistical summaries

GSVA used normalized genome-wide expression and a Gaussian kernel to calculate scores for the seven KPRMs and 50 MSigDB Hallmark gene sets. Gene sets required at least 50% coverage, with minimum sizes of two measured genes for KPRMs and ten for Hallmark pathways. Cohort-specific linear models related each Hallmark score to a KPRM score, adjusting for age standardized using BEAT2 parameters and frozen netLDA risk group. KPRM and response scores were standardized within the complete-case samples for each test. Benjamini–Hochberg adjustment covered the 350 planned module–Hallmark comparisons separately within each cohort and scoring specification.

Hallmark scores were recalculated after removing the union of all KPRM genes from each gene set. Original and overlap-removed coefficients were compared for direction and magnitude. ESTIMATE ImmuneScore, StromalScore, and ESTIMATEScore were related to KPRM scores using the same adjusted regression framework, with 21 planned tests per cohort. Untestable comparisons remained in the output with their status and reason. Supplementary workbooks contain gene coverage, model coefficients, P-values, adjusted P-values, and sample counts.

STRING version 12.0 functional and physical networks were evaluated separately at confidence thresholds of 0.400 and 0.700, without adding proteins beyond the selected genes. Gene symbols were initially mapped using the human STRING identifier API (taxon 9606). MAPK10 and VEGFA were excluded from the STRING background because their returned identifiers corresponded to DUSP16 and COL18A1, respectively, and their identities could not be verified in the inspected v12.0 protein and alias tables. The accepted candidate background therefore comprised 658 of 660 genes. All 200 selected genes remained mapped, preserving coverage of the seven KPRMs and seven singletons. These exclusions did not alter the fitted netLDA model or KPRM definitions. Connectivity enrichment used 10,000 random gene sets drawn without replacement within each draw from the mapped candidate universe. Reference sets matched size alone or additionally matched exact five-pathway KEGG membership patterns and candidate-network degree bins defined by empirical tertile cut points; coincident cut points were collapsed. Degree bins were calculated separately for each network type and threshold. Expected edge counts were the mean across random sets. One-sided upper-tail permutation P-values were calculated as (1 + the number of random sets with at least as many edges as observed)/10,001. BH correction was applied across the seven module tests separately for each network type, confidence threshold, and background model; the overall selected-network test was excluded from these correction families. STRING-native connectivity enrichment was obtained separately through the ppi_enrichment endpoint without a custom background parameter. Native node and edge counts were checked against the saved selected networks. These native tests assess network connectivity enrichment. Native P-values returned as zero were flagged rather than interpreted as exact zero probabilities.

The selected functional network was enriched in connections relative to the mapped candidate-gene universe. Compared with size-matched random gene sets, it contained 1,164 observed edges versus 919.2 expected edges, corresponding to a 1.27-fold enrichment (permutation $P = 0.00970$). Enrichment persisted after matching for exact KEGG pathway-membership patterns and candidate-network degree bins defined by tertile cut points: 943.0 edges were expected, yielding a 1.23-fold enrichment ($P = 1.00 \times 10^{-4}$). Under this matched background, KPRMs 1–5 were individually enriched, with observed-to-expected ratios of 1.47, 1.37, 1.44, 3.25, and 4.23 and corresponding BH-adjusted P-values of

0.00397, 0.0192, 0.0192, 0.00140, and 0.00280. KPRMs 6 and 7 contained no internal functional edges at this confidence threshold. Connectivity enrichment is summarized in Figure 4F.

The high-confidence physical-interaction network contained 356 edges, including 117 within-KPRM, 233 between-KPRM, and six singleton-associated edges. Overall connectivity exceeded the pathway- and candidate-degree-matched expectation of 302.9 edges (1.18-fold enrichment; permutation P = 0.0121). KPRM 1 showed significant physical-interaction enrichment, with 66 observed versus 31.1 expected edges (2.13-fold enrichment; BH-adjusted $P = 7.00 \times 10^{-4}$). The remaining modules were not significant under this matched background. Thus, functional connectivity enrichment extended across several modules, whereas significant physical-interaction enrichment was concentrated in KPRM 1.

Sensitivity analyses at a confidence threshold of 0.400 showed differences by network definition. Under the pathway- and candidate-degree-matched background, the complete functional network showed 1.07-fold enrichment (P = 0.0978), with significant module-level enrichment for KPRMs 1, 4, and 5. The complete physical-interaction network remained enriched (1.29-fold; $P = 1.00 \times 10^{-4}$), with significant enrichment for KPRMs 1, 2, and 5.

To assess whether KPRM–Hallmark associations were driven by shared genes, we repeated Hallmark scoring after removing the union of genes in all seven KPRMs from every Hallmark gene set. Adjusted coefficients from the original and overlap-removed analyses were compared across cohorts. For visualization, 18 distinct Hallmark pathways were selected by combining the three largest absolute adjusted associations for each KPRM in BEAT2 and adding the next highest-ranked pathways as needed. Cross-cohort summaries displayed the strongest overlap-removed Hallmark association for each KPRM in BEAT2; external-cohort results were not used to select these associations.

Additional functional results

GSVA identified extensive associations between KPRM scores and Hallmark biological programs after adjustment for age and predicted netLDA risk group. In BEAT2, 291 of 343 estimable KPRM–Hallmark associations were significant at $q < 0.05$. KPRM 1 was strongly associated with IL6–JAK–STAT3 signaling (standardized $\beta = 0.886$, $q = 2.46 \times 10^{-175}$), while KPRM 2 was strongly associated with apoptosis ($\beta = 0.846$, $q = 1.52 \times 10^{-143}$). Both relationships were positive and significant across all eight cohorts. Coefficients ranged from 0.570 to 0.886 for IL6–JAK–STAT3 signaling and from 0.655 to 0.846 for apoptosis, with maximum q-values of 5.46×10^{-9} and 1.56×10^{-13}, respectively. Across the seven external cohorts, 232–299 of the 350 KPRM–Hallmark comparisons per cohort were significant. BEAT2 Hallmark associations are shown in Figure 4A.

These associations were largely preserved after removing the union of KPRM genes from the Hallmark gene sets. Across 2,793 estimable paired comparisons, the median absolute change in standardized coefficients was 0.008, and 98.5% retained the same direction. The overlap-removed associations of KPRM 1 with complement, KPRM 2 with IL2–STAT5 signaling, KPRM 3 with protein secretion, KPRM 4 with allograft rejection, and KPRM 6 with myogenesis were positive and significant in all eight cohorts. Their maximum q-values across cohorts were 5.99×10^{-6}, 1.78×10^{-6}, 1.57×10^{-12}, 1.73×10^{-4}, and 1.04×10^{-5}, respectively. The associations therefore persisted after shared genes were removed. Cross-cohort overlap-removed associations and coefficient comparisons are shown in Figure 4B–C.

ESTIMATE analyses revealed distinct immune- and stromal-associated expression patterns. In BEAT2, KPRM 1 was positively associated with ImmuneScore ($\beta = 0.822$, $q = 1.32 \times 10^{-123}$) and StromalScore ($\beta = 0.795$, $q = 1.26 \times 10^{-114}$). Both associations remained significant across all eight cohorts, with coefficients ranging from 0.440 to 0.822 for ImmuneScore and from 0.175 to 0.795 for StromalScore. KPRM 4 was also positively associated with ImmuneScore in every cohort ($\beta = 0.304$–0.517; all $q \leq 2.03 \times 10^{-4}$). KPRM 6 showed a consistent positive association with StromalScore ($\beta = 0.238$–0.337; all $q \leq 0.035$), whereas KPRM 7 had no significant association with ImmuneScore. ImmuneScore and StromalScore associations are shown in Figure 4D.

STRING analysis identified extensive biological connectivity among the 200 netLDA-selected genes. At a confidence threshold of 0.700, the functional network contained 1,164 associations, including 330 within-KPRM, 787 between-KPRM, and 47 singleton-associated edges. Of these, 1,001 (86.0%) connected two network-recruited genes. TP53, MAPK3, and MAPK1 had the highest degrees, with 57, 54, and 51 connections, respectively. The functional network is shown in Figure 4E.

A7 Virtual knockout and predictive perturbation analyses

Single-cell virtual knockout used 700 balanced primary malignant CD14 cells and a 2,000-gene universe retaining expression-eligible KPRM genes. Counts were normalized to counts per million. A common wild-type network was reconstructed using scTenifoldKnk/scTenifoldNet from ten principal-component regression networks, each using 500 cells, three components, and $q = 0.9$. Rank-three tensor decomposition integrated the networks, followed by directionality processing and diagonal removal. The random seed was 1.

EIF4B, PAK1, and IFNGR2 perturbations zeroed the target's outgoing weights in the saved wild-type network. Two-dimensional manifold alignment and package differential-regulation analysis, with empirical-null estimation disabled,

quantified downstream changes. Genes with package-adjusted P <0.05 were summarized after excluding the target itself. Three controls per target were selected using detection, expression, degree, and absolute-weight strength. All bulk-selected genes and SRGN were excluded from the control pool. Module responses averaged squared manifold distances normalized by the mean squared distance among non-target genes and were compared with the matched-control mean and range.

Bulk perturbations held BEAT2 preprocessing, class contrasts, age effects, gene coefficients, and age–gene reweighting fixed except for the specified perturbation. Gene ablation removed the midpoint-centered expression contributions of the target genes. Network disconnection removed off-diagonal Ω entries incident to the targets while retaining diagonal entries; the perturbed coefficients were reconstructed from the frozen marginal contrasts. Targets included all selected genes, hub genes, network-recruited genes, and complete-module, hub-gene, and network-recruited subsets. A scaling sensitivity analysis restored the perturbed gene score to the baseline BEAT2 standard deviation before applying frozen reweighting.

Predictive effects were baseline minus perturbed AUC, summarized by cohort and as the equally weighted external mean. Cohort-specific intervals used 1,000 paired, outcome-stratified bootstrap replicates. Matched gene-set references preserved size and hub/network-recruited composition; disconnection references additionally matched edge burden within 10%. Edge-reference perturbations matched removed-edge counts and hub/network-recruited endpoint types exactly, with positive-definite references evaluated separately. Reference effects were summarized by their mean and central 95% range. Upper-tail rank probabilities were BH-adjusted across 21 planned module/component comparisons for gene ablation and 28 for each disconnection background and scaling specification.

Bulk predictive perturbation results

Ablation of all selected genes, retained hub genes, and network-recruited genes produced mean external AUC losses of 0.0271, 0.0206, and 0.0034, respectively. Corresponding network-disconnection effects were 0.0653, −0.0034, and 0.0643. After standard-deviation matching, these effects were attenuated to 0.0066, −0.0018, and 0.0063, respectively (Figure 2A), demonstrating substantial sensitivity to changes in gene-score dispersion.

Among complete-module perturbations, KPRM 1, containing 64 genes, showed the largest mean loss after raw network disconnection (0.0570), followed by KPRM 2 (0.0215) and KPRM 3 (0.0095). KPRM 1 also showed the largest loss after standard-deviation matching (0.0144), whereas its gene-ablation loss was 0.0017 (Figure 2B). Disconnecting its 62

network-recruited genes produced a similar standard-deviation-matched loss of 0.0145, compared with 0.0033 for its two hub genes.

KPRM 1 network-recruited disconnection reduced AUC in five of seven external cohorts after standard-deviation matching. Effects ranged from −0.0094 to 0.0351. Confidence intervals excluded zero in GSE37642/GPL96 (loss, 0.0264; 95% CI, 0.0095–0.0432) and GSE12417/GPL96 (0.0351; 95% CI, 0.0074–0.0620), but included zero in the other five cohorts (Figure 2C).

KPRM 1 disconnection exceeds matched-reference effects. For complete KPRM 1 disconnection, the observed standard-deviation-matched loss of 0.0144 exceeded the mean of 621 matched gene-set references (0.0036; central 95% range, −0.0003–0.0073; adjusted upper-tail rank probability, q = 0.0277). It also exceeded the mean of 1,000 positive-definite edge references (0.0039; range, 0.0012–0.0066; q = 0.0140).

The network-recruited-disconnection loss of 0.0145 similarly exceeded the mean of 504 gene-set references (0.0038; range, −0.0004–0.0075; q = 0.0277) and 1,000 positive-definite edge references (0.0039; range, 0.0014–0.0064; q = 0.0140; Figure 2D). These ranges summarize variation among reference perturbations.

None of the 21 module/component gene-ablation comparisons showed an excess loss at an adjusted upper-tail reference probability below 0.05. KPRM 1 disconnection therefore produced greater predictive losses than the specified matched-reference perturbations, with effects varying across cohorts.

A8 Cellular and spatial characterization and treatment-response sensitivity

Single-cell analyses used GSE116256 and the adult subset of GSE185381, with supplied BoneMarrowMap coordinates and cell annotations. Eligible cells passed mapping-quality review, had assignment probability ≥0.50, valid fine and broad annotations, and nonzero expression totals. Adult diagnosis-stage AML and healthy-control samples were retained. Replication microenvironment cells were classified as AML-associated nonmalignant cells. Full-module, hub-gene, and network-recruited signatures were scored with UCell maxRank = 1,500 in the common measured transcriptomic universe, requiring ≥70% signature coverage.

Scores were averaged within sample, cell state, and malignancy category, requiring at least 20 cells per group. Paired comparisons required at least three complete sample pairs; independent Welch tests required at least three samples per group and no shared identifiers. Partially overlapping groups with fewer than three pairs were not tested. State-versus-rest comparisons used each sample's other eligible malignant states as the reference. BH adjustment covered all seven

modules and eligible states separately within each dataset, signature component, and comparator. Fine-state, broad-lineage, and state-versus-rest analyses had separate testing families. Replication required concordant directions and FDR <0.05 in both studies. KPRM_001 follow-up retained the original all-module adjustment.

CellChat networks were fitted separately within AML sample units using populations with at least 30 cells and representation in at least three sample units in both datasets. Expression was normalized to 10,000 per cell and log1p-transformed before signaling-gene extraction. Communication estimation used 10% truncated means, no population-size weighting, and 100 permutations. Supported interactions had positive probability and $P < 0.05$. Recurrence required at least three eligible samples, support in at least three, and a support fraction ≥ 0.50. Associations between pathway communication and population-level KPRM scores used sample-level Spearman correlations. Hub–network-recruited gene correlations and communication correlations required at least five units, with exhaustive permutations for five to eight units and 99,999 Monte Carlo permutations with plus-one correction for larger groups. Communication associations were adjusted separately by signature component and score role across the planned testing family.

Spatial analyses used AML11 and AML12 from GSE255217 and BM1 and BM2 from GSE279576. UCell scores used the shared measured universe and the same signature coverage rule. RCTD weights were pooled into monocytic, conventional dendritic, plasmacytoid dendritic, stem/progenitor, erythroid, T, NK, and B/plasma compartments. Completed fits were compared across the primary and alternative reference definitions. Monocytic and erythroid associations formed the primary compartment family.

Spatial tests used observed nearest-neighbor links on the Visium grid and excluded components with fewer than three spots. Component-centered ranks were analyzed using Moran's I with 999 within-component permutations. Depth adjustment removed ranked total-count effects; Fisher projections also underwent a separate detected-gene adjustment. Three binomial thinning replicates targeted 500 expected counts per spot, with lower-depth spots retained without upsampling. Compartment associations used 1,999 graph-Laplacian spectral sign surrogates. BH families were defined separately by specimen, score specification, reference definition, and compartment family.

Signed spatial projections used frozen BEAT2 gene coefficients applied to log1p-normalized expression standardized within each specimen. Module and singleton contributions summed to the overall gene projection. Age and the intercept were excluded from these maps. Both measured-gene coverage and retained absolute coefficient-weight coverage had to

reach 70%. Thinned projections reused the original specimen-specific centers and scales. These spatial transformations were kept distinct from the frozen BEAT2 transformations used for bulk clinical prediction.

Bulk induction-response models compared CR/CRi with refractory disease in BEAT2 (306 eligible diagnosis-stage patients; 211 CR/CRi and 95 refractory) and GSE146173 (214 patients; 156 CR/CRi and 58 refractory). Separate logistic models evaluated overall, KPRM, hub-gene, and network-recruited-gene contributions, adjusting for standardized age and sex. Primary-endpoint score scaling was cohort-specific and retained for sensitivity analyses. Nonlinearity was assessed using natural splines with three degrees of freedom. TP53-focused models instead used the frozen BEAT2 diagnosis-reference score scaling and the primary 298-patient day-zero subset described in Section A3. Sensitivity analyses examined study-inclusion timing, TP53 alteration status, alternative handling of induction deaths, and patient influence. GSE146173/GSE106291 diagnostics harmonized shared identifiers and response labels before comparing accession-specific scores.

Gui single-cell response analyses used eight baseline patients: four published responders, including three with stable disease and one with morphologic leukemia-free state, and four with progressive disease. Patient–population means required at least 20 cells. Two-sided exact patient-level permutation tests compared scores, with BH adjustment across the planned module–population family. CD8 subtype composition and exact thinning to 1,000 UMIs over 20 draws assessed compositional and depth sensitivity. The published response grouping was retained separately from the bulk CR/CRi-versus-refractory endpoint.

A9 Reproducibility and supplementary data organization

Model objects and analysis checkpoints retained fitted parameters, gene orders, hub labels, preprocessing transformations, and prediction functions.

The gene-level supplementary workbook provides selected-gene composition, hub status, coefficients, module membership, and KEGG annotations. The functional-association workbook provides complete test results, including untestable comparisons. Statistical outputs retain sample counts, effect estimates, confidence intervals, P-values, adjusted P-values, and analysis status.

Supplementary Table S2 Guide to the supplementary analyses

Analysis	Location and contents
Network and score definitions	A1: local networks, coefficient decomposition, KPRM construction and age–gene reweighting

Analysis	Location and contents
Earlier configuration	A2 and Supplementary Table S1: archived results of the separate 15-hub configuration
TP53 clinical analyses	A3 and Supplementary Figures S2–S4: score definitions, response endpoints, survival sensitivities and accession overlap
Cohorts and comparator evaluations	A4–A5 and Supplementary Figure S6: preprocessing, sample eligibility, descriptive comparator AUCs and apparent Cox refits
Functional and perturbation analyses	A6–A7 and A10: enrichment backgrounds, gene ablation, disconnection and single-cell virtual knockout
Cellular and spatial characterization	A8 and A10; Supplementary Figures S5, A7–A9: sample-level comparisons, communication, spatial diagnostics and Gui analyses
Experimental CRISPR evidence	A11 and Supplementary Figures S10–S12: matched dependency analyses and perturbation-expression sensitivities
Clinical response and venetoclax	A12; Supplementary Figures S13–S14 and Supplementary Table S3: incremental models, specimen-matched drug screens and independent Finnish assessment

A10 Additional cellular and perturbation analyses

A10.1 Single-cell virtual knockout reconstruction and matched controls

Single-cell virtual knockout and matched-control analysis. Virtual knockout analysis used 700 balanced primary malignant CD14 cells and a 2,000-gene universe retaining expression-eligible KEGG predictive regulatory module (KPRM) genes. Raw RNA counts were normalized to counts per million. A single wild-type regulatory network was constructed using scTenifoldKnk/scTenifoldNet functions [20, 21] and reused for all target and control perturbations. Reconstruction combined 10 principal-component regression networks, each using 500 cells, with three components and a filtering parameter of $q = 0.9$. Networks were integrated by rank-three tensor decomposition, followed by directionality processing and removal of diagonal entries. The random seed was fixed at 1.

EIF4B, PAK1, and IFNGR2 were individually knocked out by setting their outgoing regulatory weights to zero in the saved wild-type matrix, without re-estimating the network. Wild-type and perturbed networks were compared using two-dimensional manifold alignment and the package differential-regulation procedure with empirical-null estimation disabled. Package-reported adjusted P-values were retained, and downstream genes were identified at FDR < 0.05 after excluding the knocked-out gene.

Three controls per target were selected before evaluating knockout outcomes. Matching incorporated detection frequency, mean log-normalized expression, incoming and outgoing degrees defined by the upper 10% of absolute network weights, and total incoming and outgoing absolute-weight strengths. Matching covariates other than detection frequency were log1p-transformed, and controls were selected by the smallest sum of squared, standard-deviation-scaled covariate differences. All bulk netLDA-selected genes and SRGN were excluded from the control pool.

For descriptive analysis, each gene's squared manifold distance was divided by the mean squared distance among non-knocked-out genes. Module scores were calculated by averaging these relative squared distances across evaluable, non-knocked-out module genes. Target scores were compared with the mean and range of their three matched controls. Target–control score comparisons were descriptive; package-derived FDR values were reported separately. Modules without evaluable genes were reported as unavailable.

Single-cell virtual knockout results. The single-cell analysis comprised three target knockouts and nine matched-control knockouts across 2,000 genes. PAK1 knockout identified one significant downstream gene, SRGN (package-reported FDR = 1.67×10^{-8}), whereas EIF4B and IFNGR2 knockouts identified no significant downstream genes. The knocked-out gene accounted for 99.05%, 96.51%, and 99.38% of total squared distance for EIF4B, PAK1, and IFNGR2, respectively, indicating that the overall perturbation was dominated by the directly altered gene.

SRGN was also significant following six of the nine matched-control knockouts, indicating a response shared across several perturbations. Module-level analyses evaluated 44 genes in KPRM 1, 26 in KPRM 2, 30 in KPRM 3, and one in KPRM 4; KPRMs 5–7 had no evaluable genes. No evaluated module gene reached FDR < 0.05 following any target knockout.

Perturbation scores varied by target and module. IFNGR2 knockout produced a mean relative squared distance of 2.017 in KPRM 1, compared with a matched-control mean of 0.763 and range of 0.106–1.920. PAK1 knockout produced a score of 0.00576 in KPRM 3, exceeding its matched-control range of 0.00039–0.00258. These scores summarize descriptive target–control contrasts.

Additional cellular localization results

After sample selection and cell filtering, the primary analysis included 16,499 cells from 17 sample units (11 AML and six healthy), comprising 8,304 malignant cells, 2,143 AML-associated nonmalignant cells, and 6,052 healthy-control cells. The replication analysis included 124,141 cells from 30 sample units (20 adult AML and 10 healthy), comprising 58,567

malignant cells, 20,443 AML-associated nonmalignant cells, and 45,131 healthy-control cells. All 193 genes across the seven KPRMs were available in the common measured gene universe, providing 100% coverage for all 21 full-module, hub-gene, and network-recruited-gene signatures. The seven modules contained 64, 51, 44, 17, 12, three, and two genes, respectively. Broad-lineage annotations and malignancy status are illustrated in Figure 5A–B.

The malignant-cell localization heatmaps included 13 matched fine states with at least three eligible sample units in each dataset (Figure 5C). KPRM_001 expression enrichment extended across progenitor and differentiated myeloid states. In the primary dataset, its highest mean UCell scores occurred in CD16 monocytes (0.1324), cycling progenitors (0.1279), and LMPP cells (0.1277). In replication, the highest scores occurred in CD14 monocytes (0.1529), CD16 monocytes (0.1505), and cDC2 cells (0.1461).

All seven modules underwent cellular characterization. KPRM 1, comprising EIF4B, PAK1, and 62 network-recruited genes, was prioritized for exploratory follow-up because its myeloid localization and CD14-monocyte contrasts were directionally consistent across datasets. Malignant-minus-AML-associated-nonmalignant differences were 0.0122 UCell units in five primary pairs (95% CI, 0.0019–0.0226; P = 0.0302; q = 0.0705) and 0.0162 in six replication pairs (95% CI, 0.0051–0.0274; P = 0.0134; q = 0.0940; Figure 5E). Neither contrast met the joint FDR replication criterion. Complete all-module comparisons and component correlations are reported in Supplementary Information Section A10.

Broad-lineage analysis preserved the positive KPRM_001 malignant–nonmalignant monocyte contrast (Figure 5D, F). The mean difference was 0.0120 in the primary dataset (P = 0.0306; FDR = 0.0715) and 0.0123 in replication (P = 0.0530; FDR = 0.4951). Within-sample comparisons also identified enrichment in malignant monocytes relative to other eligible malignant broad lineages. The mean difference was 0.0066 among seven primary sample units (P = 0.0162; FDR = 0.2270) and 0.0277 among 13 replication units (P = 3.10 × 10−5; FDR = 0.0033; Figure 5G). In the fine-state replication analysis, CD14 monocytes had higher scores than other eligible malignant states in 12 of 13 sample units (mean difference, 0.0247; FDR = 0.0147), and all leave-one-sample-out mean effects remained positive. These analyses highlighted KPRM_001 enrichment in the myeloid compartment, with the strongest FDR-adjusted localization evidence in the replication dataset.

A10.2 Hub and network-recruited score coordination

Coordination between hub and network-recruited genes was assessed using two-sided Spearman correlations between sample-level component scores within each cell state and condition–malignancy stratum. Tests required at least five sample units and nonconstant scores. P-values were calculated using exhaustive permutations for five to eight units and

99,999 Monte Carlo permutations with the plus-one correction for larger groups. Benjamini–Hochberg correction was applied across the all-module testing family separately for each dataset and fine-state or broad-lineage analysis.

Correlations between hub and network-recruited genes for KPRM_001 were positive in malignant CD14 monocytes in both datasets, although neither was significant after permutation testing and FDR correction (primary: n = 7, Spearman $\rho = 0.250$, P = 0.5948, FDR = 0.7348; replication: n = 15, $\rho = 0.418$, P = 0.1231, FDR = 0.5521). Broad monocyte correlations showed similar patterns (primary: $\rho = 0.214$, P = 0.6615, FDR = 0.8156; replication: $\rho = 0.429$, P = 0.1139, FDR = 0.5515). No fine-state correlation reached FDR <0.05 in either dataset. In the replication broad-lineage analysis, KPRM_001 hub and network-recruited scores were strongly correlated in AML-associated nonmalignant naive T cells (n = 12; $\rho = 0.888$; P = 0.00028; FDR = 0.0449).

A10.3 Recurrent communication and KPRM associations

CellChat networks [30] were estimated separately for each sample. Expression values were normalized to a total of 10,000 per cell using all measured genes and transformed using log1p before extracting signaling genes. Both datasets used the same human protein-signaling database, comprising 1,280 secreted-signaling, 424 extracellular-matrix–receptor, and 535 cell–cell-contact entries. Overexpressed signaling genes and candidate ligand–receptor interactions were identified using the CellChat workflow. Communication probabilities were calculated from unsmoothed normalized expression using a truncated mean with 10% trimming, without population-size weighting, and with 100 permutations. A sample-level interaction was considered supported when its inferred probability was positive and its CellChat permutation P-value was <0.05.

Recurrence was evaluated for each directed sender–receiver–ligand–receptor combination. The support fraction was the number of supporting samples divided by the number of successful sample fits containing both eligible populations. Recurrent interactions required at least three eligible samples, support in at least three samples, and a support fraction ≥0.50. primary-priority interactions were evaluated in replication using the same population definitions, database, inference settings, and recurrence criteria. Malignant-monocyte communication was summarized through incoming, outgoing, and within-population interactions. Representative interactions were selected using support in the primary dataset and pathway diversity.

To relate communication patterns to KPRMs, existing UCell scores for the full-module, hub-gene, and network-recruited-gene signatures were averaged within each sample and communication population. Pathway communication

scores were calculated by summing all tested ligand–receptor probabilities within each directed population pair before significance filtering. For eligible population pairs, pathways without screened interaction records received a score of zero. Associations between pathway scores and sender- or receiver-population KPRM scores were assessed using two-sided Spearman correlations, requiring at least five complete sample units and variation in both scores. Within-population interactions were tested once. P-values were obtained by exhaustive permutations for five to eight sample units and 99,999 Monte Carlo permutations with the plus-one correction for larger groups. Leave-one-sample-out correlations assessed directional stability. All testable primary associations were followed up in replication regardless of their primary P-values. Benjamini–Hochberg correction was applied separately by signature component and sender/receiver score role across the full planned family of module, pathway, and population-pair comparisons. Replicated associations required concordant directions and FDR <0.05 in both datasets.

Recurrent cell–cell communication across AML datasets. Sample-specific communication networks were successfully estimated for nine primary and 20 replication AML sample units, encompassing 7,847 and 61,643 cells, respectively, across the eligible communication populations. Malignant monocytes were represented in seven primary and 15 replication sample networks, while malignant pro-monocytes were represented in five and 11 networks, respectively (Supplementary Figure S7, panel A).

The primary analysis identified 96 recurrent directed ligand–receptor combinations representing 35 distinct ligand–receptor interactions. Of these, 89 combinations (92.7%) met the same recurrence criterion in replication. Malignant-monocyte communication showed similarly high recurrence: 39 of 42 primary-priority combinations (92.9%) satisfied the replication criterion, including 24 of 25 incoming and 18 of 20 outgoing combinations. These incoming and outgoing counts included three within-population interactions (Supplementary Figure S7, panels B–C).

Recurrent incoming communication connected malignant progenitor populations with malignant monocytes through MIF, CD99, and antigen-presentation-associated interactions. MIF–CD74/CD44 signaling from malignant pro-monocytes to malignant monocytes was supported in all five eligible primary samples and seven of ten replication samples. MIF–CD74/CXCR4 signaling along the same population pair was supported in four of five primary and five of ten replication samples. MIF–CD74/CD44 interactions from malignant early GMP and late GMP populations to malignant monocytes were supported in three of three primary samples for each pair and in seven of eight and four of five replication samples, respectively (Supplementary Figure S7, panel D).

Within malignant monocytes, CD99–PILRA and PECAM1–PECAM1 interactions were supported in five of seven primary samples and in 12 of 15 and 13 of 15 replication samples, respectively. Outgoing interactions toward AML-associated nonmalignant CD8 memory T cells included recurrent HLA–CD8 and CLEC2B–KLRB1 combinations. Three malignant-monocyte-associated priorities did not meet the replication recurrence criterion: LAIR1–LILRB4 from malignant HSC/MPP cells to malignant monocytes, CLEC2D–KLRB1 from malignant monocytes to nonmalignant CD8 memory T cells, and CD55–ADGRE5 from malignant monocytes to malignant LMPP cells. Their replication support fractions were 2/7, 1/10, and 2/4, respectively.

KPRM–communication associations were less consistent across datasets. Of 2,016 testable primary associations, 1,896 were also testable in replication, and 912 (48.1%) had concordant correlation directions. The remaining 120 follow-ups lacked variation in communication or KPRM scores. One association was nominally significant in both datasets: the KPRM_003 network-recruited signature was positively correlated with within-population malignant-monocyte IL16 communication in primary ($n = 7$; Spearman $\rho = 0.906$; $P = 0.00952$) and replication ($n = 15$; $\rho = 0.528$; $P = 0.0443$). No association met the joint FDR criterion. The strongest primary KPRM_001 associations also weakened externally; for example, its correlation with malignant-monocyte GALECTIN communication decreased from $\rho = 0.906$ in primary to $\rho = 0.219$ in replication ($P = 0.430$). Ligand–receptor recurrence was more consistent across datasets than its association with KPRM expression enrichment.

A10.4 Spatial projections and technical sensitivity

Spatial organization and technical sensitivity. Spatial neighborhoods were defined by nearest-neighbor connections on the Visium hexagonal grid. Connections were restricted to observed neighboring positions, and components containing fewer than three spots were excluded from spatial tests. Scores were ranked and centered within connected components. Positive spatial clustering was assessed using Moran's I [34], with P-values calculated from 999 permutations within connected components. Depth-adjusted analyses removed the association with ranked total counts. Fisher-projection analyses additionally included a separate adjustment for ranked detected-gene number.

Sequencing-depth sensitivity was evaluated using three binomial count-thinning replicates targeting an expected total of 500 counts per spot. Spots with fewer than 500 counts were retained without upsampling. UCell signatures were rescored in each replicate and averaged. Associations between module scores and RCTD compartment weights were evaluated using correlations of component-centered ranks; adjusted analyses removed the same depth covariate from both

variables. Spatial association P-values were calculated using 1,999 graph-Laplacian spectral sign surrogates, preserving spectral coefficient magnitudes and graph smoothness before covariate re-projection.

Benjamini–Hochberg correction was applied separately within each specimen and score specification. Moran's I testing families included seven UCell modules or nine Fisher-projection blocks. Compartment-association families included all module–compartment combinations within each reference definition and primary or exploratory compartment family. Analyses were conducted separately for each specimen, and consistency was assessed across specimens, technical sensitivity analyses, and reference definitions.

Spatial projection of netLDA gene contributions. To examine the spatial distribution of signed predictive gene contributions, we projected the frozen BEAT2 netLDA gene coefficients onto spatial expression. Full-matrix counts were normalized to 10,000 counts per spot, log1p-transformed, and standardized gene-wise across retained spots within each specimen. For spot s, the overall Fisher-style gene projection was

$$F_s = \sum_{g \in Gshared} \widehat{\beta}_g z_{gs}$$

Here β̂_g is the frozen BEAT2 gene coefficient, z_gs is specimen-standardized spatial expression, and Gshared is the shared measured-gene set. KPRM-specific projections summed the corresponding gene contributions within each module. The seven isolated selected genes formed a separate singleton block, so the seven module projections and singleton contribution summed exactly to the overall projection. Age and the discriminant intercept were excluded from these gene-contribution maps.

The same shared measured-gene set was used across all four specimens. Projection coverage was assessed by both gene count and retained absolute coefficient weight, with a minimum requirement of 70% for each. Count-thinning analyses reused the original specimen-specific gene means and standard deviations. Top outcome-associated genes were ranked using gene-wise linear models of BEAT2 expression on one-year outcome and standardized age, with Benjamini–Hochberg correction across all fitting genes. Expression maps displayed the ten highest-ranked final-model genes measured in all spatial specimens.

Maps were registered to the supplied histology and exported with tissue-focused viewports. Original and thinned projections shared a symmetric color scale within each selected specimen and module; adjusted maps displayed standardized rank residuals. Display limits were based on 99th-percentile magnitudes, with numerical scores preserved in the analysis outputs.

KPRM 3 also showed consistent compartment relationships after thinning. Its enrichment was positively associated with monocytic weights in BM1 ($\rho = 0.072$, FDR = 0.00467) and BM2 ($\rho = 0.108$, FDR = 0.00350), and negatively associated with erythroid weights in both specimens ($\rho = -0.095$, FDR = 0.00350; and $\rho = -0.066$, FDR = 0.03675, respectively). These findings placed KPRM 1 and KPRM 3 enrichment within myeloid-associated regions of the two GSE279576 specimens.

Spatial clustering of UCell enrichment was evident in 19 of 27 testable module–specimen combinations before adjustment, 12 after depth adjustment, and three after thinning. KPRM 5 retained significant clustering after thinning in both BM1 (Moran's I = 0.053, FDR = 0.00700) and BM2 (I = 0.209, FDR = 0.00700). KPRM 1 additionally remained clustered in BM1 (I = 0.027, FDR = 0.0455). Thus, depth sensitivity reduced the number of significant spatial patterns while preserving selected module–compartment relationships and localized enrichment. Thinned KPRM 5 maps are shown in Supplementary Figure S17E–F.

Signed netLDA contributions showed specimen-specific spatial organization. The Fisher-style projections revealed spatial patterns that differed from unsigned UCell enrichment (Supplementary Figure S8A–B). KPRMs 1–3 jointly contributed approximately 72–76% of overall projection variance across specimens, based on their covariance contributions to the total score. The most persistent module-level clustering occurred for KPRM 4 in BM1. Its original Moran's I was 0.053 (FDR = 0.00450), and clustering remained significant after count-depth adjustment (I = 0.035, FDR = 0.0360), detected-gene adjustment (I = 0.034, FDR = 0.0450), and thinning (I = 0.042, FDR = 0.00900).

KPRM 3 in AML12 showed the strongest original module-level Fisher clustering (I = 0.095, FDR = 0.00900). This pattern persisted after count-depth adjustment (I = 0.086, FDR = 0.0180) and detected-gene adjustment (I = 0.086, FDR = 0.00900). The thinned projection retained a positive Moran's I of 0.080, although its FDR increased to 0.0630. In BM2, the overall projection was clustered before adjustment (I = 0.040, FDR = 0.0240) and after count-depth adjustment (I = 0.041, FDR = 0.0225), but not after thinning. The AML12 KPRM 3 maps are shown in Supplementary Figure S8C.

Across all nine projection blocks and four specimens, significant clustering decreased from 10 of 35 testable comparisons in the original analysis to four after count-depth adjustment and two after thinning. Original and thinned overall-score rankings remained moderately correlated across specimens, with Spearman correlations ranging from 0.526 to 0.764.

Fisher-projection compartment associations were concentrated in BM2. KPRM 1 was negatively associated with erythroid weights in the original analysis ($\rho = -0.135$, FDR = 0.00450), after count-depth adjustment ($\rho = -0.092$, FDR = 0.00900), and after detected-gene adjustment ($\rho = -0.094$, FDR = 0.00450). Its association with monocytic weights was negative in the original analysis ($\rho = -0.093$, FDR = 0.0180) and after thinning ($\rho = -0.088$, FDR = 0.0150), but weakened after depth adjustment. These associations were significant with both reference definitions. The negative signed projections alongside positive UCell enrichment demonstrate that spatial module expression and the direction of fitted predictive contributions capture different features of the same gene sets. BM2 compartment associations are summarized in Supplementary Figure S8D.

Detection diagnostics identified particularly sparse KPRM 7 projections: the score was constant in BM1, and one spot accounted for almost all score variation in AML12. Spatial interpretation therefore centered on the larger modules.

A10.5 Gui single-cell treatment-response analysis

Single-cell analyses used eight baseline patients from Gui et al.'s GSE271406 dataset [36]: four published responders, comprising three patients with stable disease and one with morphologic leukemia-free state, and four patients with progressive disease. Full-module, hub-gene, and network-recruited-gene signatures were scored using UCell within the measured transcriptome. Scores were averaged within each patient and cell population, requiring at least 20 cells per group. Response comparisons used two-sided exact patient-level permutation tests, with BH correction across the planned module–population comparisons. Follow-up analyses examined CD8 subtype composition and sequencing-depth sensitivity, including exact thinning to 1,000 RNA UMIs per eligible CD8 effector-memory-2 cell over 20 draws. Signed projections using frozen netLDA coefficients provided complementary analyses.

In Gui's baseline single-cell cohort, CD8 memory T-cell UCell scores were higher in published responders for KPRMs 1, 4, and 5, with mean differences of 0.00549, 0.00764, and 0.02073, respectively (each exact P = 0.0286; q = 0.467). CD8 effector-memory-2 cells predominated within the accepted CD8 population. After restriction to this subtype and sequencing-depth thinning, response differences were attenuated and were not significant for KPRMs 1, 4, or 5 (P = 0.771, 0.229, and 0.143, respectively). Signed frozen Fisher projections likewise showed no significant response associations. The CD8 expression differences were sensitive to cell selection and sequencing depth. The single-cell response comparisons and sensitivity analyses are shown in Supplementary Figure S9A–D.

A11 Experimental CRISPR dependency and perturbation-expression evidence

A11.1 DepMap data and patient-unit comparisons

DepMap 24Q4 Public version 1 [37] provided CRISPRGeneEffect.csv and Model.csv. The downloaded matrix was checked against the release file size and MD5. Human gene symbols were parsed from matrix columns; ambiguous mappings were checked before restricting to the frozen 660-gene candidate universe. Of the 200 selected genes, 195 were covered; LPAR2, CSNK1E, CSF2RA, MAP3K14 and FGF16 were unavailable. AML status followed the recorded Oncotree primary-disease label. Other hematologic models comprised the myeloid and lymphoid lineages after excluding AML. These analyses used experimental Chronos fitness effects [38], with more negative values indicating greater dependency.

The screened matrix contained 28 AML models, 1,136 other cancer models and a subset of 105 other hematologic models. Models sharing PatientID were averaged gene by gene. The resulting eligible units were 26 AML, 1,093 other cancer and 103 other hematologic units; none had missing PatientID. HEL/HEL 92.1.7 and MOLM-13/MOLM-14 accounted for the two repeated AML source identifiers.

Gene-level contrasts used Welch tests with at least five finite units per group. BH correction covered all mapped candidate genes separately for each comparator, with untestable hypotheses retained in the planned family. Module analyses averaged covered gene effects within each unit, requiring at least 70% of frozen module genes to be mapped and at least 80% of mapped genes to have finite effects in a unit. Seven-module BH families were defined separately by comparator. Contributor sensitivities removed the top one, three or five KPRM 1 genes ranked by the observed AML mean dependency and assessed leave-one-gene-out effects; these were post hoc influence analyses.

A11.2 Matched candidate-gene-set benchmarks

The benchmark background contained 646 candidate genes with at least five finite units in AML, other cancer and other hematologic groups and valid non-AML mean/SD estimates. The original KEGG_AML_gene_membership.csv was verified against all frozen candidate genes and selected-gene annotations. Matching variables were the exact binary membership pattern across PI3K–Akt, JAK–STAT, MAPK, NF-κB and Wnt pathways and tertiles of mean and SD dependency among non-AML cancer units. No AML effects entered the matching strata. Size-only and baseline-only schemes were evaluated separately.

For each gene, the AML-minus-comparator contrast was the difference in patient-unit means. Module dependency strength was minus the equal-gene mean of these contrasts, so larger values indicate stronger AML-relative dependency.

This differs slightly from the earlier mean of unit-specific module scores when gene effects are missing. Modules required at least two eligible genes and at least 70% frozen-gene coverage. Conditional reference sets preserved the module gene count within every matching stratum, sampling without replacement within a set over 19,999 Monte Carlo draws. The primary pool included the observed module; a separate sensitivity excluded it. Infeasible strata were recorded without relaxing matching.

Upper-tail probabilities were P = [1 + number of reference strengths at least as large as observed]/20,000, including ties. BH correction included seven planned modules separately within each comparator, matching scheme and pool rule. Reference-distribution quantiles summarize variation among random gene sets. Matching feasibility, expected overlap and the number of possible sets were retained. Random seeds were fixed from base seed 404221 and the recorded analysis order.

KPRM 1 retained 62 of 64 genes. Against other hematologic cancers, observed strength was 0.02479 versus matched expectation 0.001348 (excess, 0.02344; P = 0.00340; q = 0.0238). The all-other-cancer benchmark gave P = 0.0107 and q = 0.0749. In the hematologic comparison, size-only and baseline-only matching gave q = 0.0343 and 0.03535; excluding KPRM 1 gave q = 0.00665 and 0.00980, respectively. Four singleton matching strata required inclusion of their KPRM 1 genes in the primary exact-match null; expected overlap was 15.26 of 62 genes. Excluding all KPRM 1 genes made exact matching infeasible. KPRM 7 had only one eligible gene and was untestable. Background sensitivities are shown in Supplementary Figure S10.

KPRM 1 patient-unit contrasts remained negative after removing RPS6, the three strongest contributors (RPS6, CDK6 and GRB2), or the five strongest contributors (also CFLAR and EIF4B). Hematologic-comparator P-values were 0.000313, 0.00644 and 0.00546, respectively. All 62 leave-one-gene-out contrasts remained negative; their maximum nominal P-value was 0.00477. These influence checks complement the seven-module comparison and matched-reference benchmark.

A11.3 GSE221578 assignments and module-response analysis

The deposited merged Seurat object contained 57,854 cells, seven recorded patient IDs and 15 sample labels. All 200 selected genes were present in the RNA counts assay. Exact reference matching verified 22,525 single-target assignments, with 4,245 NT controls concordantly labeled in both CRISPR_target and CRISPR_sgRNA. The 31,084 unassigned cells were kept separate from controls. Target and NT groups were matched within deposited patient and sample, retaining time

points and primary/PDX or in vitro/in vivo labels. Eligibility required at least 30 cells per arm, giving six CDK6 comparisons from five patients and ten comparisons each for JAK2 and FLT3 from seven patients. Guide support was assessed at ten cells per guide. The low-coverage in vivo 2263 sample was not eligible for these targets.

Unsigned UCell scores used the full measured RNA gene universe with maxRank = 1,500 [29]. CDK6 and JAK2 were excluded from their respective KPRM 1 signatures and FLT3 from its KPRM 4 signature, in both target and NT cells. All signatures retained complete measured coverage; the three- and two-gene modules were flagged as small signatures. Within-sample target-minus-NT mean differences were averaged equally across eligible samples within patient, and patients received equal weight in two-sided one-sample t-tests requiring at least three patients. BH correction covered 21 planned target–module tests. Exact patient sign-flip tests formed a separate sensitivity family.

Sensitivity analyses used 20- and 50-cell thresholds and supplied Progenitor/Mature_Myeloid annotations, with at least 20 cells per arm for within-state analyses. State tests formed a 42-comparison family. Depth sensitivity retained cells with at least 1,000 RNA UMIs, used binomial thinning to an expected total of 1,000 UMIs over three fixed-seed replicates, and averaged cell scores before patient-level contrasts. An unthinned analysis of identical retained cells separated selection effects from thinning effects. Guide-level contrasts and QC/composition summaries were descriptive; guides and repeated samples were not counted as independent patients.

No primary module response passed FDR <0.05. For KPRM 1, mean differences were +0.00129 after CDK6 perturbation (five patients; 95% CI, −0.01177 to +0.01435; P = 0.797), +0.000241 after JAK2 perturbation (seven patients; 95% CI, −0.00578 to +0.00626; P = 0.925), and −0.000806 after FLT3 perturbation (seven patients; 95% CI, −0.00374 to +0.00213; P = 0.526). The nominal CDK6–KPRM 2 decrease (−0.00762; P = 0.0387) had q = 0.539. No sensitivity family yielded an FDR-significant response (Supplementary Figures S11 and S12A).

A11.4 Patient-level pseudobulk gene responses

Raw RNA counts were summed within patient, sample and perturbation group using the same eligible comparisons. All measured genes entered aggregation. edgeR expression filtering used min.count = 10 and min.total.count = 15 across the two comparison groups. TMM normalization [39] was applied separately per target, followed by log2-CPM transformation with prior.count = 0.5. Matched target-minus-NT differences were averaged equally within patient, then analyzed with a one-sample limma model and empirical-Bayes variance moderation [40]. Patients received equal fitting weights; voom weights and cell-level tests were not used. The selected-gene BH family contained 597 planned non-target

comparisons, retaining filtered hypotheses in the family size. A separate genome-wide family contained all successfully tested expressed-gene comparisons across targets.

CDK6, JAK2 and FLT3 analyses tested 9,352, 10,086 and 9,816 expressed genes, including 152, 158 and 158 non-target selected genes, respectively. Thus, 468 of 597 selected-gene hypotheses were testable; none passed FDR <0.05, and the minimum selected-gene q-value was 0.629. Gene-response displays use the union of the five selected genes with the smallest nominal P-values for each target and are descriptive (Supplementary Figure S12B). KPRMs 5–7 had limited testable-gene coverage.

Target-transcript diagnostics gave mean paired log2-CPM changes of −0.371 for CDK6 (P = 0.151), −0.323 for JAK2 (P = 0.0248), and −1.783 for FLT3 (P = 0.0162; Supplementary Figure S12C). These target-transcript diagnostics were excluded from the primary selected-gene testing family. Guide-level expression differences and leave-one-patient-out means were saved as descriptive consistency checks.

References for this section correspond to manuscript references [29, 37, 38, 39, 40]. Complete effects, P-values, q-values, coverage flags, and analysis statuses accompany the statistical outputs.

A12 Clinical incremental value and specimen-matched drug sensitivity

A12.1 Clinical response model comparisons

The primary incremental analysis included 298 BEAT2 patients with an initial-diagnosis specimen collected on the day of study inclusion (205 CR/CRi and 93 refractory); the broader 306-patient analysis is retained in the preceding response section and Figure 6. ELN-complete comparisons included 286 of the 298 patients. GSE146173 included 214 patients (156 CR/CRi and 58 refractory). Each overall or KPRM score was evaluated separately, using fixed BEAT2-reference score standardization for these comparisons. These reference SDs differ from the earlier cohort-specific response summaries, so odds ratios are reported within their corresponding analysis specification.

Stratified five-fold cross-validation was repeated 20 times with identical folds for each clinical-only and score-added model pair. Logistic response coefficients were estimated within training folds; the expression signature and netLDA coefficients remained frozen. The comparison therefore evaluated response-model performance conditional on the established signature. Clinical covariates were age and sex, with additional ELN or FLT3-ITD/NPM1 models in BEAT2. AUC, Brier score, and calibration were calculated from held-out predictions. Quantiles across repetitions summarize sensitivity to splitting and are not sampling confidence intervals.

Mean cross-validated AUC increased from 0.605 to 0.663 after adding the overall score in BEAT2 and from 0.699 to 0.744 in GSE146173. In the same 286 ELN-complete patients, age/sex AUC increased from 0.604 to 0.671 after score addition; age/sex/ELN AUC changed from 0.718 to 0.719. The matched overall-score odds ratio attenuated from 0.566 (95% CI, 0.427–0.750) with age/sex adjustment to 0.837 (95% CI, 0.608–1.151) after ELN adjustment. Adding the score improved mean Brier score in the age/sex models but not the ELN model. These matched comparisons indicate overlap between the score and information captured by ELN.

A12.2 Drug-screen linkage and frozen-score reconstruction

Published BeatAML screens and the sample-identifier mapping [11] were linked using RNA specimen and subject identifiers, with explicit cross-checks of subject agreement. Adult initial-diagnosis specimens were selected independently of drug values and score availability. The primary set used study-inclusion day zero, selecting one specimen per patient. A broader diagnosis-stage sensitivity analysis selected the closest collection time to inclusion, with specimen-source and RNA-ID tie-breakers. Day zero denotes study inclusion. Explicit healthy-control labels were excluded. Screens required paper inclusion, convergence, finite AUC, and a valid concentration range. Exact duplicates were removed; nonidentical repeated specimen–drug screens were excluded without averaging. Increasing curves were retained with a flag and excluded in a separate sensitivity analysis. Floating-point-equivalent concentration endpoints were harmonized while preserving raw values.

The eight specified single agents were venetoclax, ruxolitinib (INCB018424), palbociclib, quizartinib (AC220), gilteritinib, crenolanib, midostaurin, and sorafenib. The primary set contained 317 patients and 1,972 screens; the broader diagnosis set contained 326 patients and 2,030 screens. The provided viability-curve AUC was analyzed continuously in its original units, with higher values indicating lower sensitivity. It is distinct from ROC AUC used for classification.

Frozen scores were completed for all eligible specimens using the original 649-specimen, 660-gene BEAT2 expression source. First-stage gene centering and sample-SD scaling used all 649 source specimens, reproducing the archived classification expression matrix. Frozen model centering/scaling was then applied to the 200 selected genes, followed by multiplication by their signed coefficients. Seven module contributions and the singleton contribution summed to the overall gene score. Scores excluded age and the intercept. Ninety-two specimens required newly reconstructed scores across the two drug populations. Maximum discrepancies were 8.88×10^{-16} for archived expression, 3.55×10^{-15} for archived gene scores, and 4.26×10^{-14} for existing specimen scores.

A12.3 Association models and sensitivity specifications

For each drug, linear models related viability AUC to each of eight scores, adjusting for age per ten years, sex, and specimen source (marrow versus peripheral blood/leukapheresis). Effects were expressed per fixed score SD across unique primary drug-analysis patients. HC3 covariance estimates and residual-degree-of-freedom t distributions provided two-sided tests and 95% confidence intervals. Models required at least 30 complete patients and 15 residual degrees of freedom. Constant nuisance covariates were recorded when omitted; rank-deficient fits or fits with extreme leverage were recorded as failed. BH adjustment retained all 64 planned drug–score hypotheses separately within each population and covariate specification. Unadjusted Spearman tests formed separate families.

Sensitivities excluded increasing curves, broadened diagnosis timing, and added ELN or FLT3-ITD/NPM1 status. Composition models used specimen-matched marrow or peripheral-blood blast percentages, recorded monocytic FAB categories (M4/M4eo/M5/M5a/M5b), or peripheral-blood monocyte percentages in blood specimens. Non-numeric ranges and inequality-coded percentages were recorded as unavailable without point imputation. Each basic and expanded comparison used identical complete patients. All 576 initial drug-association models and 384 matched-covariate comparisons succeeded. Patient-omission checks retained all patients in the reported full-data models.

Additional induction-response results

Higher frozen netLDA gene scores were associated with lower odds of CR/CRi in both bulk cohorts. In BEAT2, which included 211 CR/CRi and 95 refractory cases, the age- and sex-adjusted odds ratio per SD was 0.616 (95% CI, 0.473–0.804; q = 0.00190). In GSE146173, comprising 156 CR/CRi and 58 refractory cases, the corresponding odds ratio was 0.532 (95% CI, 0.369–0.768; q = 0.00308). Restricting BEAT2 to 298 inclusion-time-zero patients preserved the association (OR, 0.587; 95% CI, 0.445–0.774). In GSE146173, the association also persisted when deaths with residual disease were included as nonresponse (OR, 0.521; 95% CI, 0.371–0.731). All leave-one-patient-out estimates retained the same direction. Bulk response associations are shown in Figure 6A, with endpoint and inclusion-time sensitivity analyses in Supplementary Figures S20A and S1, respectively.

KPRMs 2, 3, 6, and 7 were associated with recorded response after BH correction in BEAT2. No individual KPRM passed correction in the GSE146173 primary analysis, although KPRMs 1, 3, 4, 6, and 7 did so under the expanded nonresponse definition. Hub-gene and network-recruited contributions were separately associated with response in both cohorts. In GSE146173, the hub-gene contribution added information beyond the network-recruited-gene contribution (likelihood-ratio P = 0.0110; q = 0.0154), whereas the additional network-recruited contribution did not reach significance

(P = 0.0956). None of the 200 individual-gene associations passed BH correction. Score-response curves and patient-level component distributions are shown in Figure 6B–C. Adjusted component models and response-definition comparisons are shown in Supplementary Figure S20; inclusion-time sensitivity is shown in Supplementary Figure S1.

A12.4 Differentiation-state adjustment and focused diagnostics

The official BeatAML malignant-cell-type workbook [11] supplied six RNA-derived scores: GMP-like, HSC-like, monocyte-like, progenitor-like, promonocyte-like, and cDC-like. These signed scores were not treated as fractions. Exact RNA-ID linkage provided all six scores for 306 of 317 primary patients, including 222 venetoclax-screened patients. State scores were standardized across available unique primary specimens. Models added the monocyte-like score, all six scores, or all six scores plus ELN to the basic covariates, with separate 64-test BH families. All 192 comparisons succeeded.

Overall, KPRM 1, and KPRM 6 venetoclax associations persisted after specimen-matched blast adjustment (190 patients; q = 0.0127 for each). FAB adjustment attenuated the associations. The three-gene KPRM 6 comprised PTK2, CCND1, and GNG11; removing any single signed gene contribution preserved its positive basic-model association. The overall, KPRM 1, and KPRM 6 basic associations also retained positive directions under every patient omission.

Differentiation adjustment substantially changed these patterns. The overall-score coefficient among 222 patients changed from 25.92 to −0.013 after monocyte-like adjustment (P = 0.998). KPRM 1 changed from 25.13 to −15.61 after six-state adjustment (95% CI, −26.93 to −4.29; P = 0.00710; q = 0.0576). In 217 patients with ELN, six-state-plus-ELN adjustment yielded −17.39 (95% CI, −29.18 to −5.60; P = 0.00404; q = 0.0369). KPRM 6 remained positive after monocyte-like adjustment (13.12; 95% CI, 4.90–21.34; q = 0.0336), but not FDR-significant with six-state adjustment (q = 0.251) or six-state-plus-ELN adjustment (q = 0.130).

The focused diagnostic compared all adjustment models on the same 217 patients. KPRM 1 coefficients were 24.91 in the basic model, 25.76 with ELN alone, −12.38 with the monocyte-like score, −17.65 with all six states, and −17.39 with all six states and ELN. Thus, the reversal accompanied differentiation adjustment rather than ELN alone or changes in patient eligibility. All nine focused original adjustment estimates were reproduced exactly. All 1,983 patient-omission models succeeded. For KPRM 1, six-state-plus-ELN coefficients remained negative in every omission (−19.84 to −15.99); KPRM 6 remained positive (10.04–13.45). Score variance-inflation factors ranged from 1.16 to 2.09. Partial-regression slopes reproduced the fitted multivariable coefficients. Extended diagnostics are shown in Supplementary Figure S13.

Other six-state-plus-ELN findings included a positive KPRM 7–palbociclib association (coefficient, 13.50; q = 0.0369), negative KPRM 1 associations with quizartinib, midostaurin, and sorafenib (q = 0.0369, 0.0369, and 0.0145), and negative KPRM 3–gilteritinib and KPRM 5–sorafenib associations (both q = 0.0369). Complete drug–score estimates and testing families accompany the statistical outputs. Focused sequential comparisons and omission ranges are diagnostic summaries; the original 64-test BH values remain the multiplicity reference.

A12.5 Independent Finnish venetoclax assessment

Cohort and endpoints. We used the Finnish VenEx/Helsinki data of Vänttinen et al. (manuscript reference [35]; Zenodo record 17777357). Exact study-subject identifiers linked all 74 RNA profiles to Supplementary Table S6: 53 published responders, 15 refractory patients, and six untreated patients classified as ex vivo venetoclax-resistant. The clinical endpoint retained the published response definition, including CR, CRi, CRh, and MLFS, rather than the induction CR/CRi endpoint used elsewhere in this study. The six untreated patients contributed to ex vivo analyses when DSS was measured but were excluded from clinical response models. Venetoclax DSS was available for 69 patients (50 responders, 13 refractory, and six untreated); clinical response was evaluable in 68 patients. The RNA subset comprised 36 de novo/newly diagnosed, 22 relapsed/refractory, and 16 secondary AML cases. The source describes pre–VEN–HMA drug screening; individual RNA collection dates were unavailable.

Expression processing and frozen scores. Fractional gene-level estimated read counts were retained without rounding. The 60,579 Ensembl genes shared by all 74 raw files defined the expression matrix. TMM factors were estimated from genes with CPM ≥1 in at least eight patients, using original full-file library totals; log-CPM values used an edgeR prior count of 2 [39]. Unique Ensembl-to-symbol mappings provided 199 selected genes. DAXX was restored from ENSG00000204209 through a reviewed identifier-specific mapping (NCBI Gene 1616), yielding complete coverage of all 200 model genes and all seven KPRMs. Each gene was centered and scaled across all 74 Finnish RNA profiles, followed by the frozen inner BEAT2 centering/scaling and signed coefficients. Neither age nor an intercept entered the gene scores. Score SDs were fixed across all 74 profiles before endpoint exclusions. The restored DAXX contribution exactly accounted for the changes in the overall and KPRM 3 scores, with maximum discrepancy 2.02×10^{-14}; other module scores were unchanged.

Common differentiation signatures. We extracted the top 30 genes per state from van Galen Table S3 [4], following the BeatAML signature definitions [11], including the combined GMP-like column. Genes had to be measured and variable in both cohorts. Available genes numbered 24, 27, 27, 23, 30, and 25 for GMP-like, HSC-like, Monocyte-like, Progenitor-

like, Promono-like, and cDC-like signatures, respectively. Genes were standardized separately within the BEAT2 reference expression cohort and the 74 Finnish profiles. PC1 loadings were estimated from BEAT2 expression, oriented toward the mean standardized signature expression, and frozen for Finnish projection. State scores were standardized within their respective reference cohorts. This common PCA implementation differs from the original BeatAML wave- and specimen-specific scaling and was evaluated alongside, rather than substituted for, the published-score analysis. The Monocyte-like PC1 explained 64.5% of its signature variance in BEAT2. Removing all 200 netLDA genes from the state signatures retained 24, 26, 24, 19, 30, and 25 genes, respectively.

Association models. KPRM 1 was the principal target; the overall and KPRM 6 scores were secondary targets. Finnish linear models used DSS, HC3 covariance estimates, and residual-degree-of-freedom t inference; positive coefficients indicate greater sensitivity. Clinical response models used mean bias-reduced logistic regression with Wald confidence intervals and tests; OR >1 indicates higher response odds. Basic Finnish models included RNA batch, age category (≤75 versus >75 years), and disease stage. Monocyte-like adjustment was primary; all-six-state adjustment and overlap-removed signatures were sensitivity specifications. Each base/expanded pair used identical patients. BH correction covered the three planned scores separately by endpoint, signature variant, state-adjustment specification, and base/expanded model. Comparable BEAT2 models used the frozen score SDs from the primary drug-analysis cohort and adjusted for age, sex, and specimen source. Their targeted comparisons report nominal P-values. Endpoint units were retained separately: lower BeatAML viability AUC and higher Finnish DSS both indicate greater sensitivity.

Results. All 48 Finnish and 24 BEAT2 common-state models succeeded. Among 228 BEAT2 patients, the KPRM 1 coefficient changed from 25.55 in the basic model to −9.94 after Monocyte-like adjustment (95% CI, −19.72 to −0.15; P = 0.0465) and −15.66 after six-state adjustment (95% CI, −26.78 to −4.54; P = 0.0060). Excluding netLDA genes from the state signatures retained the six-state association (−15.19; 95% CI, −26.18 to −4.20; P = 0.0070). In the Finnish cohort, corresponding KPRM 1 DSS estimates were 0.22 after Monocyte-like adjustment (95% CI, −3.51 to 3.94; P = 0.908) and 0.66 after six-state adjustment (95% CI, −4.82 to 6.14; P = 0.811). The overlap-removed six-state estimate was 1.74 (95% CI, −3.50 to 6.98; P = 0.509). The common scoring analysis retained the conditional BEAT2 association, while the Finnish estimates remained nonsignificant (Supplementary Figure S14; Supplementary Table S3).

Clinical response and phenotype sensitivities. The complete overall Finnish score was nominally associated with lower response odds without adjustment (OR, 0.489; 95% CI, 0.254–0.941; P = 0.0323; three-score q = 0.0968), but not after

clinical/technical adjustment (OR, 0.520; 95% CI, 0.263–1.031; P = 0.0610). Adding the Monocyte-like score yielded ORs of 0.520 for the overall score (95% CI, 0.263–1.028; P = 0.0600; q = 0.0917), 0.955 for KPRM 1 (95% CI, 0.459–1.986; P = 0.903), and 0.519 for KPRM 6 (95% CI, 0.261–1.031; P = 0.0611; q = 0.0917). No Finnish association passed BH correction. Separate recorded-phenotype analyses adjusted for enrichment method, FAB M4/M5 status, or CD34/CD38/CD117 category; all 36 models succeeded and none was nominally significant. The enrichment comparison excluded the two CD33-enriched RNA samples, to avoid unit leverage in the DSS models. Missing FAB values remained missing, and each sensitivity used matched base/adjusted patient subsets.

Supplementary Table S3 Finnish associations using common differentiation signatures

Endpoint / n	Score	State adjustment	Effect (95% CI)	P	BH q
DSS / 69	KPRM 1	Monocyte-like	0.22 (-3.51, 3.94)	0.908	0.990
DSS / 69	Overall	Monocyte-like	-2.14 (-5.23, 0.94)	0.170	0.511
DSS / 69	KPRM 6	Monocyte-like	-0.02 (-3.43, 3.38)	0.990	0.990
DSS / 69	KPRM 1	Six states	0.66 (-4.82, 6.14)	0.811	0.964
DSS / 69	Overall	Six states	-3.59 (-8.03, 0.85)	0.111	0.334
DSS / 69	KPRM 6	Six states	-0.12 (-5.57, 5.33)	0.964	0.964
Response / 68	KPRM 1	Monocyte-like	0.96 (0.46, 1.99)	0.903	0.903
Response / 68	Overall	Monocyte-like	0.52 (0.26, 1.03)	0.060	0.092
Response / 68	KPRM 6	Monocyte-like	0.52 (0.26, 1.03)	0.061	0.092
Response / 68	KPRM 1	Six states	1.53 (0.59, 3.98)	0.379	0.379
Response / 68	Overall	Six states	0.59 (0.26, 1.36)	0.215	0.379
Response / 68	KPRM 6	Six states	0.68 (0.29, 1.57)	0.363	0.379

Effects are DSS differences or response odds ratios per fixed score SD. Models include RNA batch, age category and disease stage plus the specified RNA states. DSS: HC3 95% CIs. Response: mean bias-reduced logistic Wald 95% CIs. BH families contain three scores per endpoint and adjustment. Higher DSS and OR >1 indicate greater sensitivity and higher response odds, respectively.

A13 Patient-grouped validation and KPRM stability

A13.1 Patient identifiers and original fold audit

The age-complete classification input comprised 568 BEAT2 specimens from 513 patients (239 in class 0 and 274 in class 1). Expression specimen identifiers were linked uniquely to dbgap_subject_id using dbgap_rnaseq_sample in the processed clinical file. All 568 specimens matched. Forty-five patients had multiple specimens; their one-year outcomes were consistent within patient. The audit of the original specimen-level reweighting folds identified 38 patients, contributing 86 specimens, whose records occurred in more than one fold. The new analysis retained the original eligibility population, including the original handling of recorded age zero, and kept all specimens from each patient together.

A13.2 Nested fitting and evaluation

Five stratified outer folds were assigned at the patient level with random seed 20261022. Within each outer training set, five further patient-grouped folds generated inner out-of-fold age and gene components. Each inner fit and the outer-training base fit reranked the fixed union of hub candidates by absolute standardized class-mean expression difference, resolving ties by gene name. Gene centers and sample standard deviations, age centers and scales, covariance/network estimates, selected genes, and coefficients were estimated only in the relevant fitting partition. The fixed settings were $\tau=20$, $\nu=200$, $\alpha=0.50$, $d=3$ and $nb=10$. All specimens from a patient stayed within a single fold at both levels. Training used equal specimen weights, matching the original model.

Inner out-of-fold component means and sample standard deviations defined the scaling for the logistic age–gene combination. Its fixed ridge penalty was $\lambda=1$, with the unpenalized intercept, unconstrained weights, and BFGS settings described in Section A1.5. The corresponding outer-training base model generated scores for outer held-out patients; the inner-derived transformations and logistic weights were then applied without using their outcomes. Full fitted objects, predictions, hub rankings, fold assignments, run status, and preprocessing parameters were checkpointed. All five outer analyses and their inner fits completed successfully.

Patient-level evaluation averaged the held-out specimen scores within each patient before calculating AUC. Specimen-level AUC was a secondary summary. We used 10,000 paired, outcome-stratified patient bootstrap resamples (seed 20261022) to calculate percentile 95% confidence intervals for AUCs and paired AUC differences, applying the same

resampled patient indices to all scores. These intervals condition on the saved out-of-fold predictions; they do not include repeated fold assignment or refitting variability. Hyperparameters, the candidate hub universe, and the processed expression input were held fixed.

A13.3 Classification results

The reweighted classifier achieved a patient-level AUC 0.746, compared with 0.647 for the original-form age-adjusted netLDA score, 0.717 for the age component, and 0.627 for the gene component (Supplementary Table S4). The paired AUC improvement was 0.099 versus the original-form score (95% CI, 0.047–0.152), 0.029 versus age (95% CI, 0.014–0.044), and 0.119 versus the gene component (95% CI, 0.063–0.176). Specimen-level AUCs were 0.728, 0.655, 0.696, and 0.638, respectively. Across the five outer folds, specimen-level reweighted AUC ranged from 0.686 to 0.755. Reweighting exceeded the original-form score in four folds and the age component in all five; in fold 5, the original-form and reweighted AUCs were 0.747 and 0.686.

Supplementary Table S4 Patient-grouped nested cross-validation performance

Score	Patient AUC	95% CI	Specimen AUC
Original-form age-adjusted netLDA	0.647	0.598–0.694	0.655
Gene component	0.627	0.578–0.675	0.638
Age component	0.717	0.673–0.760	0.696
Age–gene reweighted netLDA	0.746	0.703–0.787	0.728

Patient-level AUCs use 513 patients and mean held-out specimen scores; secondary AUCs use all 568 specimens. Confidence intervals are paired, outcome-stratified patient-bootstrap percentile intervals conditional on the saved predictions. The gene component was extracted from the age-adjusted netLDA fit.

A13.4 Module-stability assessment

We compared the five saved outer-training base models with the original frozen reference network, without additional fitting. In each fit, an undirected edge joined genes when $|\Omega_{jk}/\sqrt{(\Omega_{jj}\Omega_{kk})}|>10^{-8}$, excluding diagonal entries. The original module membership table was first verified against connected components under this definition. Gene-selection and hub-selection frequencies counted occurrence across the five fits. For each reference KPRM, retention was the fraction of its genes selected anywhere in the refitted network. Its best-matching component maximized Jaccard similarity, defined as

intersection size divided by union size. Matches were chosen independently for each reference module, allowing a refitted component to match more than one reference module and identifying splits and merges.

For each within-reference-module gene pair, unconditional co-membership was the fraction of fits selecting both genes in the same component. Conditional co-membership used only fits in which both genes were selected. Reported module summaries averaged these pair-specific frequencies. Coefficient-sign agreement was calculated among refits selecting the gene and compared with its reference coefficient sign. Stability summaries were descriptive because the five training sets overlap. Inner fits were excluded from the primary stability summary.

A13.5 Module-stability results

All five networks passed validation and selected 200 genes each, forming five to eight multi-gene components and nine to eleven singletons. Of the 200 reference genes, 62 were selected in all five fits and 136 in at least four. KPRM 1 retained a mean of 81.9% of its 64 genes, with mean best-component Jaccard similarity 0.606 (range, 0.366–0.871). Mean conditional pair co-membership was 87.7%, indicating that jointly retained genes commonly remained in the same component. EIF4B, PAK1, SIAH1, CSNK1E, and IFNGR2 were selected in all five fits; coefficient signs agreed in all fits for EIF4B, PAK1, SIAH1, and IFNGR2, and in four of five for CSNK1E (Supplementary Table S5; Supplementary Figure S15).

KPRM 5 had the highest gene retention (95.0%) but variable component boundaries (mean Jaccard, 0.492). KPRMs 2–4 had mean Jaccard similarities of 0.208, 0.276 and 0.267. Within the TP53-containing KPRM 3, TP53 was selected in four fits, DAXX in five, ATM in two and PTEN in one. All three KPRM 6 genes were retained together in three fits; two were retained in another fit and none in one. Its jointly selected pairs always shared a component, although the exact three-gene component was never recovered. The two-gene KPRM 7 was recovered exactly once. Thus, recurring genes and gene cores were more reproducible than the exact seven-component partition of the final reference model.

Supplementary Table S5 Reference KPRM composition across outer-fold refits

KPRM	Genes	Mean retention	Mean Jaccard	Conditional pair co-membership
1	64	81.9%	0.606	87.7%
2	51	64.7%	0.208	44.0%
3	44	69.5%	0.276	60.1%
4	17	68.2%	0.267	47.6%
5	12	95.0%	0.492	75.1%
6	3	73.3%	0.420	100.0%
7	2	70.0%	0.633	100.0%

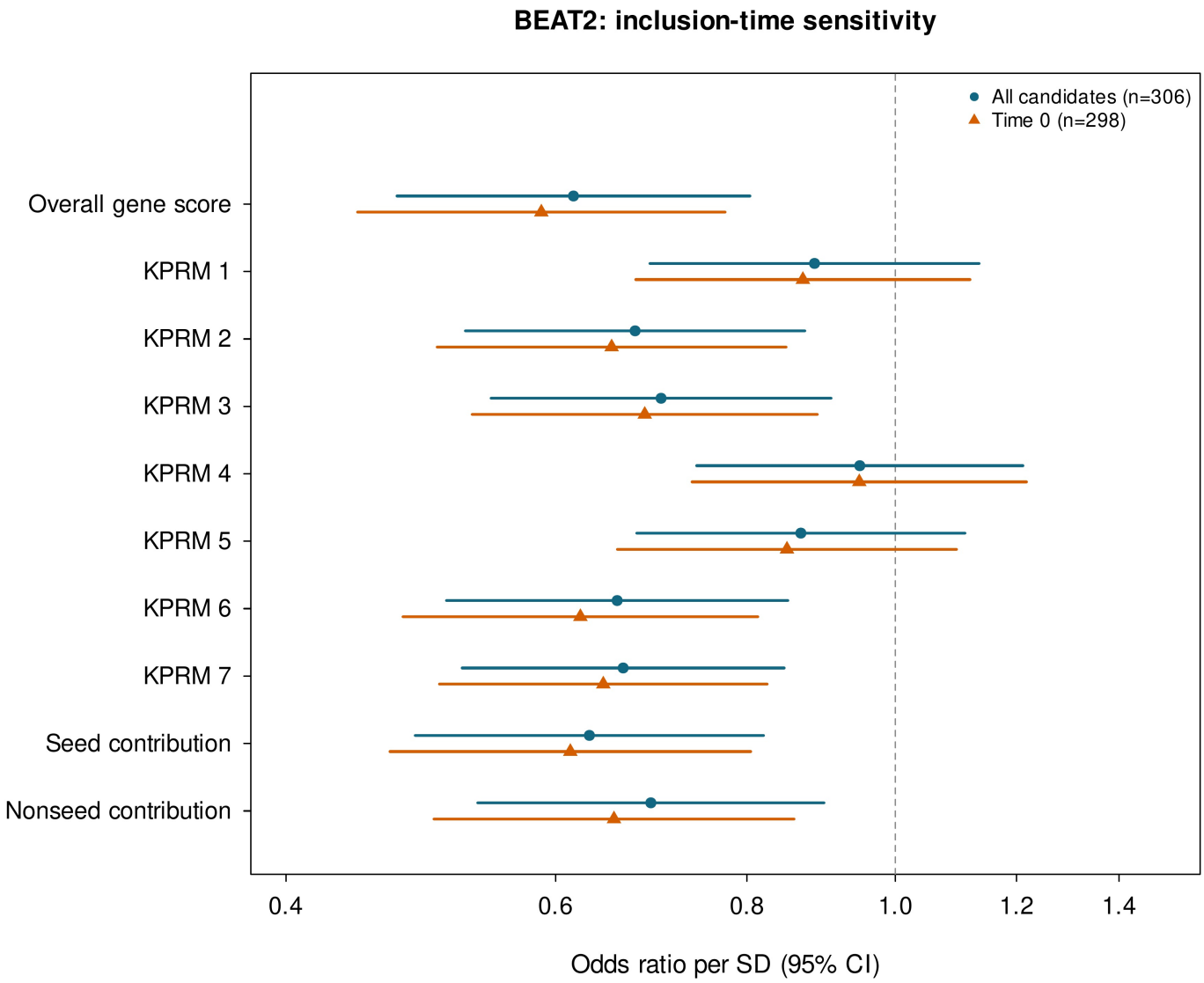
Each summary uses five outer-fold refits. Retention counts genes selected anywhere in the refit; Jaccard compares the full reference module with its best-matching component. Conditional pair co-membership averages pair-specific same-component frequencies among fits selecting both genes. Values for the two- and three-gene modules reflect few gene pairs. No multi-gene module except KPRM 7 was recovered exactly in any fit.

Abbreviations

AML	Acute myeloid leukemia
AUC	Area under the receiver operating characteristic curve
BEAT2	Beat AML model-development cohort
BH	Benjamini–Hochberg
BM	Bone marrow
cDC	Conventional dendritic cell
CI	Confidence interval
CPM	Counts per million
CRISPR	Clustered regularly interspaced short palindromic repeats
CR	Complete remission
CRi	Complete remission with incomplete hematologic recovery
DepMap	Cancer Dependency Map
ECM	Extracellular matrix
ELN	European LeukemiaNet
ESTIMATE	Estimation of STromal and Immune cells in MAlignant Tumor tissues using Expression data
FDR	False discovery rate
GEO	Gene Expression Omnibus
GMP	Granulocyte–monocyte progenitor
GSVA	Gene set variation analysis
H&E	Hematoxylin and eosin
HR	Hazard ratio
HSC	Hematopoietic stem cell
IQR	Interquartile range
KEGG	Kyoto Encyclopedia of Genes and Genomes
KM	Kaplan–Meier
KO	Knockout
KPRM	KEGG predictive regulatory module
LDA	Linear discriminant analysis
LMPP	Lymphoid-primed multipotent progenitor
ML	Machine learning
MLFS	Morphologic leukemia-free state
MPP	Multipotent progenitor
MSigDB	Molecular Signatures Database
netLDA	Network-based linear discriminant analysis

NK	Natural killer
NT	Non-targeting control
OR	Odds ratio
OS	Overall survival
PC	Principal component
PCA	Principal component analysis
pDC	Plasmacytoid dendritic cell
PH	Proportional hazards
PPI	Protein–protein interaction
PRM	Predictive regulatory module
QC	Quality control
RCTD	Robust cell type decomposition
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
scRNA-seq	Single-cell RNA sequencing
SD	Standard deviation
ST	Spatial transcriptomics
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TCGA	The Cancer Genome Atlas
TMM	Trimmed mean of M-values
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
WT	Wild type

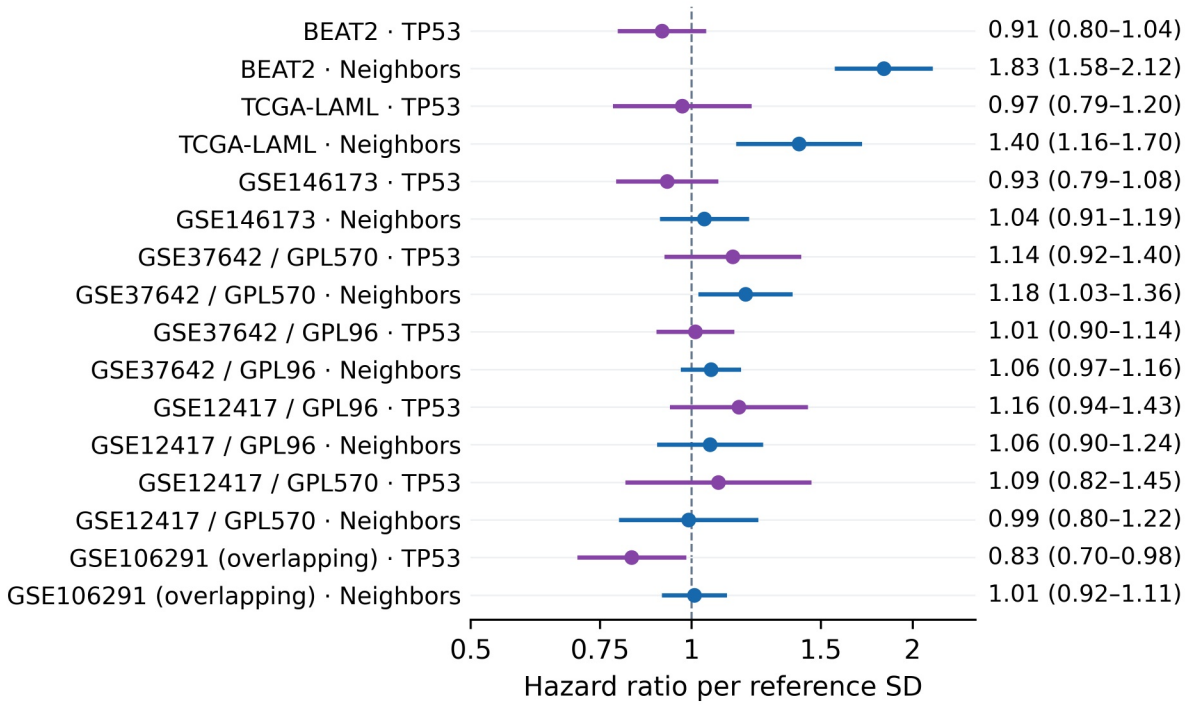
Supplementary Figure S1 Sensitivity of BEAT2 response associations to inclusion time



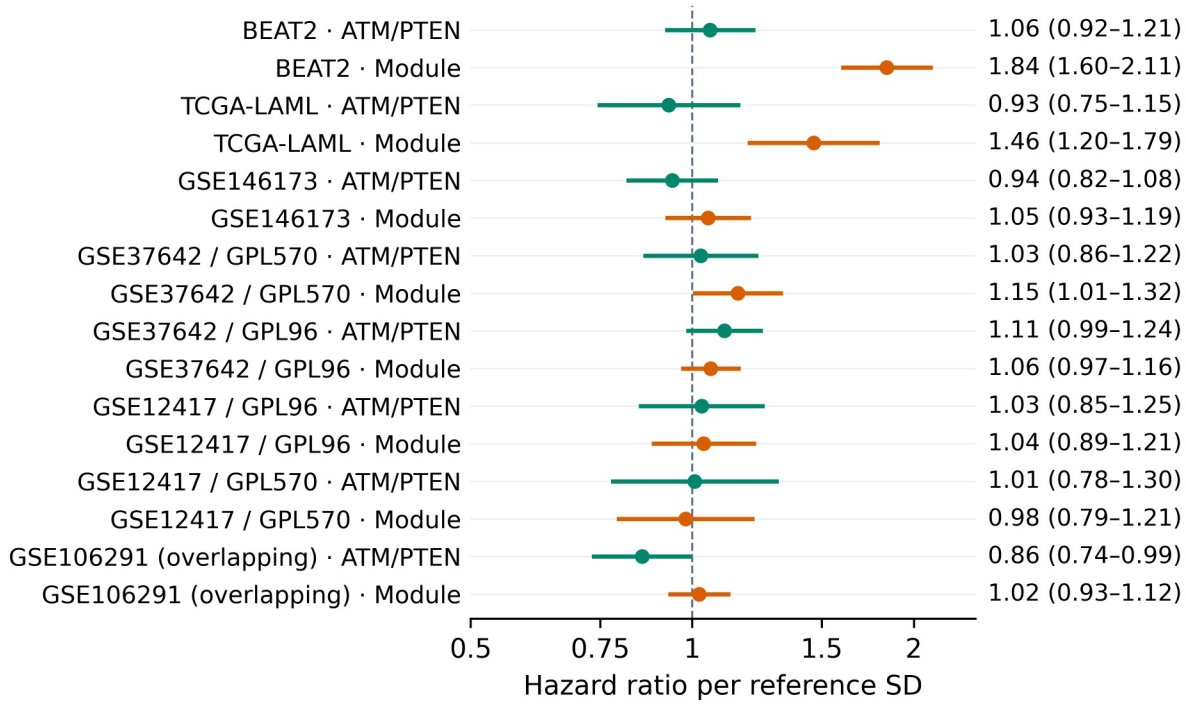
Supplementary Figure S1. BEAT2 induction-response associations by study-inclusion timing. Age- and sex-adjusted associations are shown for 306 unique diagnosis-stage patients (211 CR/CRi and 95 refractory) and their 298-patient subset with sample collection at study-inclusion time 0 (205 CR/CRi and 93 refractory). Each overall, KPRM, hub-gene or network-recruited-gene contribution was fitted separately, using the same frozen netLDA coefficients and full-candidate score standardization in both analyses. Points and horizontal bars indicate odds ratios per SD and 95% confidence intervals; the dashed line marks OR = 1. Higher scores indicate the original shorter-survival direction. Time 0 refers to study inclusion rather than verified induction-treatment timing. Full statistical results are provided in the accompanying tables.

Supplementary Figure S2 Survival associations of TP53 scores

A TP53 expression and neighbor score



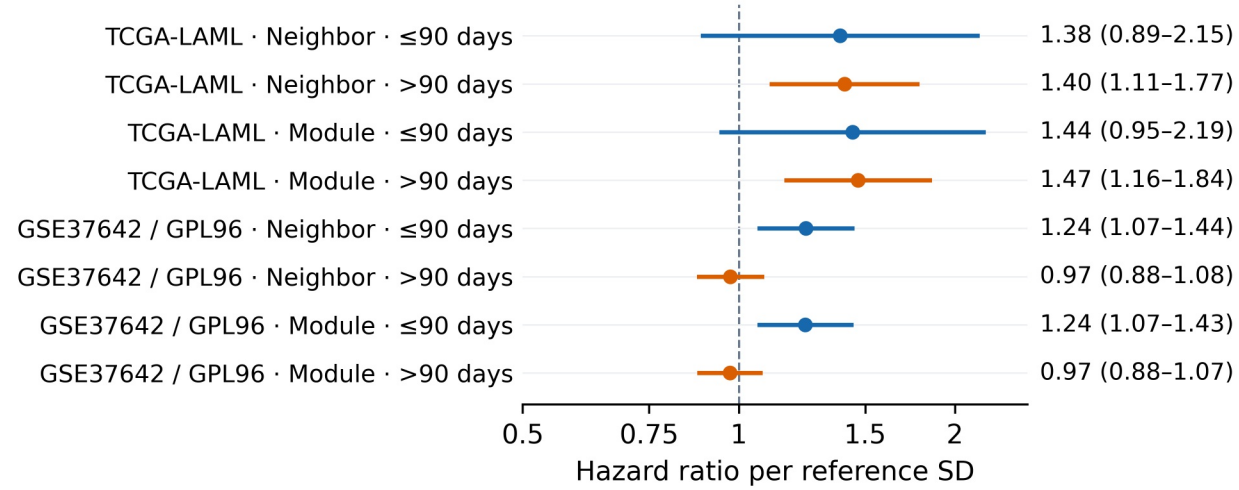
B ATM PTEN and module scores



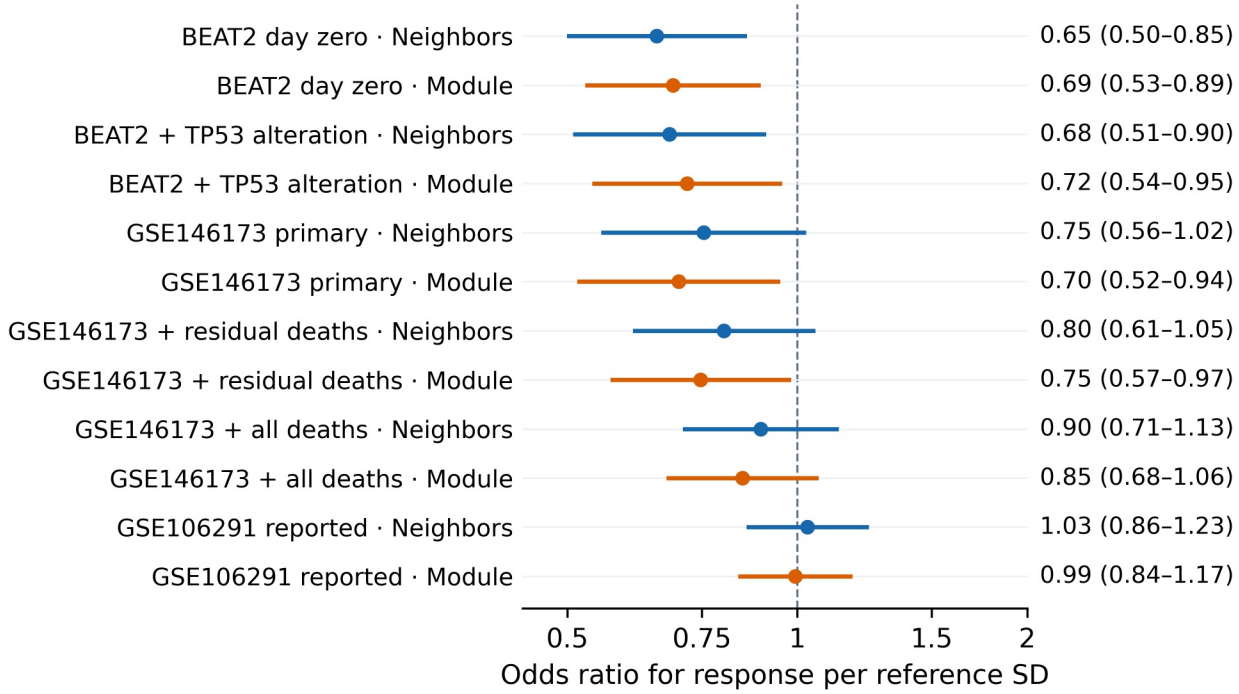
Supplementary Figure S2. Survival associations across AML cohorts. Four frozen TP53-related scores were evaluated in separate Cox models. (A) TP53 expression and the 43-gene neighbor score. (B) ATM/PTEN and the 44-gene module scores. Models adjust for age and sex in BEAT2 and age externally. Points indicate hazard ratios per frozen BEAT2-reference SD; bars indicate 95% Wald confidence intervals. GSE106291 overlaps with GSE146173. Estimates summarize full follow-up; time-specific associations are shown in Supplementary Figure S3. Gene weights were frozen, and Cox slopes were estimated within each cohort.

Supplementary Figure S3 Time and endpoint sensitivity analyses

A Early and later survival associations



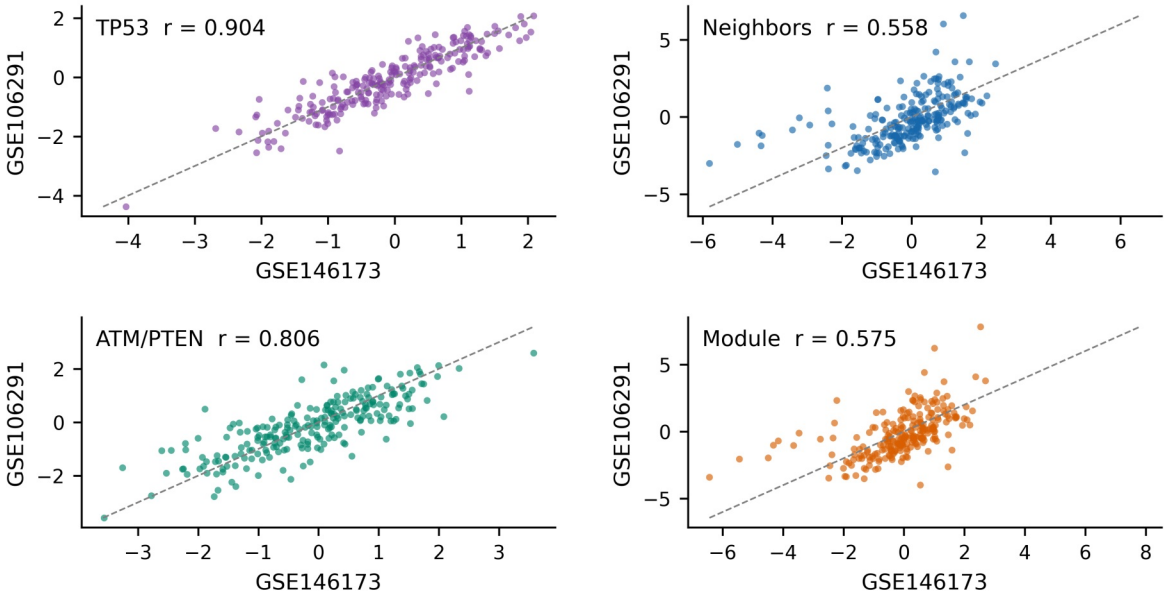
B Treatment response sensitivity analyses



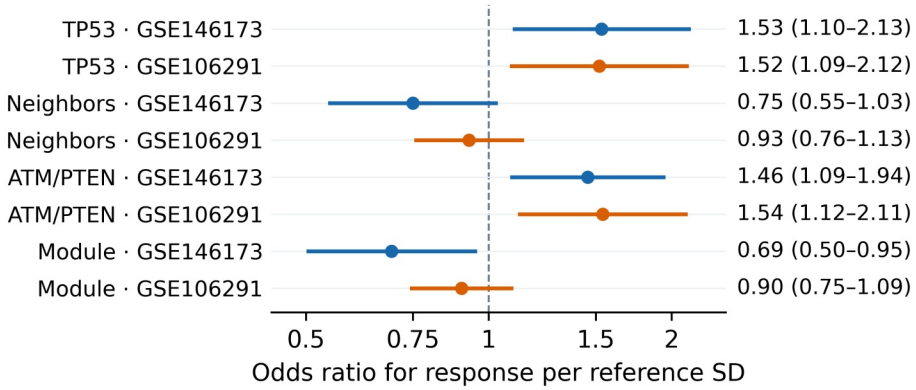
Supplementary Figure S3. Time-specific survival and response sensitivity analyses. (A) Neighbor and module hazard ratios during days 0–90 and after day 90 in TCGA-LAML and GSE37642/GPL96. Piecewise Cox models use separate interval baseline hazards, a common standardized age effect and patient-clustered robust covariance. (B) Response odds ratios from separate neighbor and module models adjusted for age and sex. BEAT2 alteration-adjusted estimates additionally include recorded TP53 alteration status among 284 eligible patients. GSE146173 endpoints comprise CR/CRi versus refractory disease (n = 214), additionally residual-disease deaths (n = 229), or all induction deaths (n = 246). GSE106291 uses reported sensitive/resistant labels (n = 235). Bars indicate 95% confidence intervals.

Supplementary Figure S4 Cross-accession and influence diagnostics

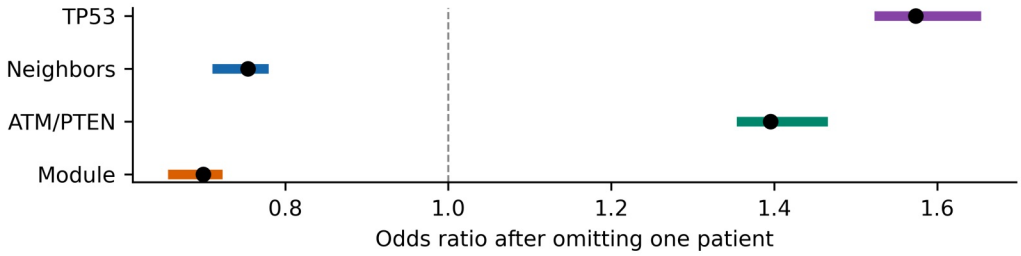
A Score agreement in shared samples



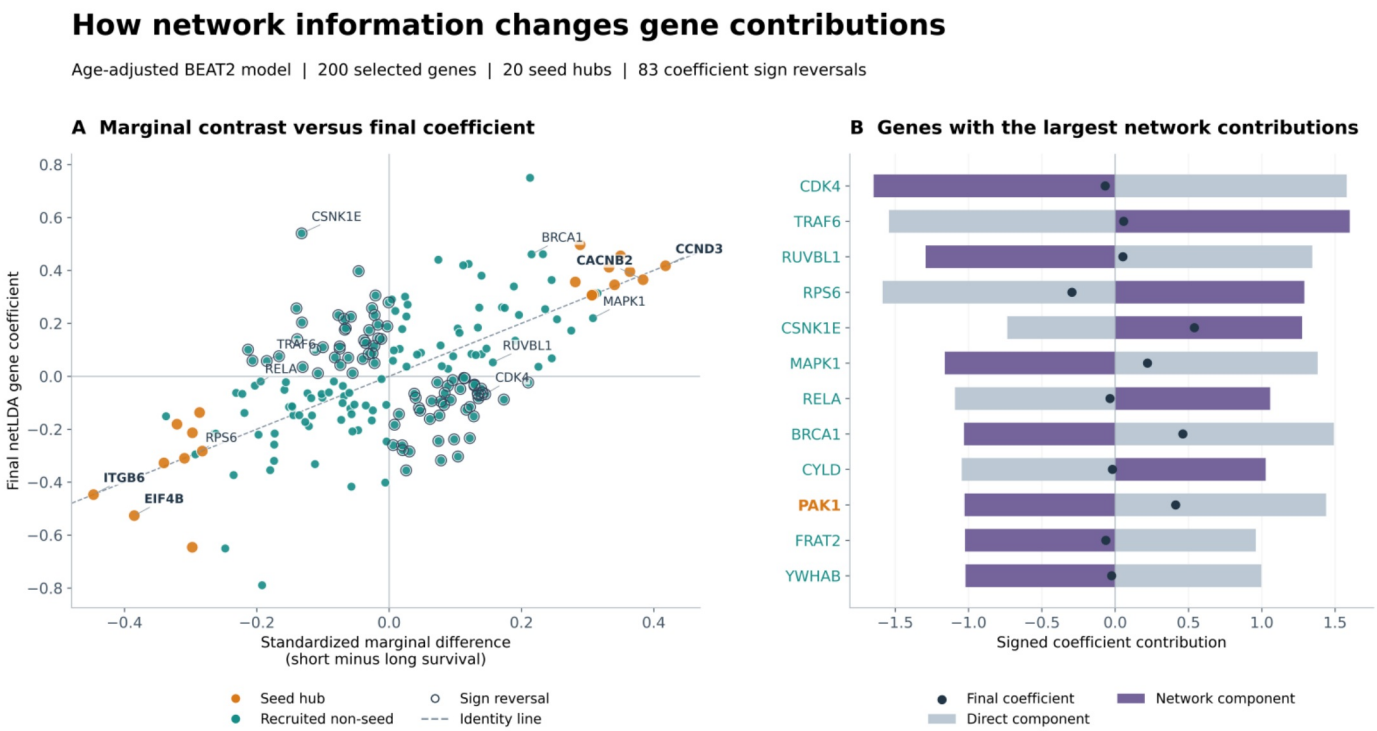
B Response associations in the same 208 patients



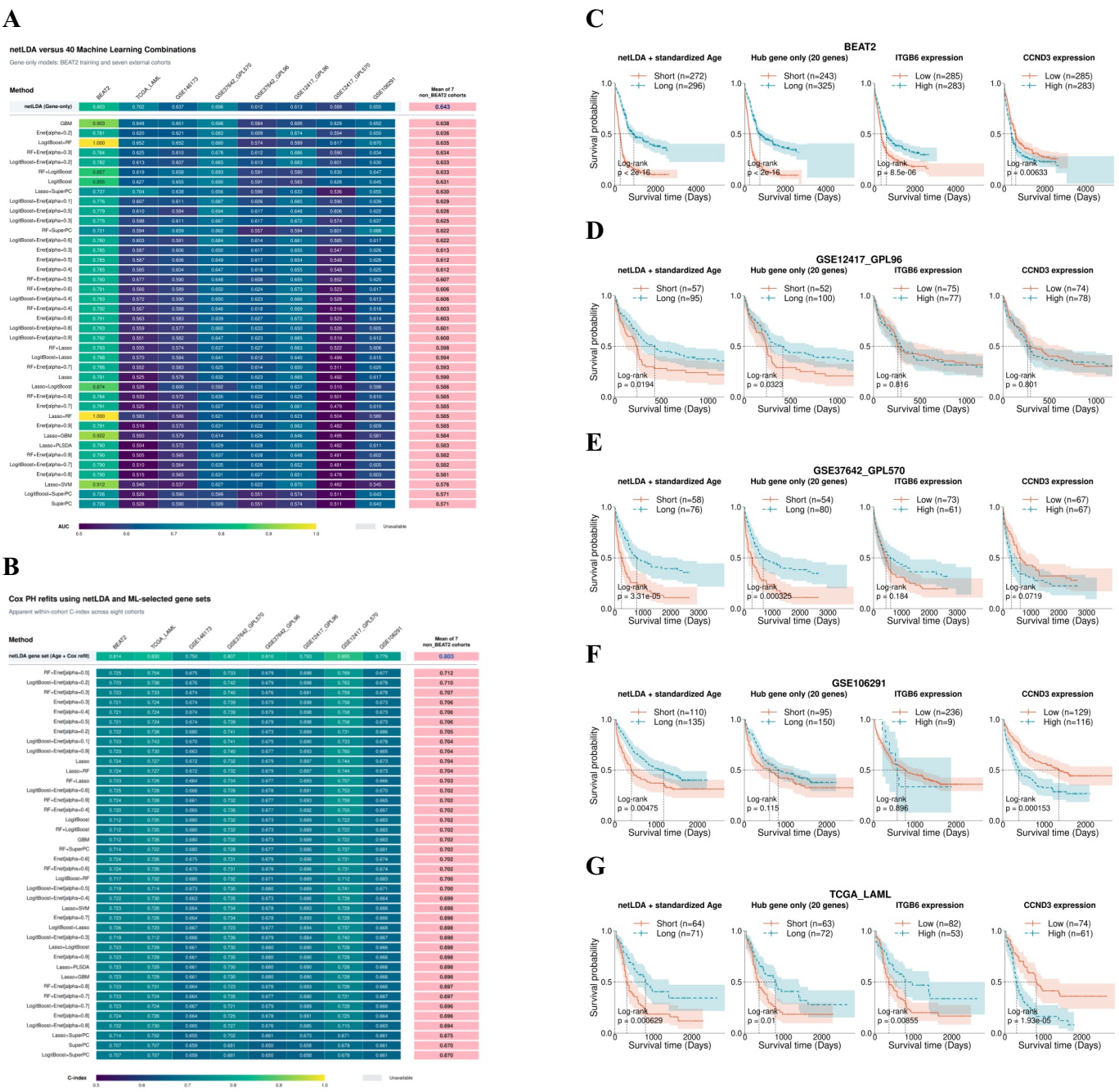
C Leave one patient out stability in GSE146173



Supplementary Figure S4. Cross-accession reproducibility and patient influence. (A) Frozen scores for 237 shared aliases with concordant age and sex; dashed lines indicate equality and r denotes Pearson correlation. (B) Response associations using the same 208 patients, GSE146173 CR/CRi-versus-refractory labels, and identical age and sex covariates, while substituting each accession's scores. Bars indicate 95% Wald confidence intervals. (C) Ranges of odds ratios across 214 leave-one-patient-out fits for each primary GSE146173 score model; black points show full-cohort estimates. These ranges are sensitivity ranges, not confidence intervals. Gene-contribution diagnostics are supplied separately.

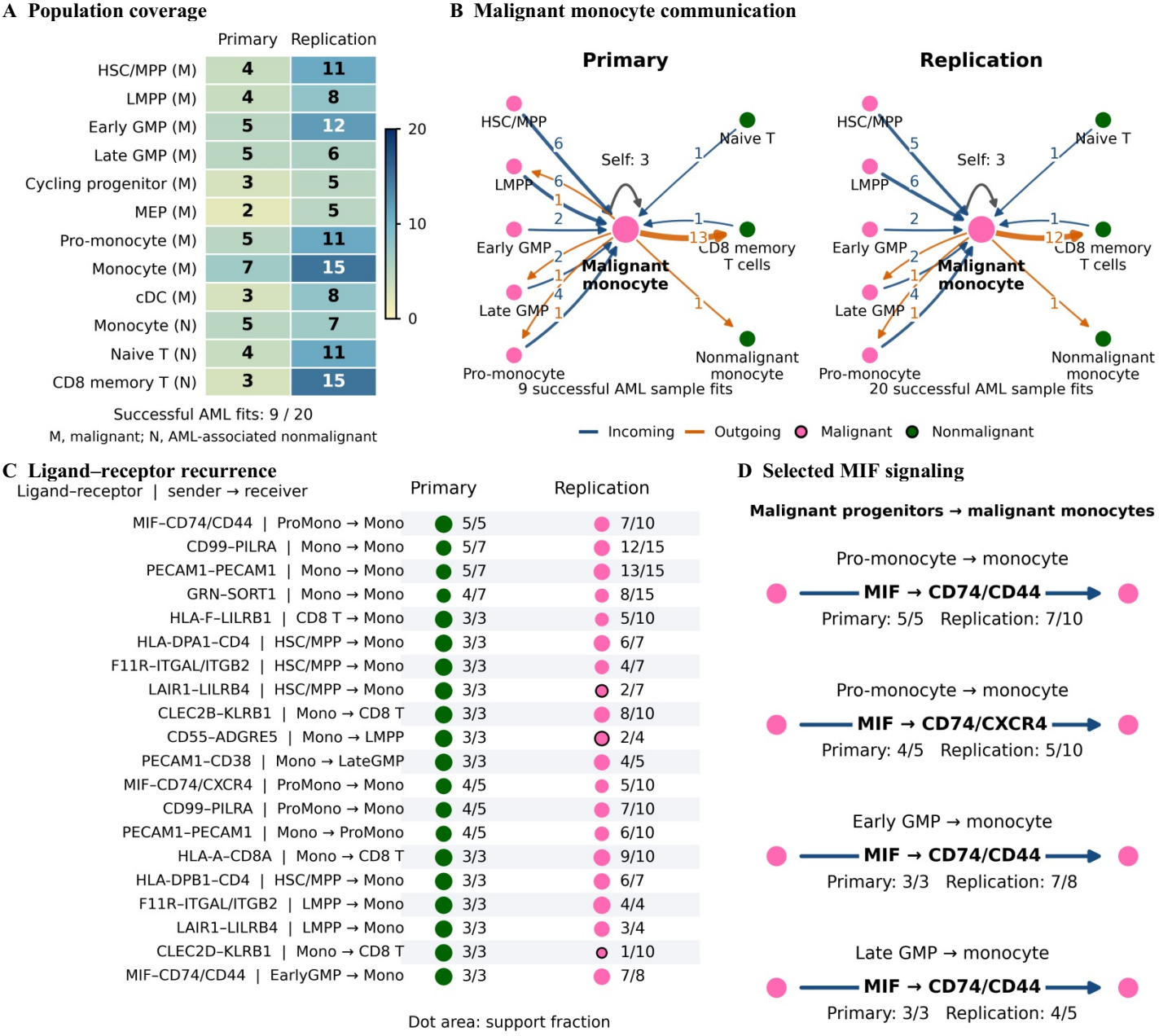


Supplementary Figure S5. Marginal and network contributions to netLDA gene coefficients. (A) Standardized marginal differences (short minus long survival) versus final coefficients for all 200 selected genes. Orange and teal distinguish hub genes and network-recruited genes; outlined points mark 83 coefficient sign reversals. Selected genes are labelled for readability. (B) Twelve genes ranked by absolute network contribution. Gray bars show the direct term $\Omega_{jj}\delta_j$, purple bars the network term $\Sigma_{k\neq j}\Omega_{jk}\delta_k$, and dots their signed sum, β_j. Same-sign terms are stacked; opposite-sign terms extend on opposite sides of zero. Age was standardized and entered the classifier separately; the gene precision network was not estimated conditional on Age. KEGG, Kyoto Encyclopedia of Genes and Genomes; KPRM, KEGG predictive regulatory module.



Supplementary Figure S6. Classification performance, Cox model discrimination, and survival stratification in AML. (A) AUCs for gene-only netLDA and 40 gene-only machine-learning (ML) combinations using up to 20 genes per ML model, trained on BEAT2 and evaluated in seven other cohorts. (B) Apparent Harrell C-indices from age-adjusted Cox proportional hazards models refitted separately within each cohort using the gene sets from gene-only netLDA and the 40 age-adjusted ML models. Age was standardized within each refit and retained separately. When the selected gene count exceeded half the usable sample size, gene-expression principal components explaining at least 95% of the variance replaced the individual genes. C-indices summarize apparent within-cohort discrimination. In both heatmaps, pink cells show the mean across seven non-BEAT2 cohorts and blue text marks the highest displayed mean. netLDA has the highest mean AUC (0.643 versus 0.638 for the leading ML comparator) and mean C-index (0.803 versus 0.712), and the highest C-index in every cohort. (C–G) Kaplan–Meier curves for BEAT2, GSE12417_GPL96, GSE37642_GPL570, GSE106291, and TCGA_LAML, respectively. Each row shows, from left to right, groups defined by netLDA with standardized age, a 20-hub-gene-only LDA classifier, ITGB6 expression, and CCND3 expression. Classifiers, age scaling, and expression cutoffs were learned in BEAT2 and reused externally. Low and high expression denote values at or below and above the BEAT2 median, respectively. Orange and teal indicate predicted short- and long-survival groups for classifiers, and low- and high-expression groups for individual genes. Shading denotes pointwise 95% confidence intervals; tick marks indicate censoring. Group sizes and unadjusted log-rank P-values are shown; time is in days. BEAT2 classification and KM results are apparent training results. AUC, area under the receiver operating characteristic curve; AML, acute myeloid leukemia. The ML panel was restricted after inspection of evaluation-cohort performance. Comparator results are descriptive; Cox estimates reflect apparent cohort-specific refit performance.

Cell–cell communication across AML datasets



C Ligand–receptor recurrence

Ligand–receptor | sender → receiver	Primary	Replication
MIF–CD74/CD44 | ProMono → Mono	5/5	7/10
CD99–PILRA | Mono → Mono	5/7	12/15
PECAM1–PECAM1 | Mono → Mono	5/7	13/15
GRN–SORT1 | Mono → Mono	4/7	8/15
HLA-F–LILRB1 | CD8 T → Mono	3/3	5/10
HLA-DPA1–CD4 | HSC/MPP → Mono	3/3	6/7
F11R–ITGAL/ITGB2 | HSC/MPP → Mono	3/3	4/7
LAIR1–LILRB4 | HSC/MPP → Mono	3/3	2/7
CLEC2B–KLRB1 | Mono → CD8 T	3/3	8/10
CD55–ADGRE5 | Mono → LMPP	3/3	2/4
PECAM1–CD38 | Mono → LateGMP	3/3	4/5
MIF–CD74/CXCR4 | ProMono → Mono	4/5	5/10
CD99–PILRA | ProMono → Mono	4/5	7/10
PECAM1–PECAM1 | Mono → ProMono	4/5	6/10
HLA-A–CD8A | Mono → CD8 T	3/3	9/10
HLA-DPB1–CD4 | HSC/MPP → Mono	3/3	6/7
F11R–ITGAL/ITGB2 | LMPP → Mono	3/3	4/4
LAIR1–LILRB4 | LMPP → Mono	3/3	3/4
CLEC2D–KLRB1 | Mono → CD8 T	3/3	1/10
MIF–CD74/CD44 | EarlyGMP → Mono	3/3	7/8

Dot area: support fraction
Black ring: recurrence criterion not met

D Selected MIF signaling

Malignant progenitors → malignant monocytes

Pro-monocyte → monocyte
MIF → CD74/CD44
Primary: 5/5 Replication: 7/10

Pro-monocyte → monocyte
MIF → CD74/CXCR4
Primary: 4/5 Replication: 5/10

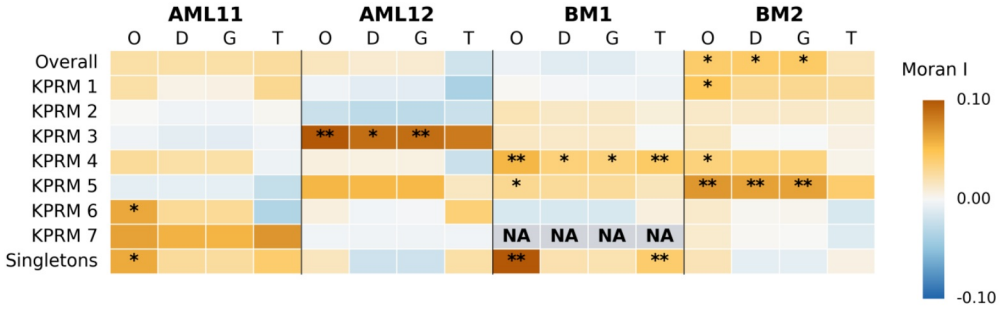
Early GMP → monocyte
MIF → CD74/CD44
Primary: 3/3 Replication: 7/8

Late GMP → monocyte
MIF → CD74/CD44
Primary: 3/3 Replication: 4/5

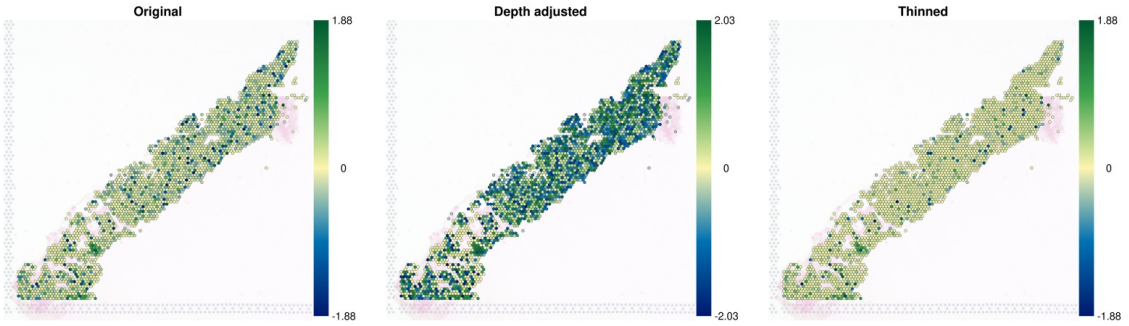
Supplementary Figure S7. Recurrent inferred cell–cell communication involving malignant monocytes in AML. (A) Number of successful AML sample fits containing at least 30 cells in each population, in GSE116256 (Primary; nine fits) and adult GSE185381 (Replication; 20 fits). (B) Networks restricted to the same Primary-priority ligand–receptor combinations involving malignant monocytes. Arrows show direction; widths and edge labels show numbers of recurrent combinations. Pink and green nodes denote malignant and AML-associated nonmalignant populations. (C) Twenty representative combinations selected by Primary support and pathway diversity. Dot area shows the fraction of eligible samples supporting each interaction; labels give supported/eligible sample counts. Black rings indicate failure to meet the replication recurrence criterion. Mono, monocyte; ProMono, pro-monocyte. (D) Selected MIF interactions directed toward malignant monocytes, with study-specific support counts. Sample-level support requires positive inferred probability and CellChat permutation $P < 0.05$; recurrence requires at least three eligible samples, at least three supporting samples, and support in at least 50% of eligible samples. In Replication, 39 of 42 monocyte-associated priorities and 89 of 96 priorities across the complete network met this criterion. Recurrence is descriptive and distinct from the KPRM correlation FDR analysis.

Spatial organization of netLDA gene contributions

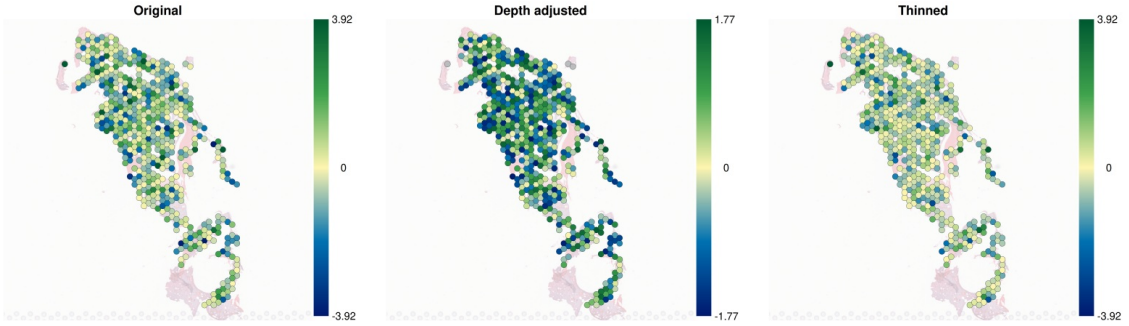
A Spatial autocorrelation across specimens



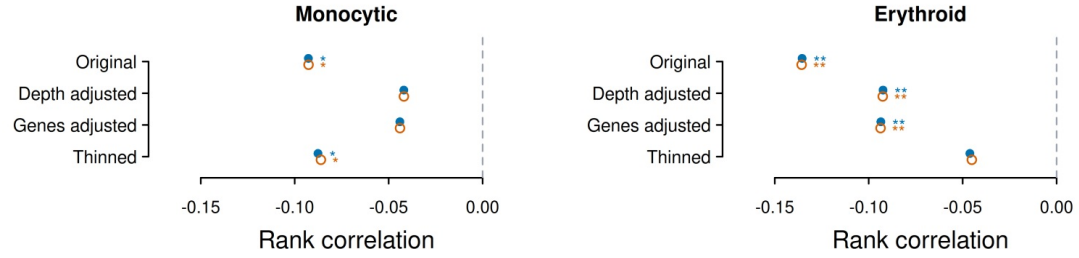
B KPRM 4 in BM1



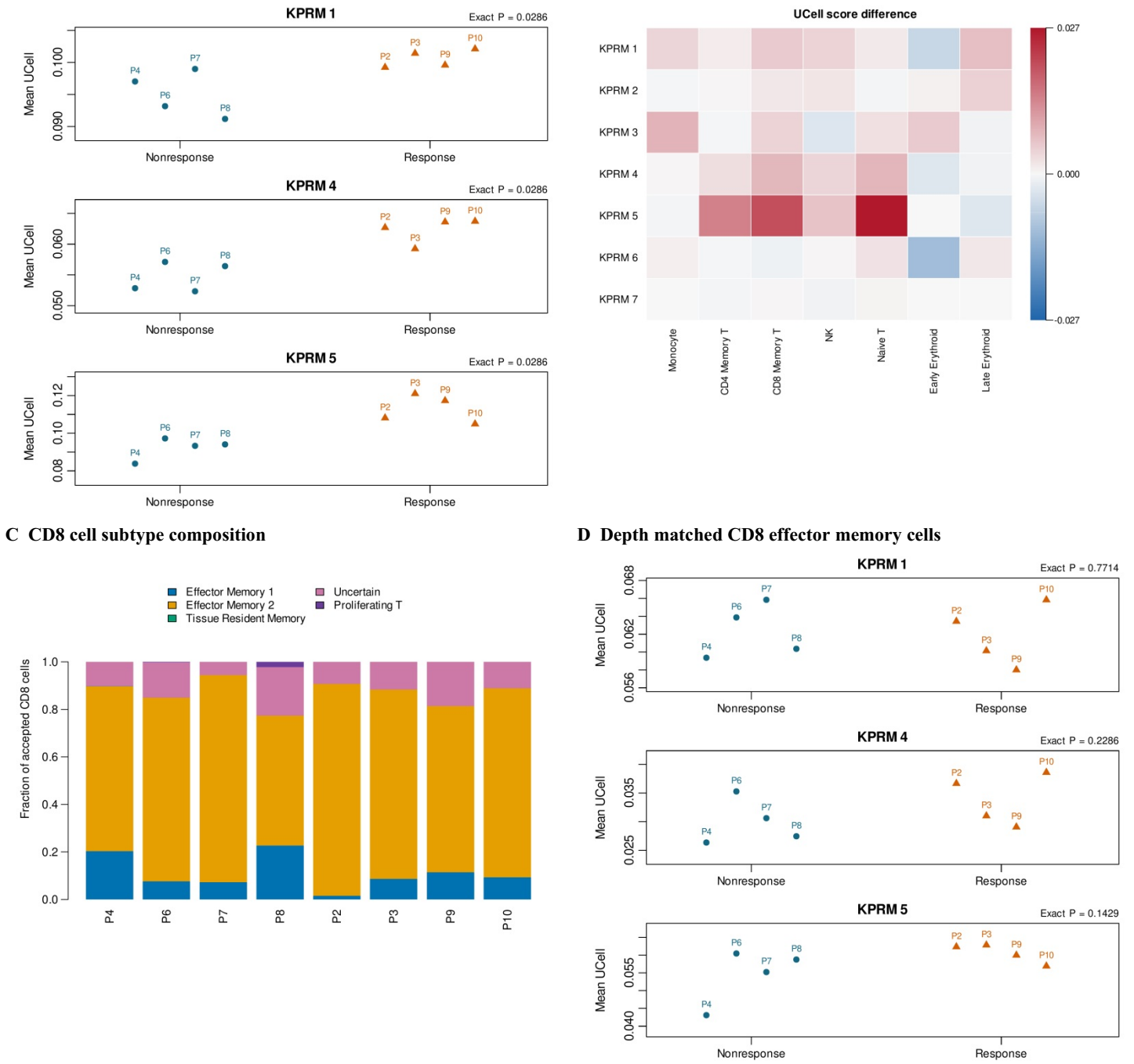
C KPRM 3 in AML12



D KPRM 1 compartment associations in BM2



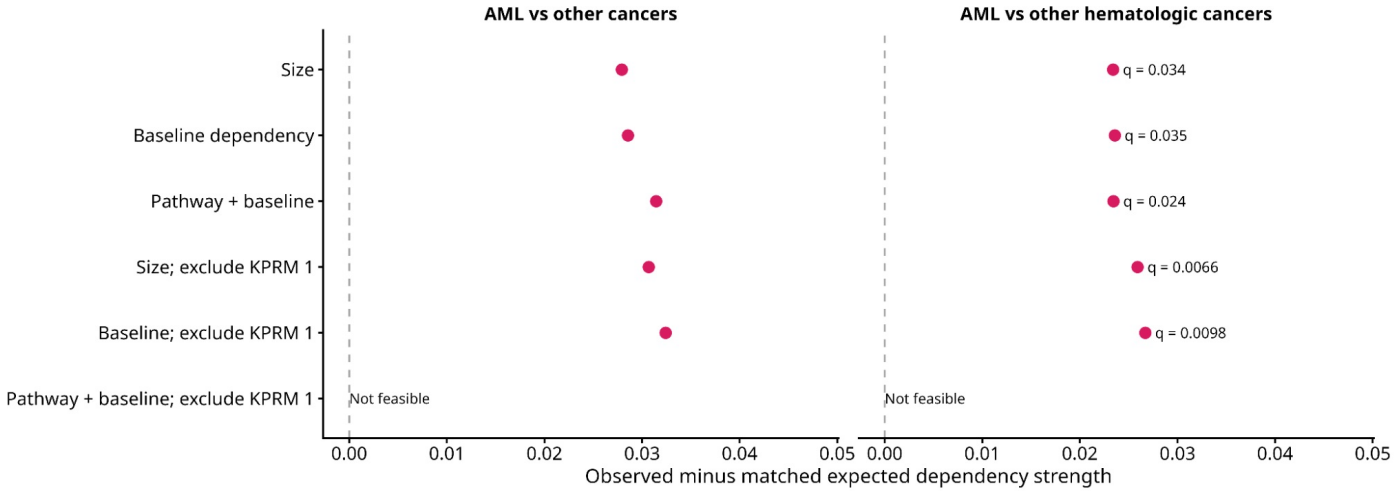
Supplementary Figure S8. Spatial distribution of netLDA gene contributions in AML bone marrow. (A) Moran autocorrelation of the overall gene projection, seven KPRMs and singleton block across four specimens. O, original; D, count-depth adjusted; G, detected-gene adjusted; T, thinned. (B) KPRM 4 in GSE279576 BM1 and (C) KPRM 3 in GSE255217 AML12, shown as original, depth-adjusted and thinned maps. Original and thinned maps share a symmetric color scale within each panel; adjusted maps show rank residuals in SD units. (D) KPRM 1 associations with monocytic and erythroid RCTD weights in BM2 under the primary and sensitivity references. Filled blue and open orange points denote the primary and sensitivity references, respectively. Stars indicate BH-adjusted significance: * FDR <0.05, ** <0.01, *** <0.001; gray NA cells are untestable. Clustering tests used 999 within-component permutations; compartment tests used 1,999 spectral sign surrogates. Gene projections use frozen BEAT2 coefficients and specimen-specific expression scaling, excluding age and the intercept. Maps B and C were selected after reviewing the spatial diagnostics. Viewports encompass the largest tissue component and other components containing at least 20 spots or 1% of retained spots, whichever is larger; display exclusions are audited.



Supplementary Figure S9. KPRM expression enrichment and treatment response in the Gui single-cell cohort. Baseline GSE271406 analyses included eight patients: four in the published responder group (three stable disease and one morphologic leukemia-free state) and four with progressive disease. (A) Patient mean full-module UCell scores for KPRMs 1, 4 and 5 in accepted CD8 memory T cells. Each point is one patient; P labels identify patient IDs. Displayed P-values are two-sided exact patient-level permutation tests. (B) Responder-minus-nonresponder differences in full-module UCell scores across seven KPRMs and seven cell populations. Red indicates higher scores in responders; blue indicates lower scores; gray indicates unavailable comparisons. (C) Fine-state composition within accepted CD8 cells, ordered by response group (P4, P6, P7, P8: progressive disease; P2, P3, P9, P10: published responders). (D) Patient mean scores in CD8 effector memory 2 cells after exact thinning to 1,000 RNA UMIs per eligible cell, averaged across 20 draws. Tests use the eight patient means. UCell quantifies unsigned expression enrichment. The three focus modules were chosen from the initial all-module results; full-family BH results are provided in the accompanying tables. Panels A and D use different cell selections and sequencing-depth conditions.

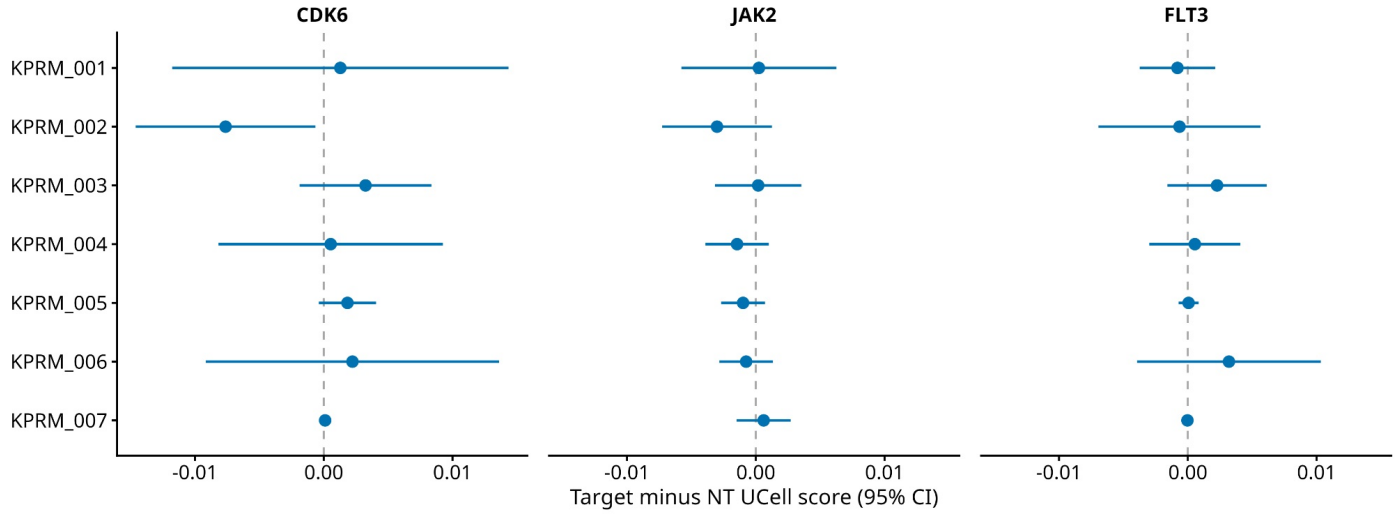
Experimental dependency background and module responses

Supplementary Figure S10 Matched-background sensitivity of KPRM 1



Supplementary Figure S10. KPRM 1 dependency relative to matched candidate-gene backgrounds. Points show observed minus expected strength for size-only, baseline-dependency and exact pathway-plus-baseline matching, using reference pools that include or exclude KPRM 1. Positive values indicate excess AML-relative dependency. Feasible nulls used 19,999 draws; labels show q <0.05 across seven modules separately per comparator, matching scheme and pool rule. Exact matching without KPRM 1 was infeasible. The displayed ranges summarize conditional gene-set benchmarks.

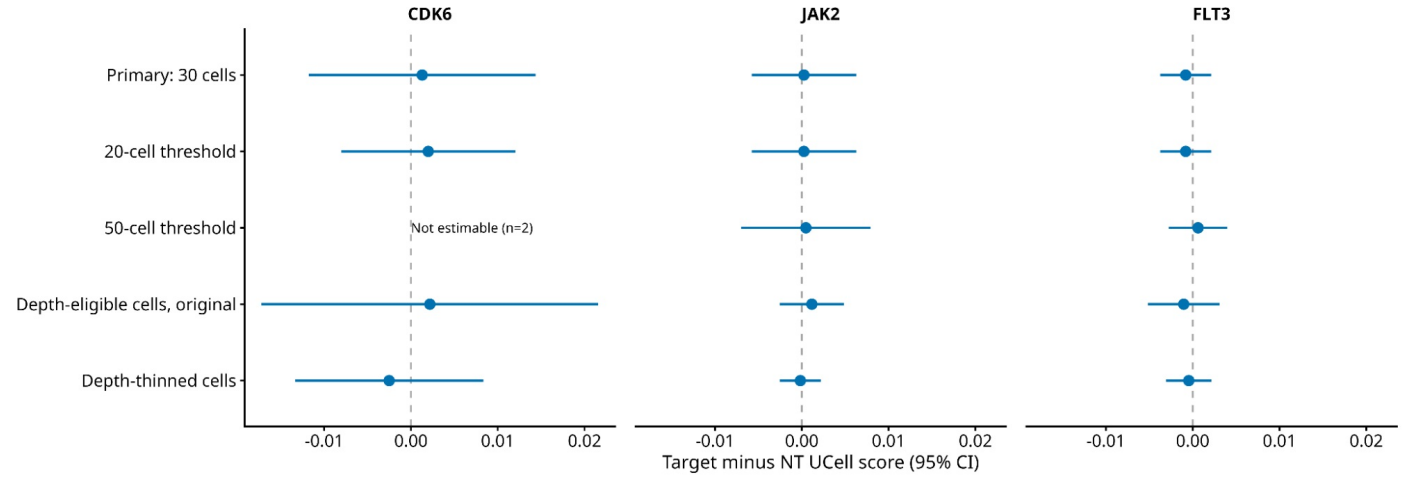
Supplementary Figure S11 Perturbation-expression responses across seven KPRMs



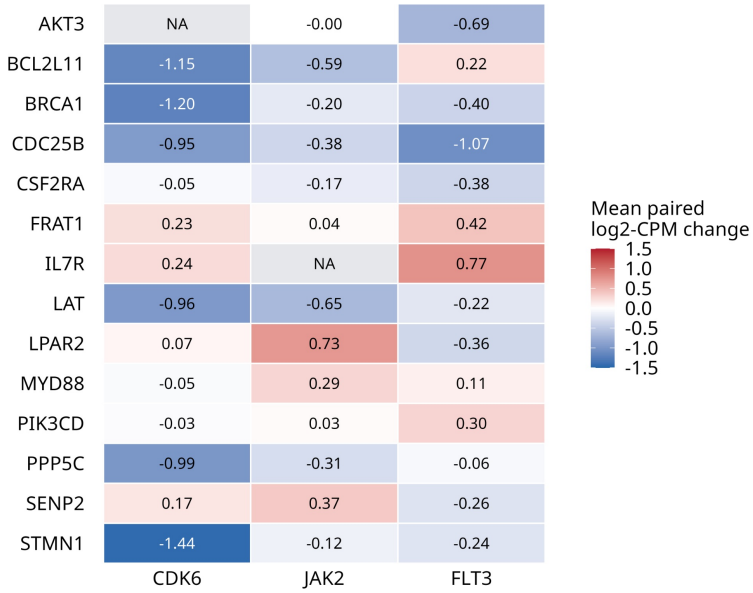
Supplementary Figure S11. GSE221578 module-enrichment responses to CDK6, JAK2 and FLT3 perturbation. Target-minus-NT UCell differences were matched within sample and averaged within patient, with five, seven and seven eligible patients, respectively. The targeted gene was removed from its own module signature in both arms. Bars show 95% t confidence intervals. No comparison passed BH correction across the 21 primary tests. Complete P-values and q-values are provided with the statistical outputs.

Perturbation-expression sensitivity and gene diagnostics

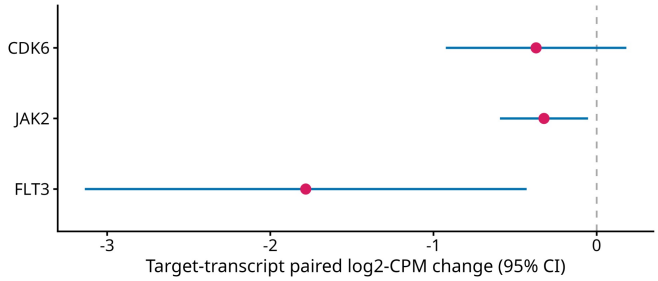
A KPRM 1 cell-count and depth sensitivity



B Descriptive gene responses



C Target-transcript changes



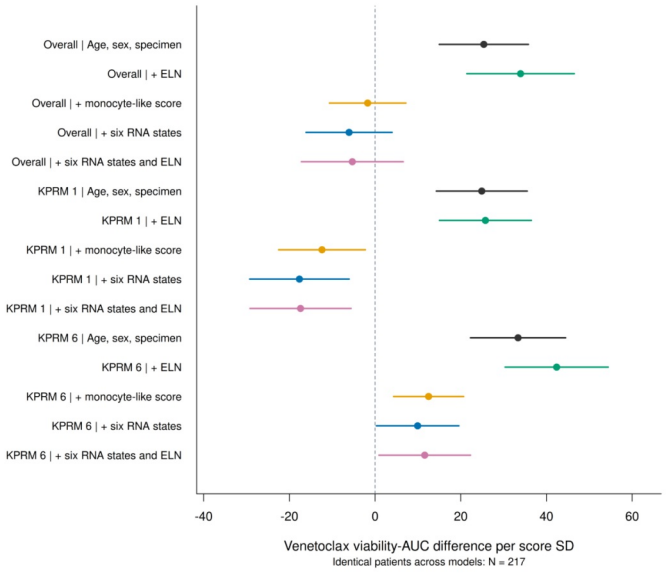
Supplementary Figure S12. Sensitivity analyses and gene-level responses in GSE221578. (A) KPRM 1 UCell contrasts under alternative cell thresholds and depth sensitivity. Thinning targeted an expected 1,000 UMIs over three replicates; the depth-eligible original analysis used identical cells without thinning. Bars show 95% confidence intervals; insufficient-patient comparisons are labeled. (B) Mean patient-level paired log2-CPM changes for the union of the five non-target selected genes with the smallest nominal P-values per target. Colors saturate at ±1.5; numerical labels retain the fitted changes. Gray cells were untestable. No gene passed correction across the 597 planned selected-gene comparisons. (C) Target-transcript changes with limma moderated 95% confidence intervals. Target-transcript rows were diagnostic and excluded from the selected-gene testing family.

Extended venetoclax differentiation diagnostics

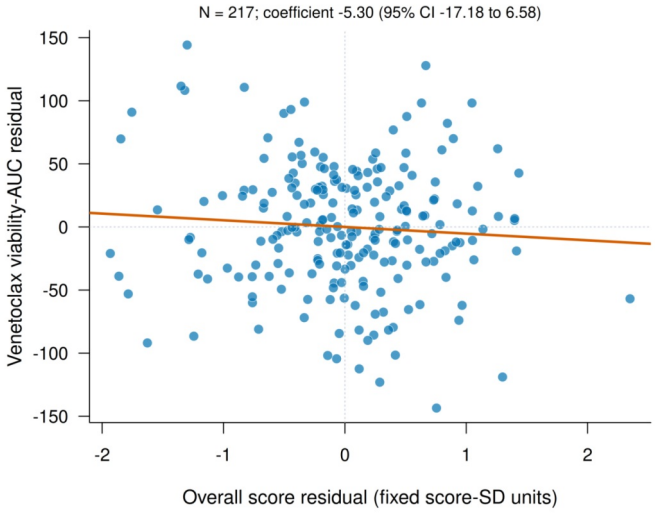
A All gene scores and differentiation states



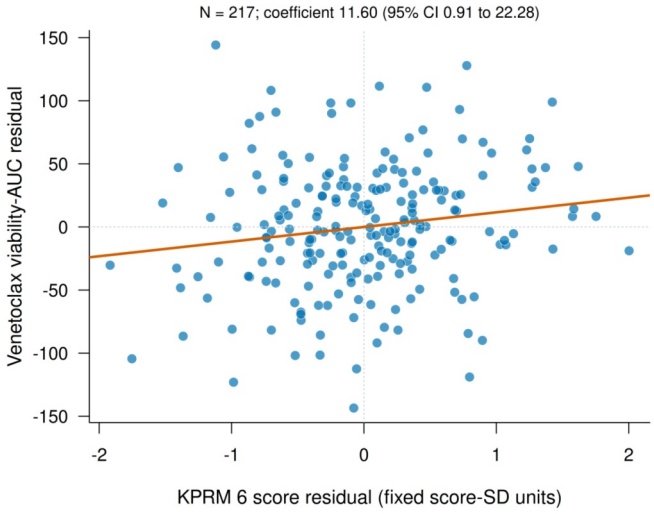
B Complete adjustment comparisons



C Adjusted overall score association



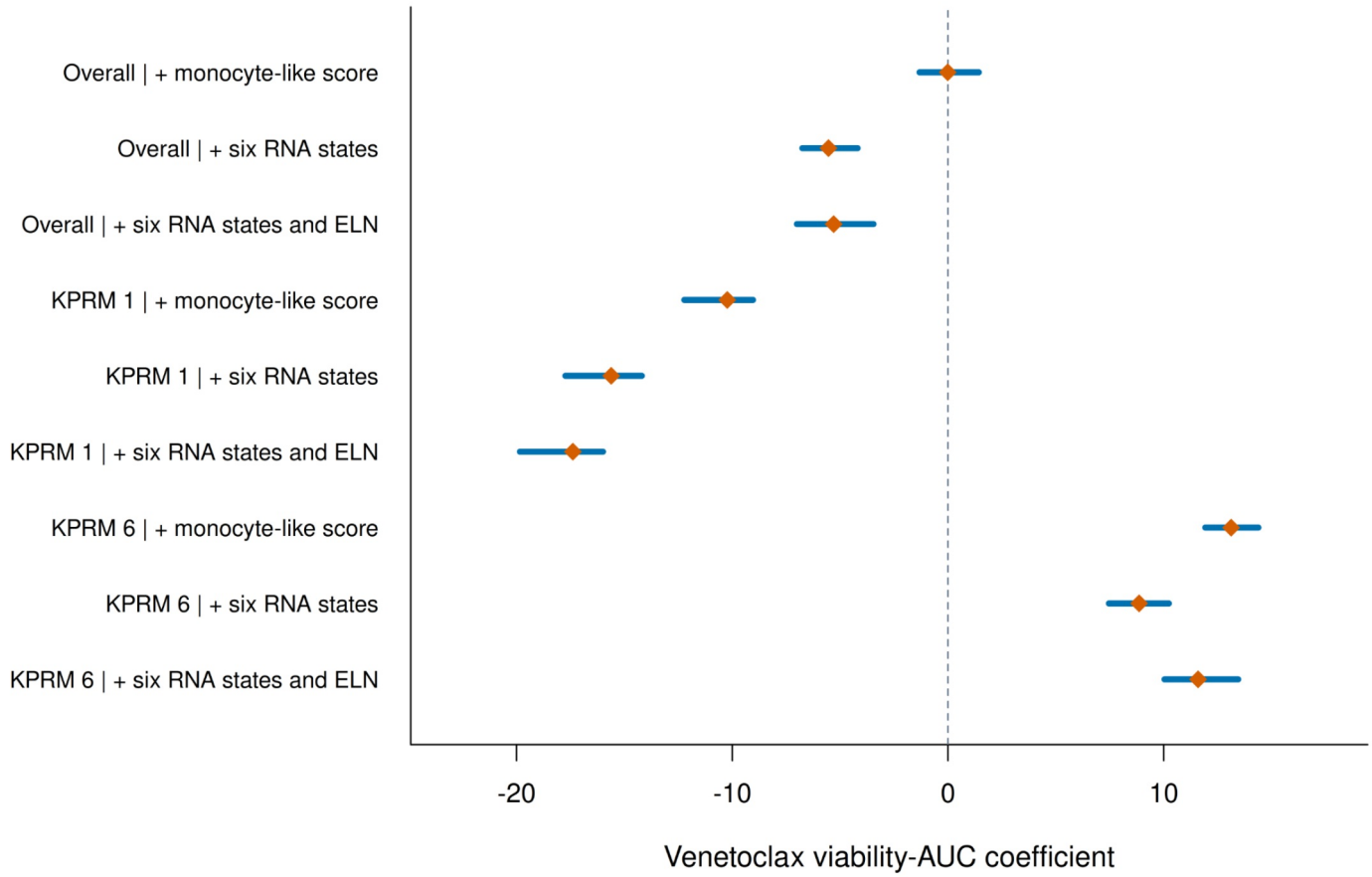
D Adjusted KPRM 6 association



Supplementary Figure S13. Extended differentiation-state diagnostics for the BeatAML venetoclax analysis. (A) Pearson correlations of all eight frozen gene scores with six published RNA-derived differentiation scores among available venetoclax specimens (n = 222). (B) Coefficients and HC3 95% confidence intervals from models fitted to the same 217 complete patients, with basic, ELN, monocyte-like, six-state, and six-state-plus-ELN adjustment. Basic covariates are age, sex, and specimen source. (C–D) Partial-regression plots for the overall and KPRM 6 scores after adjustment for basic covariates, six differentiation scores, and ELN. Higher viability AUC indicates lower ex vivo sensitivity; effects use the fixed primary-cohort gene-score SD. Panel E continues on the next page.

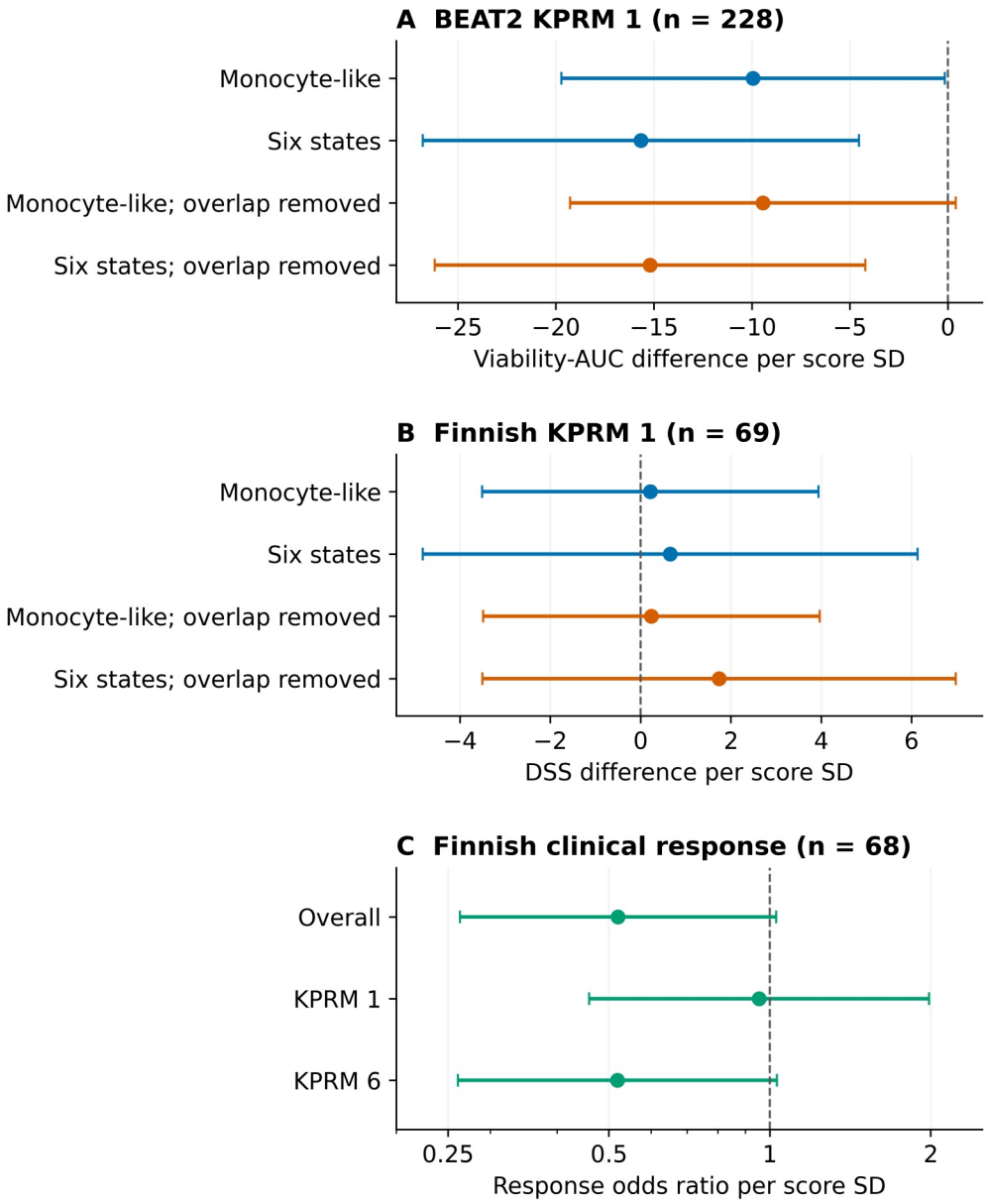
Extended venetoclax differentiation diagnostics continued

E Patient omission stability across adjustment models



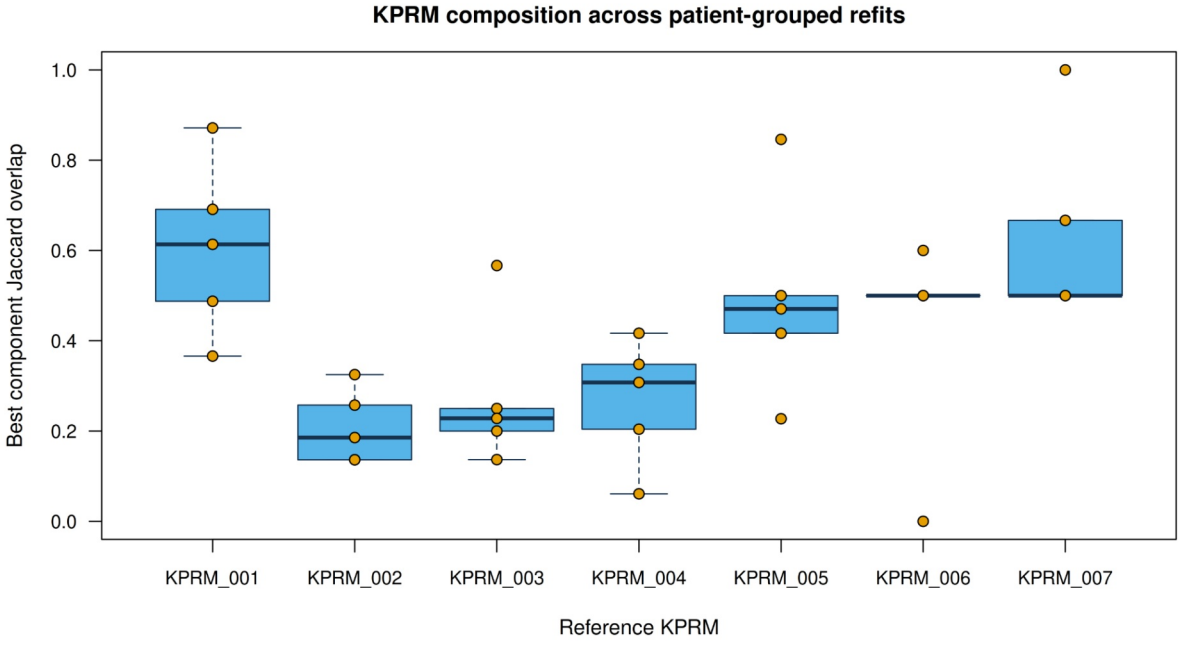
Supplementary Figure S13 continued. (E) Leave-one-patient-out coefficient ranges for the overall, KPRM 1, and KPRM 6 scores under monocyte-like, six-state, and six-state-plus-ELN adjustment. All models include age, sex, and specimen source. Each specification uses the same patients as its corresponding full-data model (n = 222 without ELN and n = 217 with ELN). Blue segments show the range across patient omissions; orange diamonds mark full-data estimates. These ranges summarize patient influence rather than sampling confidence intervals. Gene scores retain signed frozen gene contributions without an age contribution. Full numerical estimates, P-values, and BH-adjusted results accompany the statistical tables.

Supplementary Figure S14 Independent Finnish venetoclax assessment

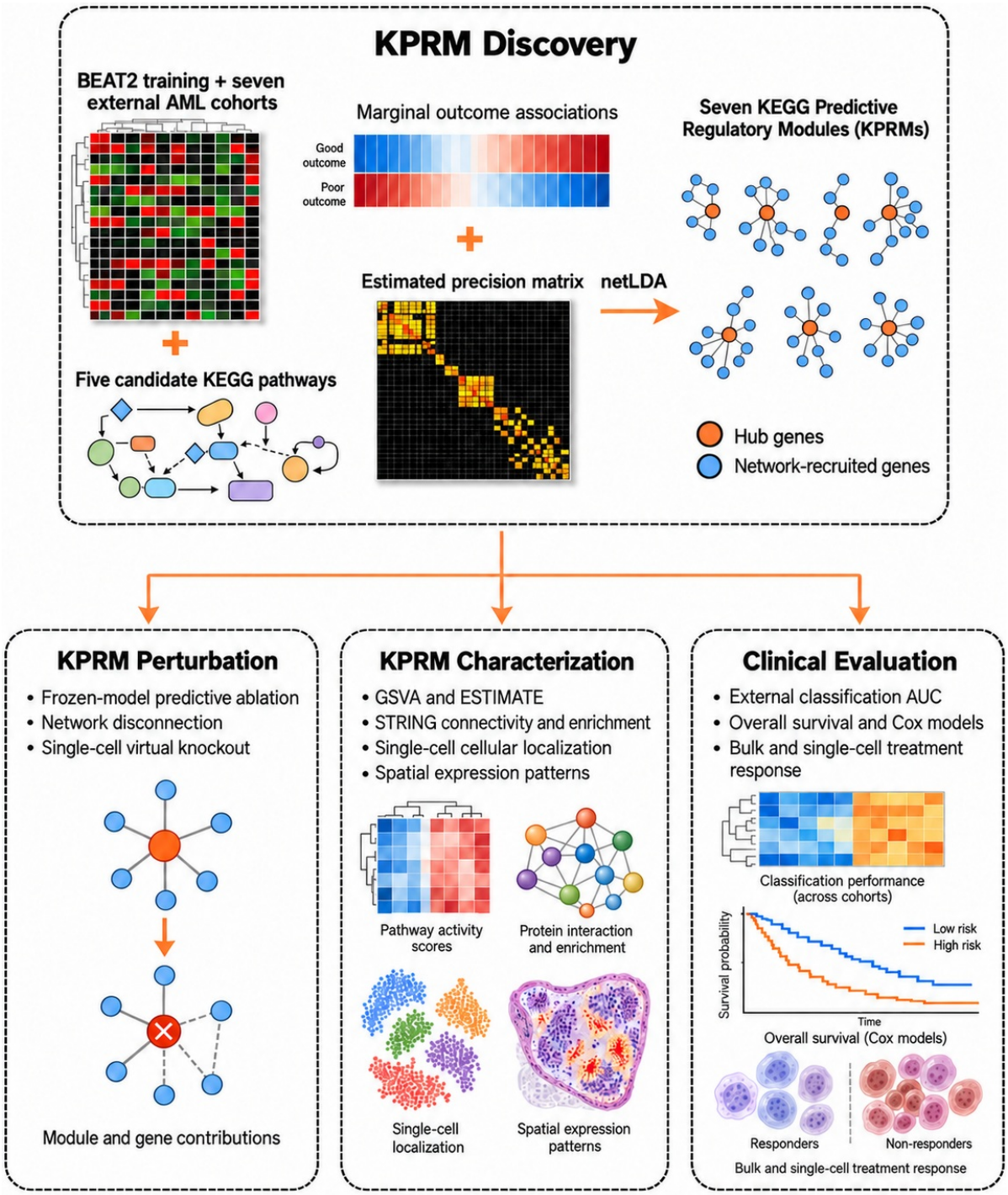


Supplementary Figure S14. Common differentiation-state adjustment in BEAT2 and the Finnish cohort. (A–B) KPRM 1 coefficients and 95% confidence intervals after Monocyte-like or six-state adjustment, using common published signatures (blue) or signatures excluding netLDA genes (orange). BEAT2 includes 228 patients; Finnish DSS models include 69. Axes retain assay-specific units: lower viability AUC and higher DSS indicate greater sensitivity. (C) Finnish response odds ratios and 95% confidence intervals after Monocyte-like adjustment in 68 patients (53 responders and 15 refractory). BEAT2 models include age, sex and specimen source; Finnish models include RNA batch, age category and disease stage. State loadings were derived in BEAT2 and frozen for Finnish projection after cohort-specific gene standardization. Gene coefficients remained frozen. Panels A–B use HC3 intervals; panel C uses mean bias-reduced logistic Wald intervals. Numerical P-values and Finnish three-score BH q-values are provided in Supplementary Table S3 and accompanying outputs.

Supplementary Figure S15 KPRM composition across patient-grouped refits



Supplementary Figure S15. Overlap between each original KPRM and its best-matching connected component in five patient-grouped outer-training refits. Orange points show fold-specific Jaccard similarities; coincident values overlap. Boxes show medians and interquartile ranges, with standard boxplot whiskers. Jaccard similarity penalizes both loss of reference genes and inclusion of additional genes. Component matches were determined independently for each reference module, allowing merges and splits. All models used τ=20, ν=200, α=0.50, d=3 and nb=10, with hub ranking repeated within the fitting partition. The five training sets overlap; the display summarizes composition sensitivity rather than independent sampling uncertainty.



Supplementary Figure S16. Overview of the netLDA framework for identifying and characterizing KEGG predictive regulatory modules in acute myeloid leukemia. Gene-expression profiles for 660 candidate genes from five KEGG signaling pathways—PI3K–Akt, JAK–STAT, MAPK, NF-κB, and Wnt—were analyzed in BEAT2 to classify one-year overall survival, incorporating standardized age. netLDA combined marginal outcome contrasts with local gene-network information to select 200 genes, comprising 20 retained hub genes and 180 network-recruited genes. Connected components of the selected network defined seven multi-gene KPRMs, with seven additional genes retained as singletons. The modules were evaluated through three approaches: predictive perturbation, using frozen-model gene ablation, network disconnection, and single-cell virtual knockout; functional and cellular characterization, using GSVA, ESTIMATE, STRING protein-association networks, single-cell transcriptomics, cell–cell communication, and spatial transcriptomics; and clinical evaluation, examining classification performance, survival stratification, and treatment-response associations. The BEAT2-trained classifier was evaluated in seven external cohorts, while cohort-specific Cox refits assessed the prognostic information carried by the selected gene sets. AML, acute myeloid leukemia; KEGG, Kyoto Encyclopedia of Genes and Genomes; KPRM, KEGG predictive regulatory module; netLDA, network-based linear discriminant analysis.

Spatial characterization of AML KPRMs

A Depth-adjusted monocytic associations

	AML11	AML12	BM1	BM2
KPRM 1	0.04	0.05	0.19**	0.21**
KPRM 2	0.01	0.04	-0.01	0.07*
KPRM 3	0.01	0.02	0.05	0.15**
KPRM 4	-0.08	-0.01	-0.05	0.10**
KPRM 5	-0.04	-0.01	0.00	-0.22**
KPRM 6	-0.01	-0.08	-0.03	-0.02
KPRM 7	-0.06	-0.06	NA	0.02

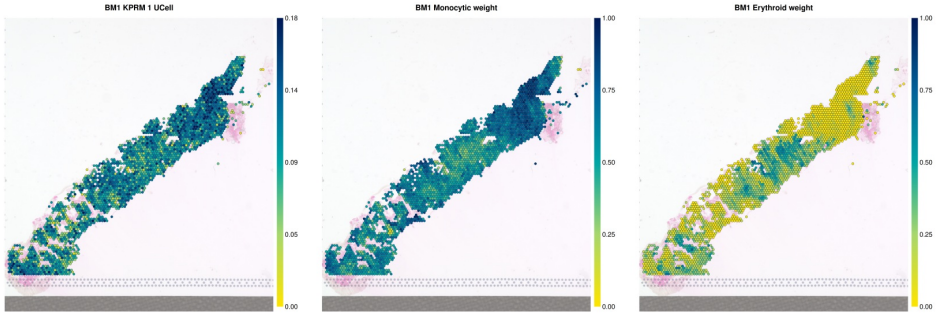
−0.3 0.0 0.3

B Thinned-score monocytic associations

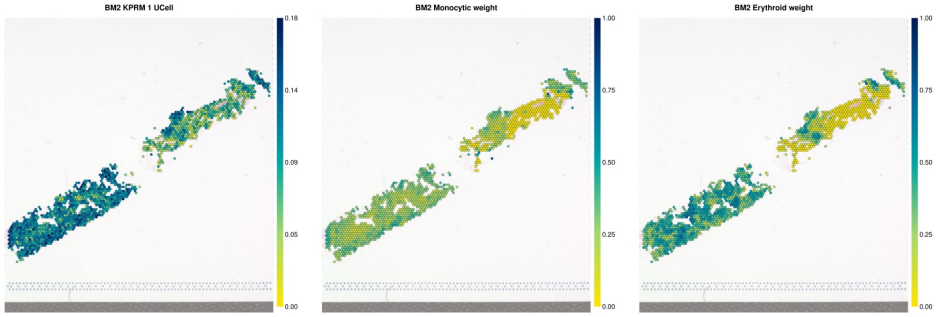
	AML11	AML12	BM1	BM2
KPRM 1	0.02	0.02	0.13**	0.11**
KPRM 2	-0.04	0.03	0.02	0.11**
KPRM 3	-0.05	0.10	0.07**	0.11**
KPRM 4	-0.07	0.03	-0.01	0.04
KPRM 5	0.02	0.01	0.01	-0.21*
KPRM 6	-0.02	-0.05	-0.00	-0.03
KPRM 7	-0.04	-0.06	NA	0.02

−0.3 0.0 0.3

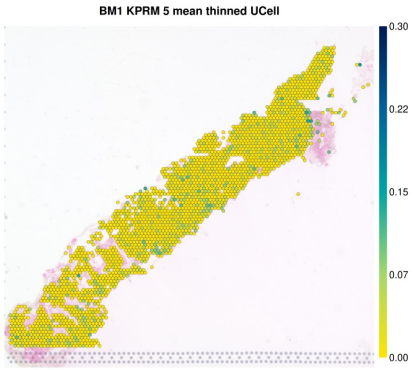
C BM1 KPRM 1 and cellular compartment weights



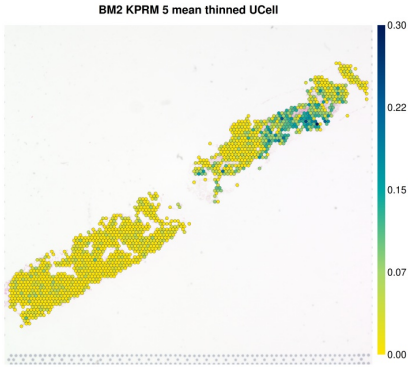
D BM2 KPRM 1 and cellular compartment weights



E BM1 KPRM 5 after count thinning



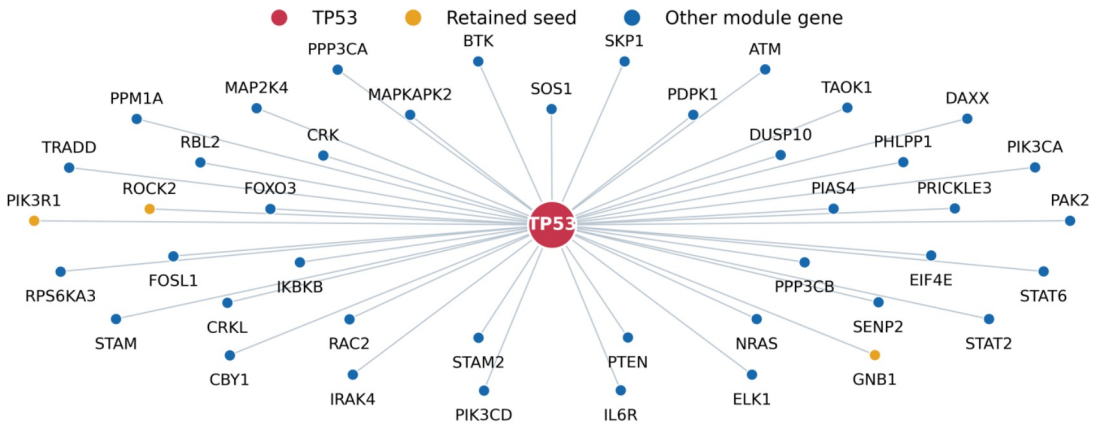
F BM2 KPRM 5 after count thinning



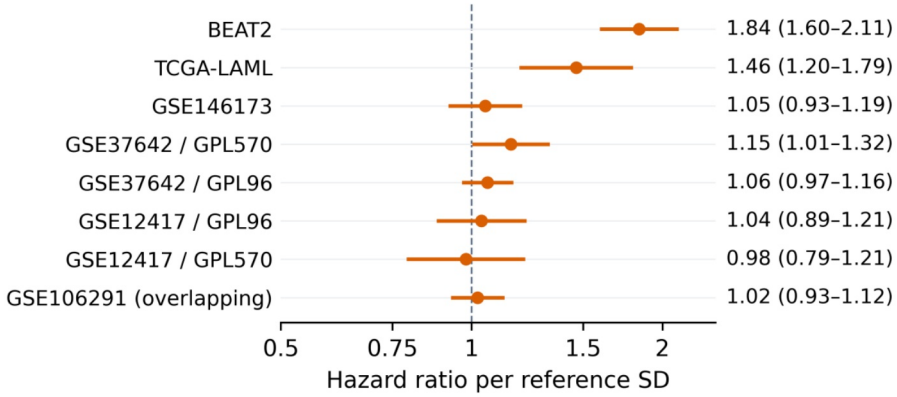
Supplementary Figure S17. Spatial characterization of KEGG predictive regulatory modules. (A–B) Associations of all seven KPRMs with monocytic RCTD weights across four AML marrow specimens, using depth-adjusted score ranks (A) or mean thinned UCell scores (B). Values are component-centered rank correlations; stars denote BH-adjusted results from the approximate graph-spectral null: * q < 0.05, ** q < 0.01, *** q < 0.001. NA indicates an untestable signature. (C–D) Original KPRM 1 UCell scores and monocytic and erythroid weights overlaid on BM1 and BM2 histology. (E–F) KPRM 5 mean UCell scores after three count-thinning replicates. RCTD maps use the supported reference excluding malignant-pDC cells. Each map has its own labeled scale. KPRM 1 monocytic associations persisted after depth adjustment and thinning in BM1/BM2; KPRM 5 retained thinned-score spatial clustering in both specimens. Focused maps were selected after reviewing the all-module results.

Clinical associations of the TP53 associated module

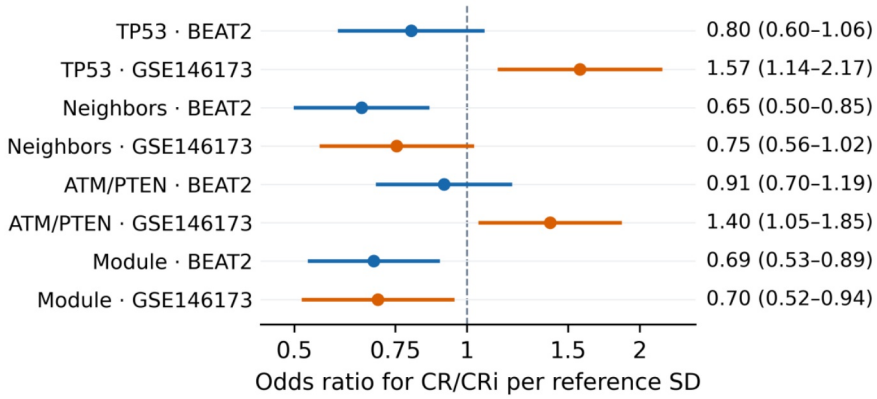
A TP53 centered view of KPRM 003



B Module score and overall survival



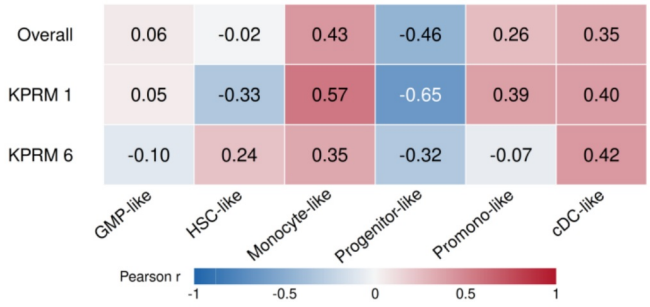
C Scores and induction response



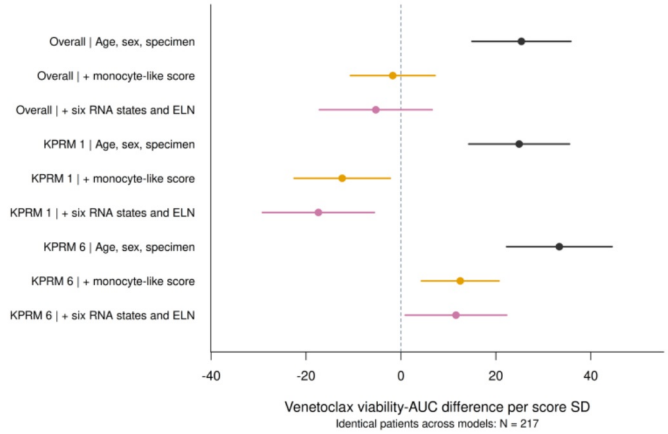
Supplementary Figure S18. Clinical associations of the TP53-associated predictive module. (A) TP53 and its 43 neighbors in the frozen netLDA inverse-covariance network. Only TP53-incident edges are shown; edges among neighbors are omitted. Gold nodes identify retained hub genes. (B) Overall-survival hazard ratios for the 44-gene module score across eight cohorts. BEAT2 models adjust for age and sex; external models adjust for age. (C) Induction-response odds ratios in BEAT2 day-zero samples (n = 298; 205 CR/CRi, 93 refractory) and GSE146173 (n = 214; 156 CR/CRi, 58 refractory), adjusted for age and sex. Each point represents a separate score model; horizontal bars indicate 95% Wald confidence intervals. Score units use the frozen BEAT2 diagnosis-reference SD. GSE106291 overlaps GSE146173 and is displayed separately.

Differentiation dependent venetoclax associations

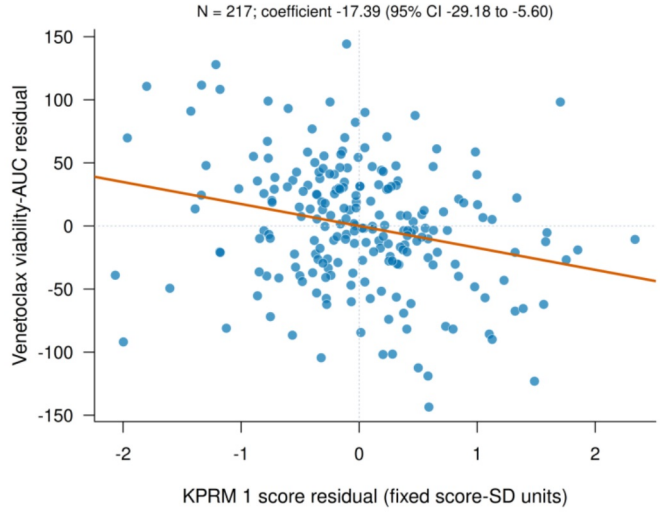
A Gene scores and differentiation states



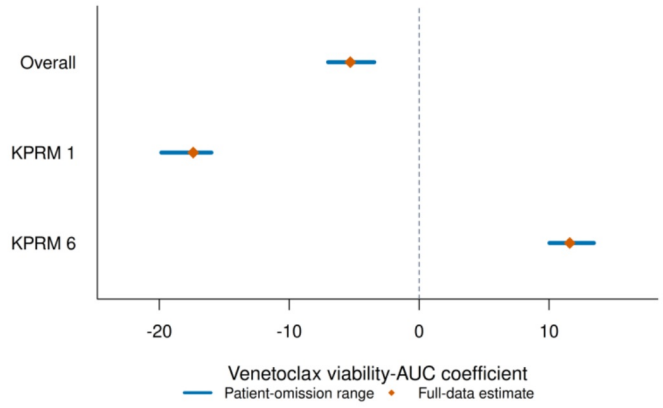
B Matched adjustment comparisons



C Adjusted KPRM 1 association



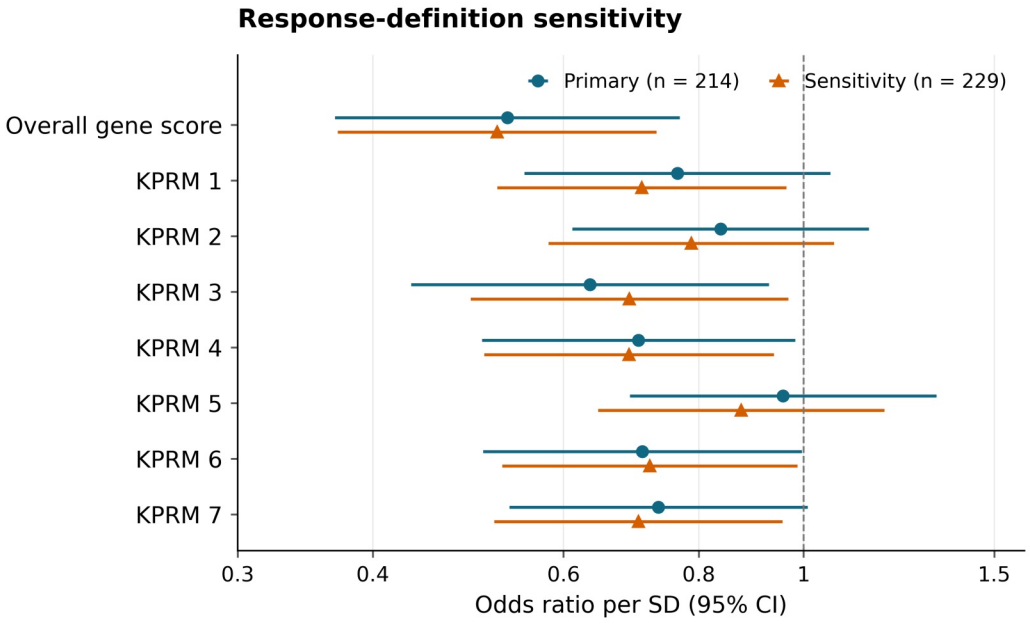
D Patient omission stability



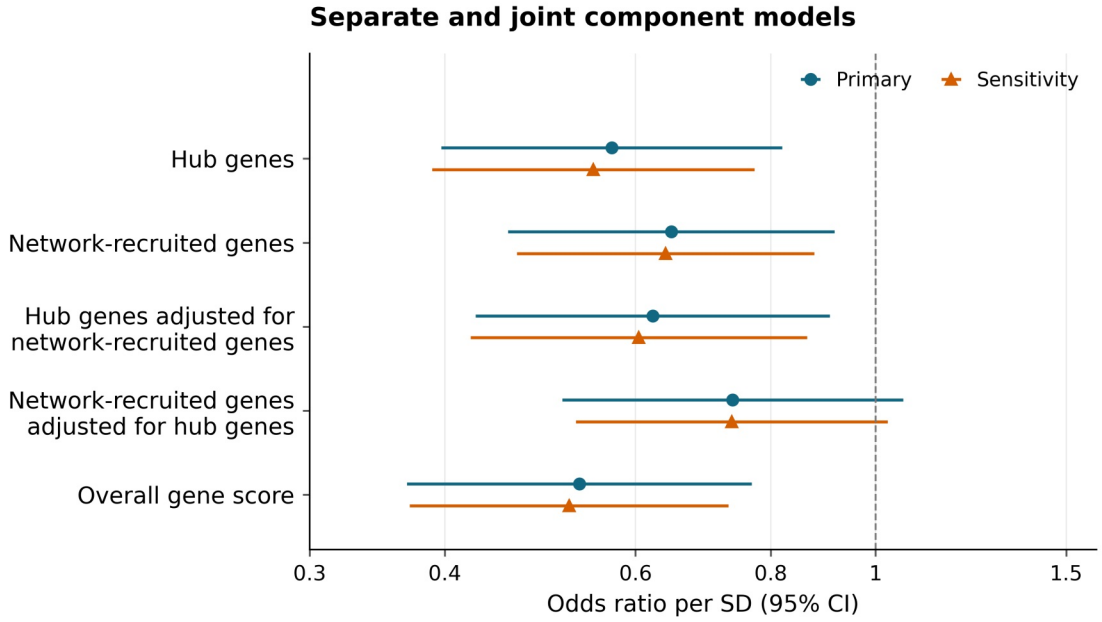
Supplementary Figure S19. Differentiation-dependent associations between netLDA gene scores and ex vivo venetoclax sensitivity in BeatAML. (A) Pearson correlations of overall, KPRM 1, and KPRM 6 frozen gene scores with six published RNA-derived differentiation scores among available venetoclax specimens. (B) Gene-score coefficients and HC3 95% confidence intervals from models fitted to the same 217 complete patients. Basic models include age, sex, and specimen source; additional adjustment includes the monocyte-like score or all six differentiation scores and ELN risk. (C) Partial-regression plot of KPRM 1 and venetoclax viability AUC after adjustment for age, sex, specimen source, six differentiation scores, and ELN. Points represent patients; the line represents the adjusted coefficient. (D) Leave-one-patient-out coefficient ranges for the six-state-plus-ELN models; diamonds mark full-data estimates. Omission ranges are stability summaries, not confidence intervals. Higher viability AUC indicates lower drug sensitivity. Gene-score effects are expressed per fixed SD from the primary diagnosis-specimen cohort. Scores contain signed frozen gene contributions without an age contribution. Numerical P-values and BH-adjusted results are supplied in the accompanying tables.

Supplementary Figure S20 Bulk response endpoint and component sensitivity

A



B



Supplementary Figure S20. Response-definition sensitivity and component associations in GSE146173. (A) Age- and sex-adjusted associations under the primary response definition (156 CR/CRi and 58 refractory; n = 214) and the sensitivity definition (156 CR/CRi and 73 nonresponders, including death with residual disease; n = 229). Each overall or KPRM score was fitted separately. (B) Hub-gene and network-recruited component associations in separate and joint age- and sex-adjusted models under both response definitions. Joint models adjust each component for the other; the hub and network-recruited components contain 20 and 180 genes, respectively. Points and bars show odds ratios per SD and 95% confidence intervals. Primary-endpoint reference scaling was retained for sensitivity comparisons. Full P-values and BH-adjusted values are provided in the accompanying statistical tables. CR/CRi, complete remission or complete remission with incomplete hematologic recovery. BEAT2 inclusion-time sensitivity is shown in Supplementary Figure S1.