Supplementary Information

A therapy-aware multimodal dynamical scaffold reveals residual persistence and relapse-associated escape in pediatric leukemia

Seung-Hwan Kim[1,2*]

[1]Department of Biology, Fisher College, Boston, MA, USA
[2]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
[*]To whom correspondence should be addressed: `seung-hwan.kim@fisher.edu`

This document contains Supplementary Notes 1–4 and Supplementary Figures S1–S7 supporting the manuscript *A therapy-aware multimodal dynamical scaffold reveals residual persistence and relapse-associated escape in pediatric leukemia*. Supplementary Fig. S7 provides a contextual cross-lineage transfer analysis using GSE227122. This analysis was performed as a supplementary robustness and transferability check only and was not used to train the scaffold or to support the primary discovery, external AML calibration, or serial bulk validation conclusions. Supplementary Data 1–6 are provided separately as Excel workbooks and are indexed in a Supplementary Data Guide. Together, the supplementary materials document raw-data processing, diagnosis-state construction, reduced equilibrium diagnostics, longitudinal displacement and branch-transition robustness, heavy-tail model comparison, external AML calibration, cross-lineage transfer, and conservative serial bulk validation.

Supplementary Note 1. Raw-data download, extraction, parsing, and per-sample h5ad generation

This note summarizes the preprocessing workflow used to convert raw downloaded archives into analysis-ready per-sample `h5ad` files and cohort-level sample summary tables. Raw GEO archives were downloaded and extracted into sample-level directories. Per-sample matrix, gene, and barcode files were inspected, decompressed where necessary, and read into `AnnData` objects. Cell barcodes were assigned to `obs`, gene identifiers were assigned to `var`, and sample-level metadata were added to each object. Where applicable, optional quality-control summaries, including numbers of detected genes, total counts, and mitochondrial fraction, were computed and retained for downstream filtering and reproducibility.

The final preprocessing output consisted of one annotated `h5ad` file per sample and a machine-readable sample summary table. These files provided the starting point for diagnosis scaffold construction, longitudinal projection, external transfer, and conservative bulk-validation analyses. The full raw-data-to-`h5ad` workflow is shown in **Supplementary Figure S1**.

Supplementary Note 2. Diagnosis-state variables, baseline scaffold structure, and reduced equilibrium diagnostics

Diagnosis-state variables were constructed to provide interpretable, low-dimensional summaries of malignant-state, ecological, regulatory, and immune-context structure before longitudinal displacement analyses. These variables included calibrated lineage-composition contrasts, auxiliary hematopoietic programs, and ecological principal-component coordinates used to define the diagnosis-anchored scaffold. The goal was not to replace the full single-cell representation, but to obtain compact coordinates that could be transferred across cohorts and interpreted in terms of disease-state organization.

For each diagnosis sample, Euclidean distance from the cohort centroid was computed as a summary of baseline state-space displacement. Pairwise correlations among diagnosis-state variables were then used to characterize covariance, redundancy, and contrast among the scaffold axes. Leukemia-versus-control distributions of transferable validation variables in the serial bulk validation cohort are shown in **Supplementary Figure S2**. Baseline diagnosis heterogeneity and covariance structure are summarized in **Supplementary Figure S3**.

A reduced equilibrium model was used to evaluate whether interpretable diagnosis-state axes could be approximated from coarse ecological and subgroup variables. Coefficient structures and leave-one-out prediction diagnostics are shown in **Supplementary Figure S4**. These analyses support the biological interpretability of the reduced axes while indicating that the reduced summaries are intended primarily for scaffold interpretation and transfer, rather than individual-level prediction.

Supplementary Note 3. Branch-transition, displacement, transfer, and validation robustness analyses

Longitudinal DX→REL intervals were analyzed in the frozen diagnosis-defined scaffold to quantify branch-continuous and branch-switching trajectories. For each matched interval, total displacement and block-wise malignant-state and tumor-microenvironment displacement were computed. Branch-continuous intervals remained within the same coarse disease-state basin, whereas branch-switching intervals moved between basins. Additional branch-transition and displacement diagnostics are shown in **Supplementary Figure S5**.

Upper-tail displacement analyses were performed to test whether relapse-associated movement was compatible with constrained Gaussian fluctuation or instead showed evidence of punctuated departure. Robust z-scores were computed relative to the branch-continuous baseline, and jump-candidate scores were defined from Gaussian-tail surprisal. Branch-switching intervals showed larger total and block-wise displacement and were enriched among upper-tail jump candidates. These analyses are summarized in **Supplementary Figure S6A–D**.

Formal model comparison was performed using nested Gaussian and Student-t displacement models. The Student-t family was used as a tractable heavy-tailed surrogate for jump-capable, Lévy-like increment behavior. Candidate models compared pooled versus branch-aware structure and Gaussian versus heavy-tailed residual behavior. The branch-aware heavy-tail model provided the strongest support by corrected Akaike information criterion, and case-wise model improvement was concentrated among extreme-displacement cases. These analyses are shown in **Supplementary Figure S6E–H**.

Cross-lineage transfer was evaluated by projecting GSE227122 samples into the frozen AML-derived ecological coordinate system. The strict transfer analysis documents timepoint organization, paired longitudinal movement, and the per-sample normal-cell support used for transfer. These results are summarized in **Supplementary Figure S7**.

Supplementary Note 4. Therapy-aware ecological OU–Lévy–Branching extension

This note gives a mathematical extension of the ecological OU–Branching model used in the main text. The extension explicitly includes treatment or clinical phase, allowing attractor location, restoring strength, diffusion scale, jump propensity, and branch-transition behavior to vary across diagnosis, response-associated, residual, and relapse states.

Let X_t denote the reduced leukemia state vector, e_t the ecological context, b_t the latent branch state, and τ_t the treatment or clinical phase. The phase variable τ_t may represent diagnosis, on-therapy state, EOI/REM, or relapse. We write the local attractor as $\mu_{(b_t,e_t,\tau_t)}$, the restoring-strength matrix as $\Theta_{(b_t,e_t,\tau_t)}$, the diffusion matrix as $\Sigma_{(b_t,e_t,\tau_t)}$, and the jump process as J_t. A therapy-aware OU–Lévy–Branching process can then be written as

$$\mathrm{d}X_t = \Theta_{(b_t,e_t,\tau_t)}\big(\mu_{(b_t,e_t,\tau_t)} - X_{t^-}\big)\,\mathrm{d}t + \Sigma^{1/2}_{(b_t,e_t,\tau_t)}\,\mathrm{d}W_t + \mathrm{d}J_t. \tag{1}$$

Here X_{t^-} denotes the left limit of the càdlàg process X_t. Away from jump times, $X_{t^-} = X_t$. The first term describes OU-like stabilization toward a phase-, branch-, and ecology-specific attractor. The second term represents gradual stochastic fluctuation within the active regime. The third term represents rare punctuated departures, allowing the transition distribution to exhibit heavy tails.

On an interval $[s,t]$ during which the branch, ecological context, and treatment phase remain fixed, say $b_r = b$, $e_r = e$, and $\tau_r = \tau$ for $r \in [s,t]$, Eq. (1) reduces to

$$\mathrm{d}X_t = \Theta_{(b,e,\tau)}\big(\mu_{(b,e,\tau)} - X_{t^-}\big)\,\mathrm{d}t + \Sigma^{1/2}_{(b,e,\tau)}\,\mathrm{d}W_t + \mathrm{d}J_t. \tag{2}$$

Let $\Delta = t - s$. Conditional on the fixed regime (b,e,τ) over $[s,t]$, the mild solution is

$$X_t = \mu_{(b,e,\tau)} + \exp\big(-\Theta_{(b,e,\tau)}\Delta\big)\big(X_s - \mu_{(b,e,\tau)}\big) + \int_s^t \exp\big(-\Theta_{(b,e,\tau)}(t-r)\big)\Sigma^{1/2}_{(b,e,\tau)}\,\mathrm{d}W_r + \int_s^t \exp\big(-\Theta_{(b,e,\tau)}(t-r)\big)\,\mathrm{d}J_r. \tag{3}$$

This representation separates four interpretable mechanisms. First, $\mu_{(b,e,\tau)}$ defines the local attractor selected by branch, ecology, and treatment phase. Second, $\Theta_{(b,e,\tau)}$ controls the strength of local restoration

toward that attractor. Third, $\Sigma_{(b,e,\tau)}$ controls within-regime stochastic spread. Fourth, J_t captures rare discontinuous departures from local equilibrium structure.

Connection to empirical dynamic summaries

The main-text dynamic summaries are empirical reductions of the model components above. Effective restoring strength, denoted θ_{eff}, summarizes the degree to which a sample remains constrained toward a local attractor. Attractor displacement, denoted μ-shift, summarizes the distance between an inferred sample-level attractor and the diagnosis-defined baseline. Effective instability or diffusion proxy, denoted σ_{eff}, summarizes local spread or reduced constraint. Jump-sensitive score summarizes the degree to which an interval resembles an upper-tail departure relative to the branch-continuous baseline.

In this interpretation, response-like constrained states have relatively low attractor displacement and stronger restoration. Residual persistent states retain disease-state structure but do not fully collapse back to the diagnosis baseline. Escape-prone relapse-like states show larger displacement, enrichment for branch switching, and elevated jump-sensitive behavior.

Compound-Poisson special case

A biologically interpretable specification is a compound-Poisson jump process,

$$J_t = \sum_{k=1}^{N_t} Y_k, \qquad N_t \sim \text{Poisson}\big(\lambda_{(b,e,\tau)} t\big), \tag{4}$$

where $\lambda_{(b,e,\tau)}$ is the phase-, branch-, and ecology-specific jump intensity and Y_k are jump sizes. Under this specification,

$$X_t = \mu_{(b,e,\tau)} + e^{-\Theta_{(b,e,\tau)}\Delta}\big(X_s - \mu_{(b,e,\tau)}\big) + \int_s^t e^{-\Theta_{(b,e,\tau)}(t-r)}\Sigma_{(b,e,\tau)}^{1/2}\,\mathrm{d}W_r + \sum_{\tau_k \in (s,t]} e^{-\Theta_{(b,e,\tau)}(t-\tau_k)} Y_k. \tag{5}$$

Each jump produces an abrupt displacement of the latent disease state, after which OU-like drift can restore the trajectory toward the active attractor. In the pediatric leukemia setting, the Brownian term represents gradual within-branch fluctuation, the jump term represents rare treatment-associated or evolutionary shocks, and the branch process represents structural redirection into alternative disease-state basins.

Branch-transition component

The branch process can be written as a discrete transition model,

$$\Pr(b_{t^+} = b' \mid b_t = b, e_t = e, \tau_t = \tau) = \Pi_{b \to b'}(e, \tau), \tag{6}$$

where $\Pi_{b \to b'}(e, \tau)$ is a branch-transition probability that may depend on ecological context and treatment phase. Branch-continuous trajectories correspond to $b' = b$, whereas branch-switching trajectories correspond to $b' \neq b$. In the empirical analysis, branch switching was used as an observable coarse-grained signature of structural redirection.

Heavy-tail surrogate used for model comparison

Direct inference of a continuous-time Lévy process is not identifiable from the available sparse clinical sampling. Therefore, the formal model-comparison analyses used Student-t displacement models as tractable

heavy-tailed surrogates for Lévy-like increments. Gaussian models represent constrained fluctuation around a local displacement distribution, whereas Student-t models allow heavier tails. Branch-aware models allow displacement distributions to differ between branch-continuous and branch-switching intervals. The model-comparison results therefore test whether both branch structure and heavy-tail behavior are needed to describe the observed DX→REL displacement distribution, rather than proving a unique mechanistic jump process.

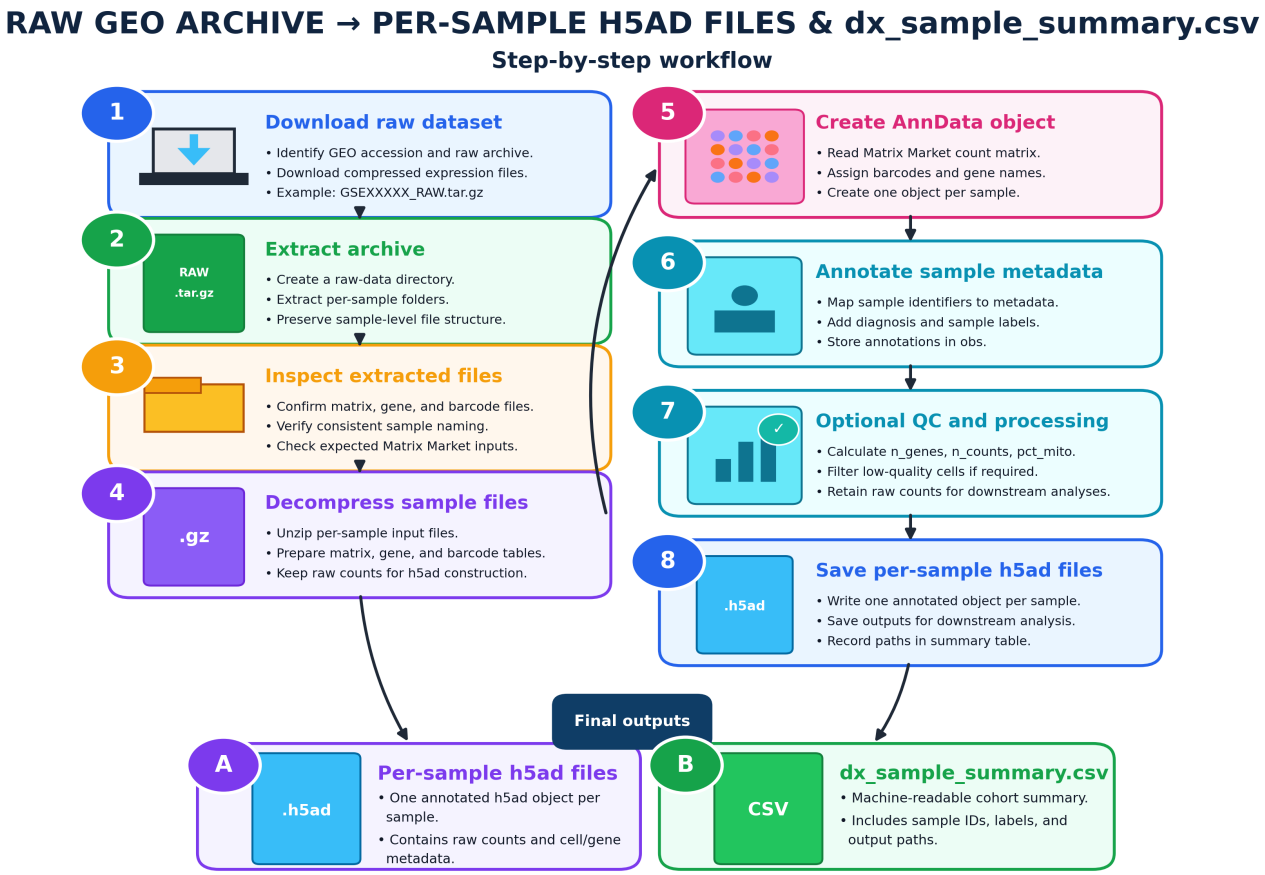
Parameter block for future Bayesian inference

For future Bayesian inference, the full therapy-aware within-regime parameter block may be written as

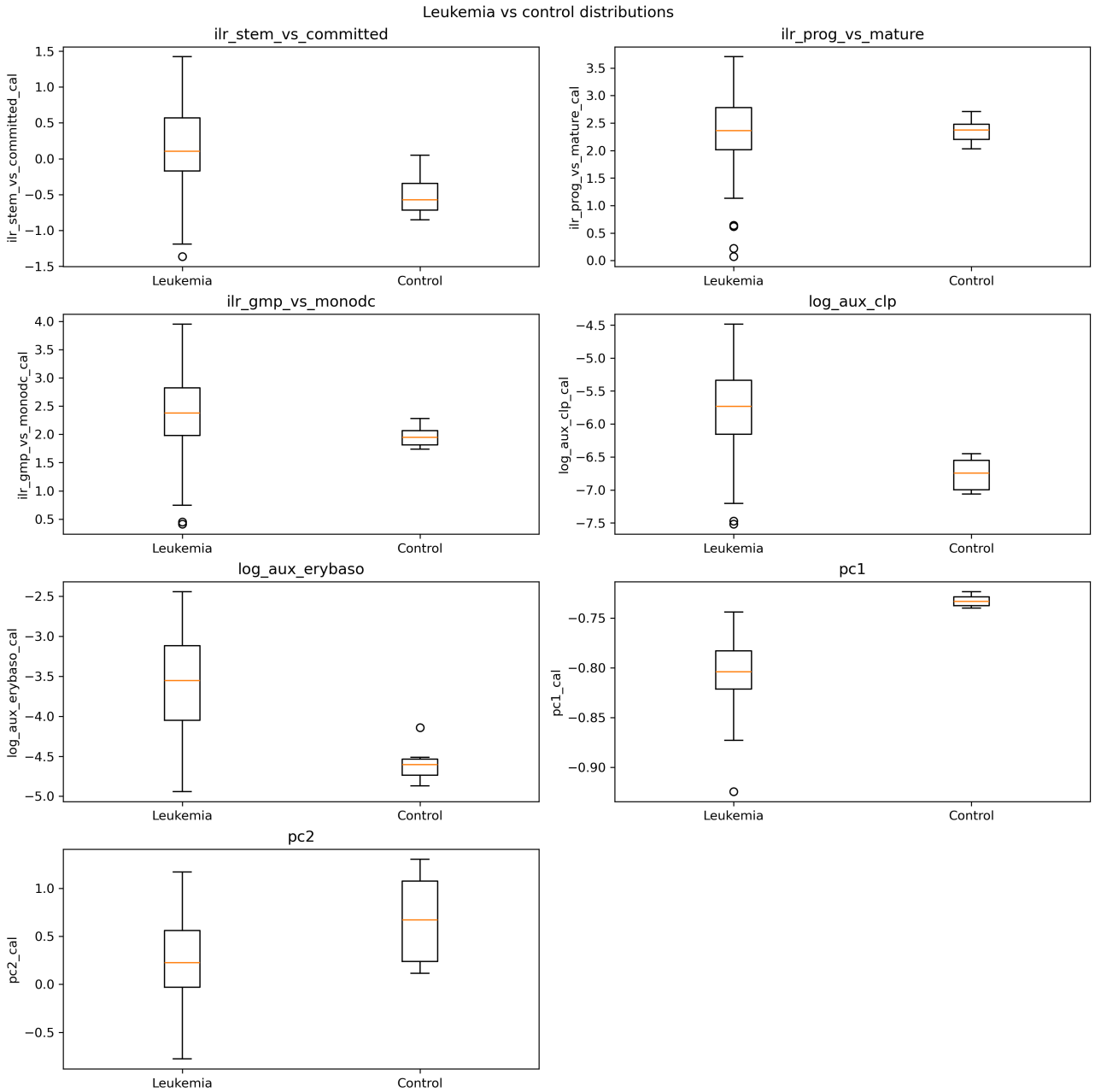
$$\Psi_{(b,e,\tau)} = \left\{\mu_{(b,e,\tau)},\, \Theta_{(b,e,\tau)},\, \Sigma_{(b,e,\tau)},\, \phi_{(b,e,\tau)},\, \Pi_{(b,e,\tau)}\right\}, \tag{7}$$

where $\phi_{(b,e,\tau)}$ collects jump parameters, such as jump intensity and jump-size parameters, and $\Pi_{(b,e,\tau)}$ collects branch-transition probabilities. This parameterization preserves the attractor-based biological interpretation of the main model while allowing therapy phase to modulate stabilization, persistence, and relapse-associated escape.

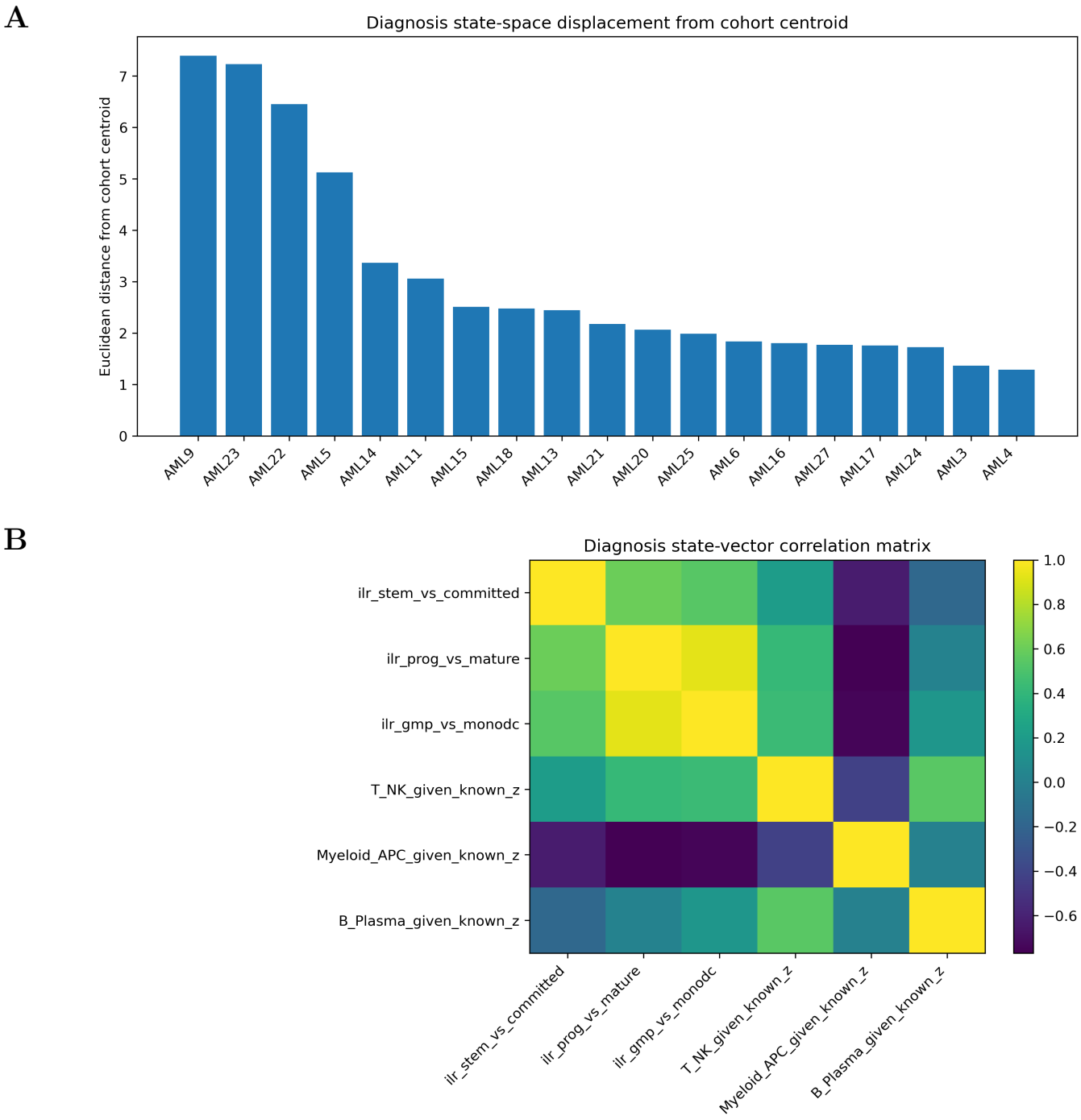
Supplementary Figures



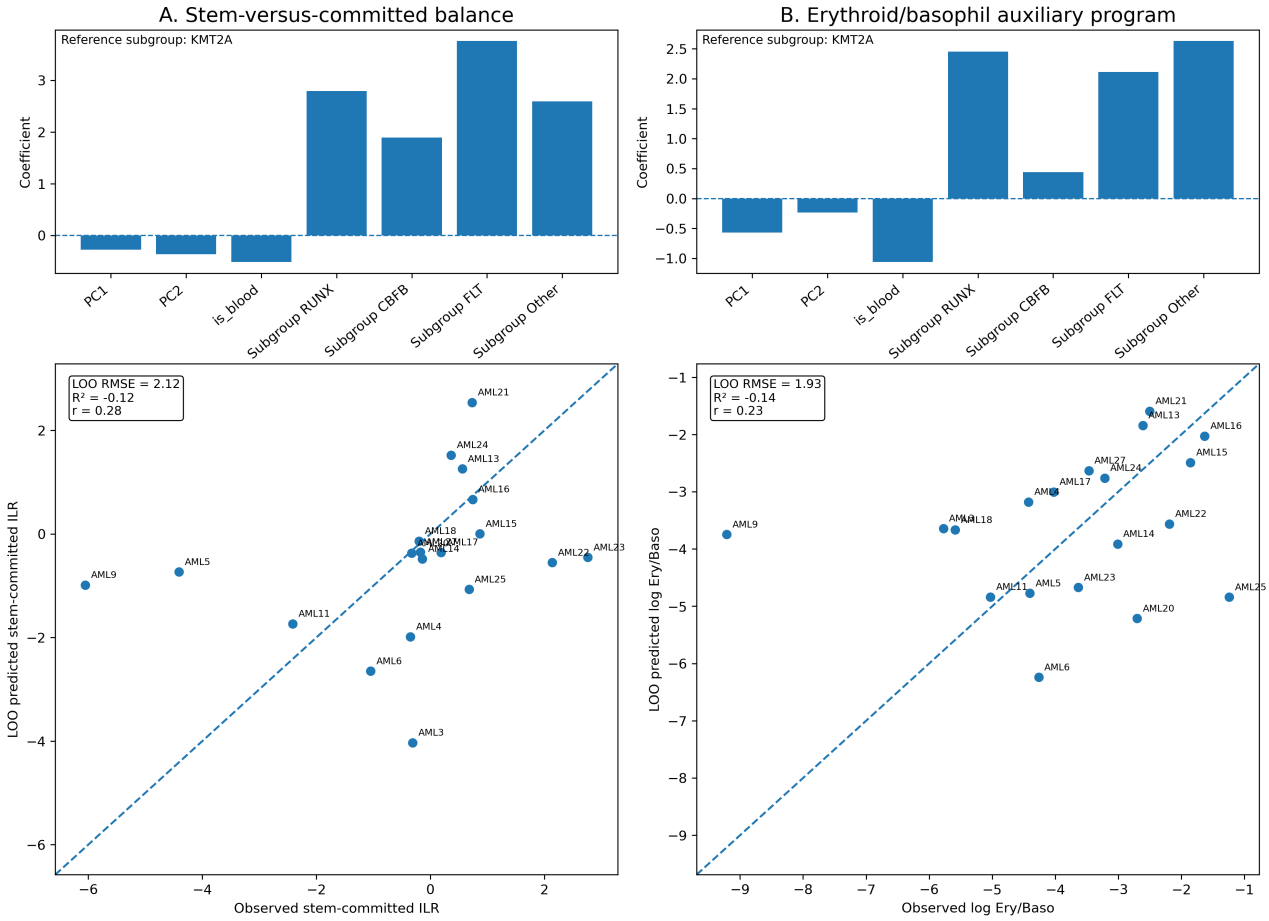
Supplementary Figure S1. Raw-data processing and per-sample object construction workflow. Step-by-step workflow used to convert raw GEO archive files into per-sample AnnData objects and a machine-readable sample summary table. The workflow proceeds through raw archive download, archive extraction, file inspection, sample-file decompression, AnnData object creation, sample-metadata annotation, optional quality control and processing, and final export of per-sample `h5ad` files together with `dx_sample_summary.csv`.



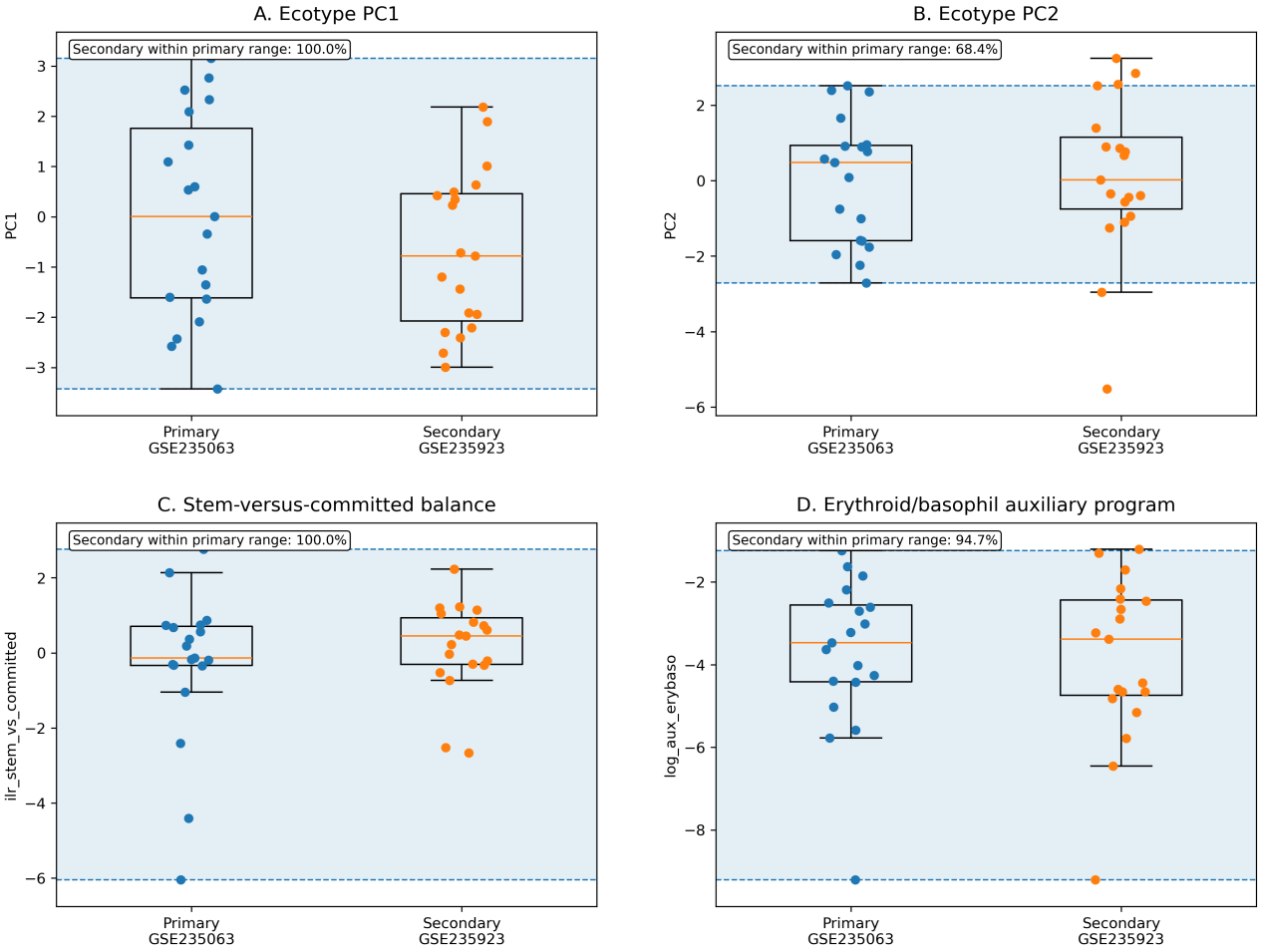
Supplementary Figure S2. Leukemia-versus-control distributions of transferable validation variables. Boxplots compare leukemia and control distributions for transferable state variables used in reduced validation analyses. Variables include immature-versus-committed balance, progenitor-versus-mature balance, GMP-versus-Mono/DC balance, auxiliary CLP and erythroid/basophil programs, and ecological principal-component axes. These distributions show that transferred or bulk-compatible variables retain disease-associated structure and provide a basis for conservative lower-resolution validation.



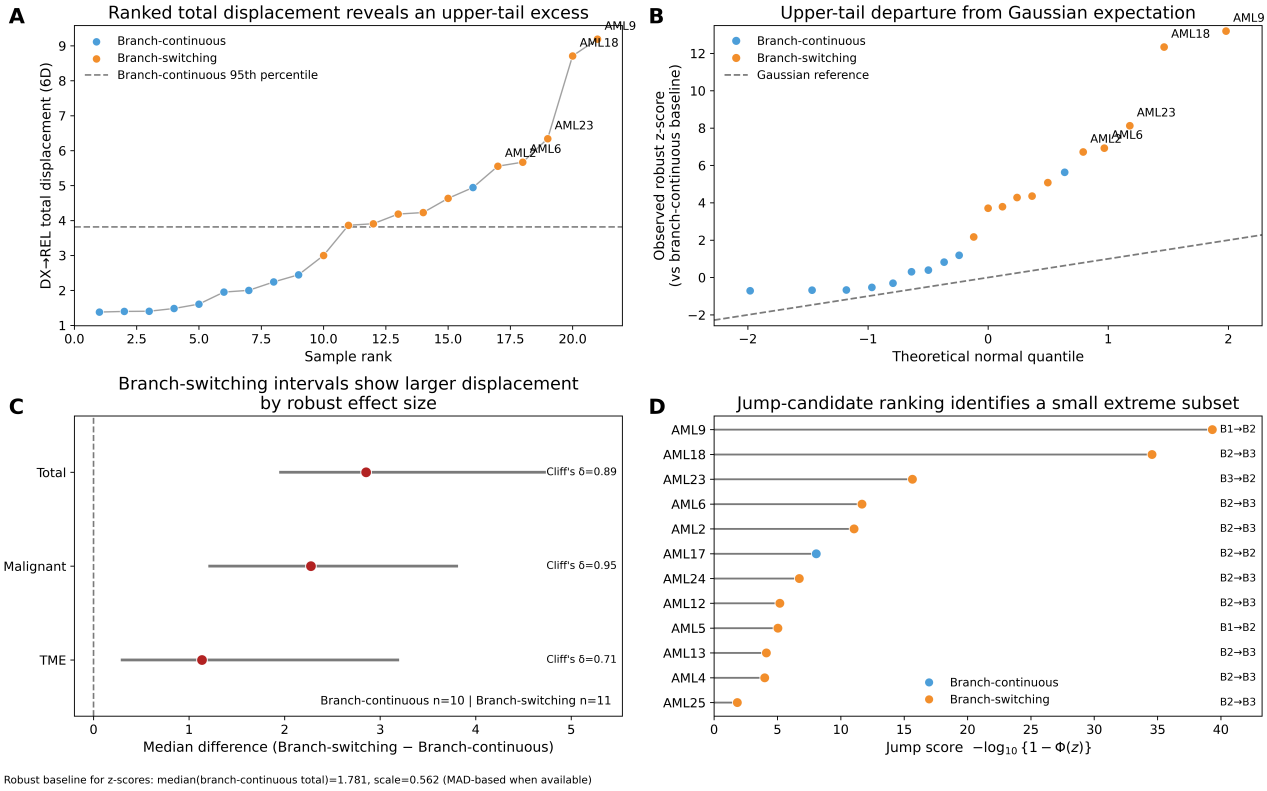
Supplementary Figure S3. Baseline diagnosis state-space heterogeneity and covariance structure. (A) Diagnosis state-space displacement from the cohort centroid, ranked across primary AML diagnosis samples. Larger values identify samples occupying more peripheral regions of the diagnosis-defined scaffold. **(B)** Diagnosis state-vector correlation matrix showing covariance among malignant-state, immune-context and ecological variables used to define the reduced diagnosis state space. Together, these analyses support the presence of structured heterogeneity within the pretreatment scaffold rather than a single homogeneous diagnosis state.



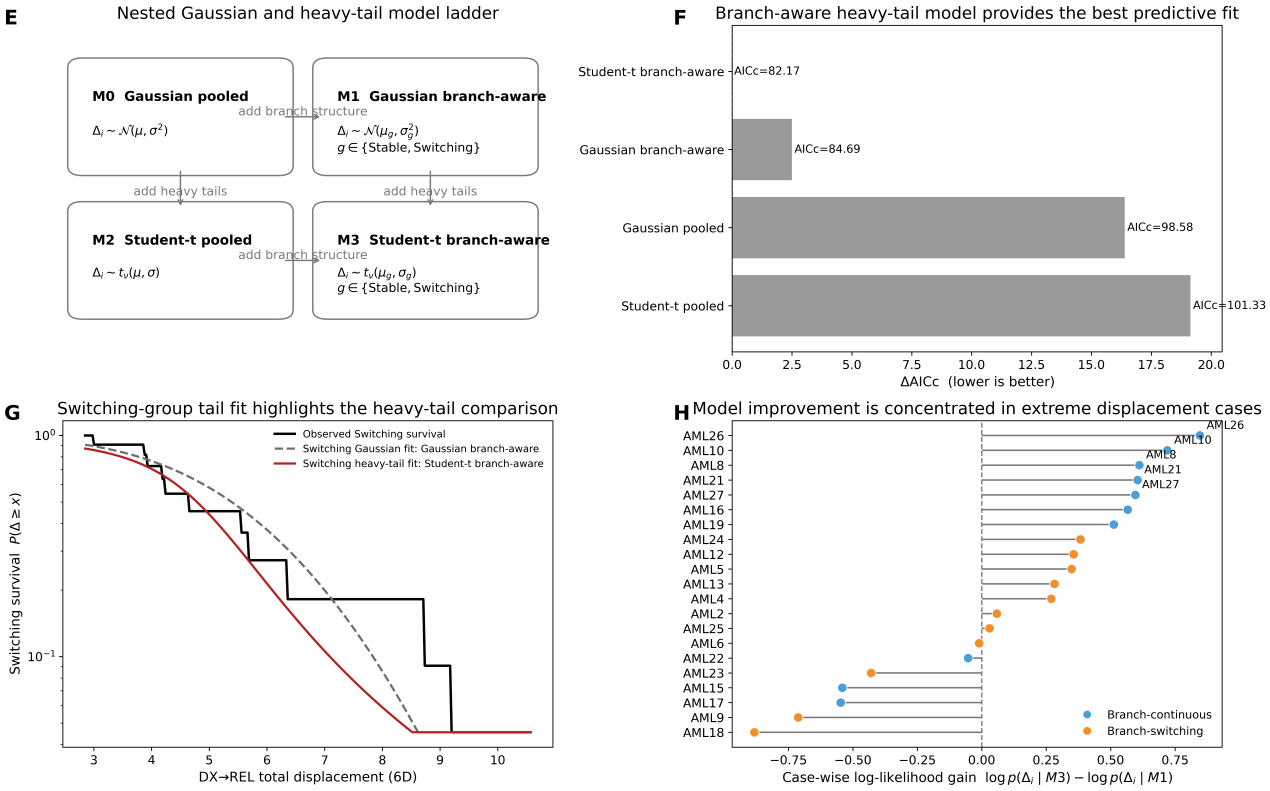
Supplementary Figure S4. Reduced equilibrium model and leave-one-out diagnostics in the primary AML cohort. (A) Coefficient structure for the stem-versus-committed balance axis relative to the KMT2A reference subgroup. **(B)** Coefficient structure for the erythroid/basophil auxiliary-program axis relative to the KMT2A reference subgroup. **(C)** Leave-one-out predicted versus observed values for the stem-versus-committed axis. **(D)** Leave-one-out predicted versus observed values for the erythroid/basophil auxiliary-program axis. These diagnostic analyses support the interpretability of the reduced equilibrium axes while indicating that the low-dimensional summaries are intended primarily for biological interpretation rather than individual-level prediction.



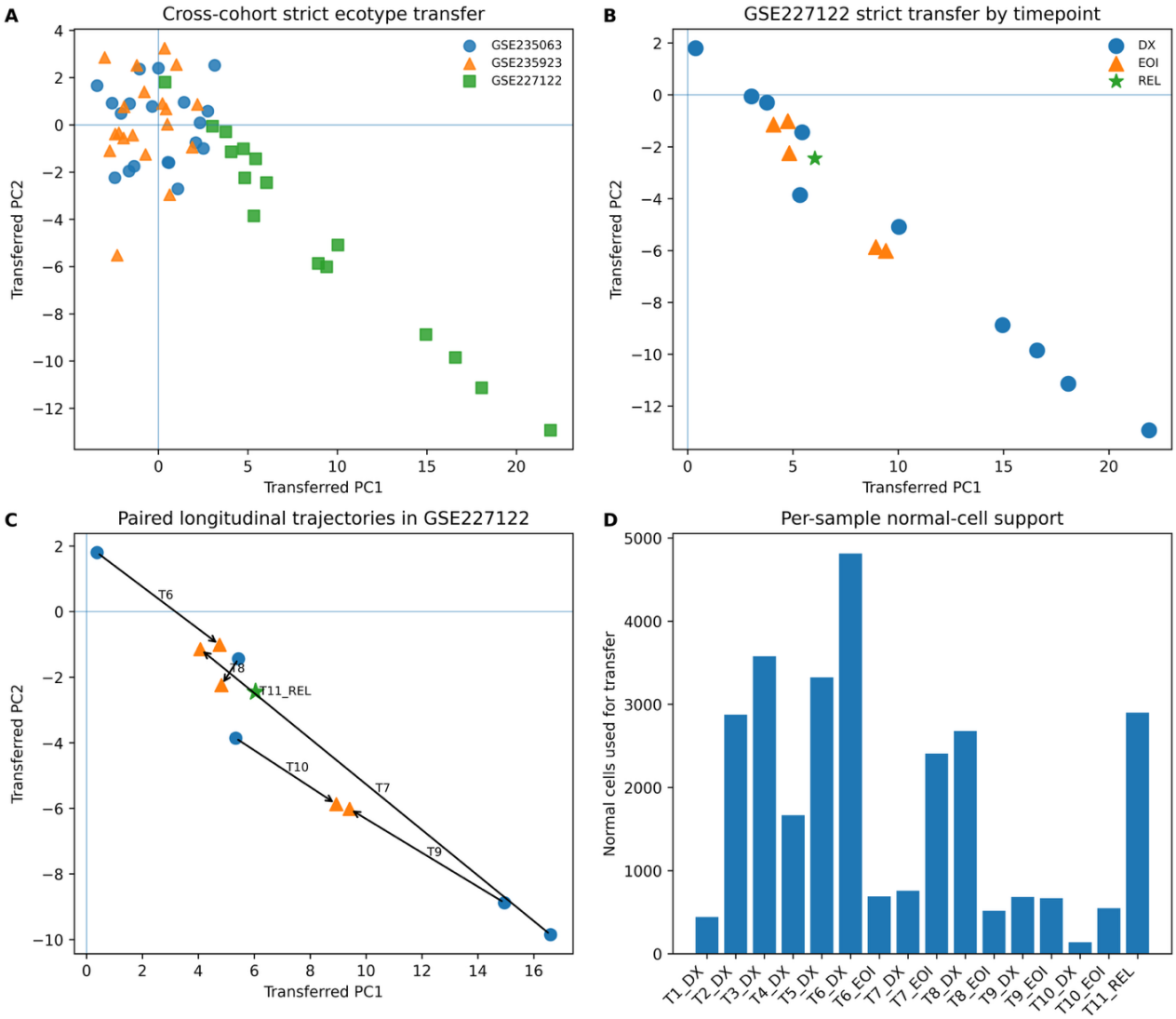
Supplementary Figure S5. Cross-cohort comparison of projected ecotype and malignant-state summaries between primary and secondary AML cohorts. Boxplots and overlaid sample points compare ecotype PC1, ecotype PC2, stem-versus-committed balance, and erythroid/basophil auxiliary program between GSE235063 and GSE235923. Ecotype PC1 and PC2 represent the first two principal-component coordinates of the projected ecotype summary space learned in GSE235063 and applied without refitting to GSE235923. Shaded bands and dashed boundaries indicate the primary-cohort reference range. The proportion of secondary samples falling within the primary range is annotated for each axis.



Supplementary Figure S6. Upper-tail displacement and heavy-tail model support for punctuated relapse-associated escape. **(A)** Ranked total DX→REL displacement across retained matched intervals. The dashed line indicates the branch-continuous 95th percentile, used as an empirical constrained-motion reference. **(B)** Upper-tail departure from Gaussian expectation after robust standardization relative to the branch-continuous baseline. **(C)** Robust effect-size summaries for total, malignant-state and tumor-microenvironment displacement, shown as median differences between branch-switching and branch-continuous intervals with bootstrap confidence intervals. **(D)** Jump-candidate ranking based on Gaussian-tail surprisal, identifying a small extreme subset enriched for branch-switching transitions.



Supplementary Figure S6 continued. Heavy-tail model support for punctuated relapse-associated escape. (E) Nested Gaussian and heavy-tail model ladder comparing pooled and branch-aware displacement models. Gaussian models represent constrained fluctuation, whereas Student-t models provide a tractable heavy-tailed surrogate for jump-capable behavior. **(F)** Corrected Akaike information criterion comparison across the four candidate models. The branch-aware heavy-tail model provides the best supported fit. **(G)** Branch-switching tail fit comparing observed survival behavior with Gaussian and heavy-tail fitted survival curves. **(H)** Case-wise log-likelihood gain of the branch-aware heavy-tail model over the branch-aware Gaussian model. Model improvement is concentrated in a subset of extreme-displacement cases, supporting a punctuated escape interpretation for the upper tail.



Supplementary Figure S7. Strict cross-lineage ecotype transfer and support metrics. **(A)** Cross-cohort strict ecotype transfer into the frozen AML-derived ecological coordinate system, including the primary AML cohort, independent AML cohort and cross-lineage GSE227122 cohort. **(B)** GSE227122 samples stratified by clinical timepoint after projection into the frozen coordinate system. **(C)** Paired longitudinal trajectories in GSE227122, showing directional movement across available timepoints. **(D)** Per-sample normal-cell support used for transfer. These analyses support the reuse of the ecological scaffold outside the primary AML discovery cohort and provide context for cross-lineage transfer interpretation.

Supplementary Figure Abbreviations

DX, diagnosis; EOI, end of induction; REM, remission; REL, relapse; AML, acute myeloid leukemia; HSC, hematopoietic stem cell-like; GMP, granulocyte–monocyte progenitor-like; Mono/DC, monocyte/dendritic cell-like; TME, tumor microenvironment; OU, Ornstein–Uhlenbeck; AICc, corrected Akaike information criterion.

Supplementary Data Guide

Supplementary Data 1. Cohort metadata and preprocessing summaries. This workbook contains sample identifiers, cohort labels, clinical timepoints, preprocessing outputs, per-sample `h5ad` paths, and sample-level inclusion summaries.

Supplementary Data 2. Diagnosis scaffold variables and baseline structure. This workbook contains GSE235063 diagnosis-state variables, ecotype summaries, branch assignments, scaffold coordinates, centroid distances, state-vector covariance summaries, and reduced equilibrium diagnostics.

Supplementary Data 3. Longitudinal displacement and branch-transition summaries. This workbook contains matched DX→REL intervals, diagnosis and relapse branch identities, branch-continuous and branch-switching annotations, displacement metrics, relapse-escape scores, branch ecological summaries, scaffold program summaries, and branch-conditioned escape-propensity outputs.

Supplementary Data 4. Therapy-aware dynamic parameters and state tiers. This workbook contains effective restoring strength, attractor displacement, effective instability, jump-sensitive scores, branch behavior, dynamic clinical map variables, representative dynamic-state scorecard rows, and state-tier assignments.

Supplementary Data 5. External AML calibration and cross-lineage transfer. This workbook contains GSE235923 external AML projection outputs, dynamic-parameter summaries, full-triad annotations, primary-versus-secondary AML calibration ranges supporting Supplementary Figure S5, and GSE227122 strict-transfer coordinates and support metrics.

Supplementary Data 6. Serial bulk validation and model-comparison outputs. This workbook contains GSE163634 bulk-validation summaries, transferred state variables, serial response comparisons, Gaussian and heavy-tail model-comparison tables, parameter summaries, tail-fit grids, and case-wise log-likelihood gains.