Bridging Bulk Genomics and Physics-Constrained Tumor Simulation via Knowledge-Grounded Clonal Coarse-Graining

Per Magnus Swedenborg[1,*]

[1] DNAI Biotech * Correspondence: per.swedenborg@dnai.bio | ORCID: 0009-0008-2750-9735

Abstract

Background: Mechanistic models of tumor evolution require fixed-dimensional state variables to simulate competitive dynamics. However, mapping bulk sequencing data to these simulators is an ill-posed problem. Standard clonal deconvolution yields variable numbers of nested clones, systematically loses clinically critical minor resistant populations during dimensionality reduction, and produces under-identified drug sensitivity parameters. **Methods:** We present a reproducible computational interface layer bridging clonal deconvolution with a physics-constrained Lotka-Volterra Neural ODE. We introduce a "Resistance Sentinel" strategy that maps variable clones to a fixed macro-state ($K = 4$), utilizing probabilistic guardrails to preserve minor resistant subclones. Nested Cancer Cell Fractions (CCFs) are transformed into disjoint population fractions via a formal nesting Directed Acyclic Graph (DAG) subtraction algorithm, followed by constrained quadratic programming on the simplex using Condat's $\mathcal{O}(K \log K)$ algorithm. Finally, clone-specific drug sensitivities (β) are anchored using curated associations, providing a regularization prior for under-identified parameters via structured abstention backed by a 328d multi-omics latent space. **Results:** Evaluated across a cohort of 9415 TCGA patients, the interface layer passed a software verification suite of 72/72 deterministic invariant tests for mass conservation and non-negativity. As a verification of the rule, the Resistance Sentinel successfully retained 23/23 curated canonical resistance mutations present in minor subclones, whereas a naive Top-K approach retained 0/23 (McNemar's exact test, $p = 2.4 \times 10^{-7}$). Addressing structural identifiability in synthetic treated trajectories, we observed an empirical β/ρ ratio of 0.0017 (IQR [0.0008, 0.0031]), confirming severe practical non-identifiability without anchoring. In system-level context, the full multimodal architecture achieved a global Concordance Index of 0.704 [0.691, 0.717] and a trajectory emulator R^2 of 0.997. **Conclusion:** Framing clonal deconvolution as a functional macro-state mapping provides a mathematically coherent initialization for mechanistic simulation. This knowledge-grounded interface resolves critical mass-conservation failures and provides a stable, reproducible precondition for future longitudinal treatment forecasting.

1. Introduction

The development of mechanistic digital twins in oncology aims to enable the simulation of tumor evolution under selective therapeutic pressure. While deep learning models have achieved high prognostic accuracy for static risk stratification, predicting the temporal dynamics of acquired resistance requires eco-evolutionary differential equations that explicitly model competitive dynamics between tumor subclones.

However, a critical translational gap exists between the outputs of standard genomic bioinformatics pipelines and the inputs required by mechanistic simulators. Bulk DNA sequencing provides variant allele frequencies (VAFs) representing an admixture of normal tissue, dominant tumor clones, and

minor subclones. While algorithms can deconvolute these VAFs into latent clonal clusters, they output a variable number of clones with nested Cancer Cell Fractions (CCFs). In contrast, stable numerical simulation of multi-compartment Lotka-Volterra systems requires a fixed-dimensional state vector representing disjoint, mass-conserving population fractions.

Bridging this gap naively introduces three severe failure modes. First, standard dimensionality reduction (e.g., selecting the top K clones by prevalence) systematically discards low-prevalence subclones. If a minor subclone harbors a canonical resistance mutation, discarding it eliminates the simulator's ability to model relapse. Second, directly initializing differential equations with nested CCFs violates the law of mass conservation, leading to unphysical population dynamics. Third, attempting to learn clone-specific drug sensitivity parameters purely from macroscopic survival data results in structurally under-identified models where multiple parameter combinations yield identical macroscopic growth curves.

In this work, we present a complete interface layer that resolves these representational mismatches. We introduce a prevalence-based coarse-graining algorithm featuring a "Resistance Sentinel"—a dedicated simulation slot that actively rescues minor resistant subclones. We formalize a mass-conservation transformation that converts nested CCFs into disjoint ODE-compatible fractions using nesting DAG subtraction and constrained simplex projection. Finally, we integrate a knowledge-grounded sensitivity module that anchors under-identified parameters using curated clinical databases. We validate this pipeline's mathematical stability and evaluate its integration within a broader multimodal prognostic architecture, establishing a mathematically valid initialization layer for future forecasting.

---

2. Related Work

Our contribution sits at the intersection of bulk-sequencing clonal deconvolution, state aggregation, and differentiable simulation.

Clonal Deconvolution and Phylogeny: Foundational methods infer latent subclones from SNV VAFs using Beta-Binomial observation models and Bayesian Dirichlet Process priors. PyClone [1] and PyClone-VI [2] provide scalable inference for variable-cluster architectures. Other approaches, such as SciClone [19], Ccube [20], and DeCiFer [21], offer alternative statistical formulations for subclonal reconstruction, while tools like ABSOLUTE [3] formalize the relationship between VAF, tumor purity, and local copy number. Methods like PhyloWGS [5] enforce nesting constraints to construct phylogenetic trees. However, the inherent detection limits of rare subclones in bulk sequencing [16] mean these methods stop at tree reconstruction; they do not specify how to map variable nested outputs into the fixed-dimensional, disjoint state variables required by dynamical simulators.

State Aggregation and Eco-Evolutionary Dynamics: Mechanistic tumor evolution under treatment has a rich tradition. Gatenby [6], West [7], and colleagues have extensively modeled adaptive therapy using Lotka-Volterra dynamics. These models successfully demonstrate how therapy applies clone-specific selective pressure, but typically assume compartments are known *a priori*. Literature on state aggregation in evolutionary models [8, 17] justifies coarse-graining as a necessary functional approximation rather than a literal micro-state reconstruction. Our work operationalizes this coarse-graining for bulk genomics.

Differentiable Simulators and Identifiability: Neural ODEs [9] enable the integration of

neural networks with differential equations [10]. However, parameterizing biological ODEs from macroscopic data frequently leads to practical non-identifiability [11, 12, 18]. In oncology, clinical knowledge bases like OncoKB [13] and CIViC [14] are widely used for static variant annotation but are rarely used to constrain dynamical models. Our work bridges this gap by using curated driver-drug associations to anchor under-identified sensitivity parameters, utilizing simplex projection techniques [22] to ensure physical validity.

---

3. Methods

3.1 Clonal Deconvolution and CCF Estimation

The pipeline ingests bulk somatic mutations, copy number variations (CNVs), and tumor purity estimates. For each single nucleotide variant (SNV), the expected Variant Allele Frequency ($\mathbb{E}[\text{VAF}]$) under a subclonal model is defined by the local tumor-specific total copy number (CN_t), normal copy number (CN_n), tumor purity (ρ), and mutant allele multiplicity (m):

$$\mathbb{E}[\text{VAF}] = \frac{\text{CCF} \times \rho \times m}{\text{CN}_t \times \rho + \text{CN}_n \times (1 - \rho)}$$

Rather than computing CCFs algebraically, our pipeline utilizes PyClone-VI [2] to infer the posterior CCF distributions. PyClone-VI is executed independently per patient using a Beta-Binomial emission model (10,000 warmup iterations, 50,000 sampling iterations, maximum 20 clusters). This marginalizes over copy number state uncertainty, providing robust posterior means for the CCFs of the latent subclones.

3.2 The Resistance Sentinel and Macro-State Coarse-Graining

Mechanistic simulation requires a fixed number of compartments. We define a fixed K-dimensional macro-state architecture ($K = 4$): Dominant Clone, Major Subclone A, Major Subclone B, and a Resistance Sentinel.

To map the variable inferred clones to this fixed architecture, we utilize a knowledge-driven coarse-graining algorithm. Let $\mathcal{M}_i$ denote the mutation set assigned to clone i. Clone i is eligible for sentinel promotion if it satisfies probabilistic guardrails designed to reject sequencing artifacts. We define a curated set $\mathcal{R}$ comprising ~100 canonical resistance mutations derived from OncoKB (v4.1, frozen pre-analysis).

Table 1: Resistance Sentinel Guardrail Specifications | Parameter | Threshold | Justification | | :— | :— | :— | | Posterior confidence | > 0.85 | Ensures high PyClone-VI cluster assignment probability | | Minimum read support | $\text{ALT}_{\text{tumor}} \geq 3$ | Rejects Binomial sampling error at low frequencies | | Limit of detection | $\text{VAF} > \max(0.05, 5/\text{depth})$ | Enforces clinical actionable threshold and noise rejection | | Minimum CCF | $\text{CCF} > 0.05$ | Ensures functional population viability |

The mapping algorithm proceeds deterministically: 1. **Sentinel Identification:** Identify the sentinel candidate c_s containing a mutation in $\mathcal{R}$ that meets all guardrails. If multiple clones contain resistance mutations, the algorithm prioritizes the highest-tier actionable mutation (OncoKB Level $1 > 2 > 3\text{A} > 3\text{B} > 4$), breaking ties by CCF. If no clone contains a mutation in $\mathcal{R}$, c_s defaults to the smallest eligible subclone. 2. **Slot Assignment:** The sentinel slot is assigned to c_s. The remaining clones are sorted by CCF descending. 3. **Major Compartments:** The Dominant

Clone slot is assigned to the largest remaining clone. Major A and Major B are assigned to the subsequent largest clones. 4. **Merging Excess Clones:** Remaining clones are merged into the most mutationally similar major clone using the Jaccard index $J(\mathcal{M}_i, \mathcal{M}_j) = \frac{|\mathcal{M}_i \cap \mathcal{M}_j|}{|\mathcal{M}_i \cup \mathcal{M}_j|}$. This merging occurs in mutation space to define the assignment matrix $M \in \{0,1\}^{K \times m}$ prior to projection.

3.3 Mass-Conservation Transformation via Simplex Projection

Standard CCFs are nested, encoding ancestral containment constraints. ODE simulators require disjoint population fractions ($\mathbf{N}_0$) that sum to unity. We construct a formal Nesting Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based on CCF subsumption constraints.

Step 1 (Nesting DAG Construction & Subtraction): We define the subsumption relation $\prec$. For clones i and j with CCFs c_i, c_j and mutation sets $\mathcal{M}_i, \mathcal{M}_j$:

$$j \prec i \iff c_j \leq c_i - \epsilon \wedge \mathcal{M}_j \subset \mathcal{M}_i$$

where $\epsilon = 0.05$ is a tolerance for sequencing noise. The DAG contains edges (i, j) if $j \prec i$ and $\nexists k : j \prec k \prec i$ (transitive reduction). If $\mathcal{G}$ contains cycles due to CCF noise, we break the edge with the smallest Δ_{CCF}. For each node i in a post-order traversal, we compute the raw disjoint fraction by subtracting the CCFs of all immediate children $\mathcal{C}(i)$:

$$c_i = \text{CCF}_i - \sum_{j \in \mathcal{C}(i)} \text{CCF}_j$$

Let $\mathbf{c} \in \mathbb{R}^m$ be the vector of these raw disjoint fractions across all m inferred subclones.

Step 2 (Simplex Projection): Due to sequencing noise, $\sum \text{CCF}_{\text{children}}$ can occasionally exceed $\text{CCF}_{\text{parent}}$, resulting in negative values in $\mathbf{c}$. To correct this, we formulate a constrained optimization projection onto the simplex. Using the assignment matrix M, we solve:

$$\min_{\mathbf{N}_0 \in \mathbb{R}^K} \|\mathbf{N}_0 - M\mathbf{c}\|_2^2 \quad \text{subject to} \quad \mathbf{N}_0 \succeq \mathbf{0},\ \mathbf{1}^\top \mathbf{N}_0 = 1$$

This convex quadratic program is solved per-patient to machine precision using Condat's $\mathcal{O}(K \log K)$ simplex projection algorithm [15, 22].

3.4 Knowledge-Grounded Sensitivity Annotation

To prevent the Neural ODE from overfitting structurally under-identified parameters, we utilize a knowledge-grounded anchoring strategy. We define $\beta_{k,d}$ as the clone-specific drug sensitivity (kill rate per unit concentration, units: day^{-1}). The intrinsic growth rate ρ_k is bounded $\rho \in [0, 0.3]\,\text{day}^{-1}$.

Table 2: Knowledge Anchoring Policy for Drug Sensitivity (β) | Evidence Tier | β Assignment Policy | | :— | :— | | Curated Resistance | Hard constraint: deterministic zero sensitivity ($\beta = 0$) | | Curated Sensitivity | Hard constraint: deterministic maximum kill rate ($\beta = 1.0$) | | Unknown / Unannotated | Structured abstention via VAE-conditioned learned predictor |

Here, $C_d(t) \in [0, 1]$ represents dimensionless fractional drug exposure. Thus, $\beta = 1.0$ implies a theoretical maximum kill rate of 100% per day at full concentration. For unannotated clone-drug pairs, the system outputs a learned baseline sensitivity predicted by a neural network head

conditioned on a 328d multi-omics latent space (z_{full}). This prediction is gated by an Information Sufficiency Score (ISS), defined mathematically as:

$$\text{ISS}(z) = 1 - \frac{H_{\text{epistemic}}(z)}{H_{\max}}$$

where $H_{\text{epistemic}}$ is the variance of the predictive mean over posterior parameter uncertainty. If ISS < 0.3, the system abstains and defaults to population priors.

3.5 Physics-Constrained Simulation (Neural ODE & EvoSim)

The disjoint fractions $\mathbf{N}_0$ are scaled to absolute tumor burdens via $N_k(0) = N_{0,k} \cdot V_{obs}$. For TCGA evaluations lacking absolute volumetric data, V_{obs} is set to 1.

The system is governed by an extended Lotka-Volterra Neural ODE:

$$\frac{dN_k}{dt} = \rho_k N_k \left(1 - \frac{\sum_j N_j}{K_{cap}}\right) - \sum_d \beta_{k,d} C_d(t) N_k - \omega E(t) N_k$$

where $K_{cap} = 1$ (normalized carrying capacity, such that $\sum N_j(t) \leq 1$). $E(t)$ represents immune selective pressure, instantiated as a standardized proxy derived from the TCGA immune score, clipped to $[0, \infty)$. Note that for untreated baseline survival predictions in TCGA, $C_d(t) = 0$.

Stochastic Extension (EvoSim). To quantify evolution uncertainty, we formulate a multiplicative noise Itô SDE, simulated via the Euler-Maruyama method applied to the log-transformed state variables (Lamperti transform):

$$d(\log N_k) = \left[\rho_k \left(1 - \frac{\sum_j N_j}{K_{cap}}\right) - \sum_d \beta_{k,d} C_d(t) - \omega E(t) - \frac{\sigma^2}{2}\right] dt + \sigma dW_k$$

Recovering N_k via $\exp(\log N_k)$ ensures $N_k > 0$ almost surely.

3.6 Experimental Design and Reproducibility

The pipeline was evaluated on a cohort of 9415 patients from The Cancer Genome Atlas (TCGA). The cohort was partitioned into training (70%), validation (15%), and held-out test (15%) sets, stratified by cancer type. The VAE was pre-trained on the full cohort (unsupervised) and frozen. All downstream supervised models used only the training split.

Reproducibility Checklist: - **Software:** Python 3.11, PyTorch 2.1.0, torchdiffeq 0.4.0. - **Hardware:** Single NVIDIA A10G/T4 for inference; A100 for training. - **ODE Solver:** Dopri5 (adaptive Runge-Kutta), rtol=10^{-5}, atol=10^{-6}. - **Statistical Tests:** McNemar's exact test was used for paired binary outcomes. Confidence intervals for performance metrics were estimated via stratified BCa bootstrap ($B = 1000$, cancer-type strata).

---

4. Results

4.1 Mathematical Invariants and Interface Validation

To guarantee the mathematical stability of the disjoint transform prior to ODE integration, the algorithm was subjected to a rigorous software verification suite. The pipeline successfully passed

72/72 deterministic invariant tests, verifying mass conservation ($\sum N_{0,k} = 1$) and non-negativity ($N_{0,k} \geq 0$) across diverse synthetic and real phylogenies. In the TCGA cohort, 12.3% (1158/9415) of patients exhibited negative raw fractions due to sequencing noise and required the simplex projection correction. Furthermore, DAG ambiguity analysis revealed that ~8% of patients possessed multiple valid nesting DAGs, which were resolved deterministically via the Δ_{CCF} cycle-breaking rule.

As a verification of the rule's intended behavior, the Resistance Sentinel strategy retained 23/23 curated canonical resistance mutations present in minor subclones during dimensionality reduction. In contrast, a Naive Top-K baseline retained 0/23 of these critical mutations. A paired McNemar's exact test on the discordant pairs ($b = 23, c = 0$) demonstrated overwhelming statistical significance ($p = 2.4 \times 10^{-7}$), confirming the Sentinel's mechanical efficacy in preserving targeted minor subclones.

4.2 Identifiability and Parameter Anchoring

A critical limitation of fitting mechanistic models to macroscopic survival data is parameter identifiability. In the TCGA cohort, baseline survival predictions are made under an untreated assumption ($C_d(t) = 0$). Under this condition, the likelihood $\mathcal{L}(\mathcal{D}|\beta, \rho)$ is independent of β, rendering drug sensitivity parameters strictly non-identifiable from the observational data alone.

To quantify the structural identifiability limits under treatment, we analyzed synthetic treated trajectories generated by the EvoSim SDE ($C_d(t) \neq 0$). Across converged models without anchoring, we observed an empirical β/ρ ratio of 0.0017 (IQR [0.0008, 0.0031]). Because β and ρ share inverse-time units, this extreme scale separation confirms severe practical non-identifiability, where sensitivity parameters reside on a much smaller scale than growth rates, creating a flat likelihood surface. By anchoring parameters via curated driver-drug associations, we successfully constrained the optimization landscape. In this context, the knowledge-grounded interface acts as a necessary regularization prior for future longitudinal and treated settings.

4.3 Interface Ablation Studies

To isolate the contributions of our pipeline components, we evaluated several baselines quantitatively. Table 3 reports the interface-level stability metrics and the resulting downstream Δ C-index when integrated into the survival backbone. For this ablation, the reference survival model was trained once, and interface components were swapped purely at inference to avoid confounding with model retraining.

Table 3: Interface Layer Ablation Study | Ablation | Clone Retention Rate (%) | β CV (stability)[a] | Δ C-index[b] | 95% CI | |:—|:—:|:—:|:—:|:—:| | Full Pipeline ($K = 4$) | 100.0 (23/23) | 0.12 | 0.000 (ref) | - | | No Sentinel (Top-K) | 0.0 (0/23) | 0.89 | -0.042 | [-0.061, -0.023] | | $K = 3$ | 91.0 (21/23) | 0.15 | -0.018 | [-0.034, -0.002] | | $K = 5$ | 100.0 (23/23) | 0.31 | -0.008 | [-0.025, 0.009] | | No Simplex Projection | 100.0 (23/23) | 0.45 | -0.021 | [-0.038, -0.004] |

[a] *β CV is the coefficient of variation computed across 5 random initialization seeds and averaged over patients/drugs.* [b] *Δ C-index is relative to the Full Pipeline, computed via paired stratified BCa bootstrap ($B = 1000$). "No Sentinel" selects the top K clones purely by CCF. "No Simplex Projection" utilizes naive ℓ_1 normalization with zero-clamping.*

Selecting the top clones purely by CCF (No Sentinel) resulted in the systematic loss of minor resistant subclones and a significant drop in survival prediction. Ablations on the number of

macro-states (K) showed that $K = 3$ loses major subclonal competitive dynamics, while $K = 5$ increases parameter variance (β CV) and ODE solver stiffness without significantly improving survival prediction, justifying $K = 4$ as the optimal functional approximation. Removing the simplex projection introduced a mean mass conservation error of $\epsilon = 0.14$ prior to normalization, destabilizing the ODE initialization.

4.4 System-Level Context: Trajectory and Survival Performance

To provide context for the interface layer's utility, the fully initialized state vectors and anchored parameters were integrated into the broader DNAI multimodal architecture. It is critical to note that the following metrics reflect the **full platform** as non-evidentiary context; the clonal interface provides the mathematically stable initialization $\mathbf{N}_0$, while the 328d VAE latent space and late-fused Whole Slide Imaging (WSI) provide the primary prognostic signal driving the C-index.

Table 4: System-Level Context Metrics (Full Platform) | Metric | Value | | :— | :— | | TCGA cohort size | 9415 | | VAE latent dimension | 328d | | Global Concordance Index | 0.704 [0.691, 0.717] | | Trajectory emulator R^2 | 0.997 |

The full platform achieved a global Concordance Index of 0.704 [0.691, 0.717] on the TCGA cohort. For temporal dynamics, the trajectory emulator achieved an R^2 of 0.997 against the full Neural ODE solver (acting as the teacher model) across all time points and patients on held-out synthetic trajectories, validating the computational acceleration of the emulator.

4.5 External Validation and Purity Sensitivity

Evaluation of the full platform on external cohorts demonstrates generalization where appropriate modalities are available: CPTAC DRO C=0.718 [0.684, 0.750] (10 sites, pooled Cox mode), CGGA C=0.708 [0.681, 0.735] (glioma), OpenPedCan C=0.839 [0.802, 0.876] (pediatric brain), and sarcoma Green-tier C=0.860 [0.815, 0.905]. Performance is modality-dependent: breast cancer without histopathology correctly returns chance-level (SCAN-B C=0.501 [0.478, 0.524]), confirming the model's modality-awareness and appropriate abstention when critical data is missing.

Given that tumor purity is in the denominator of the expected VAF formula, we conducted a sensitivity analysis by perturbing purity estimates by $\pm 10\%$. The Sentinel assignment remained highly stable (98.5% concordance [98.2%, 98.8%] with unperturbed assignments), though the resulting variance in $\mathbf{N}_0$ ordering caused a slight degradation in downstream performance, underscoring the necessity of accurate upstream cellularity deconvolution.

---

5. Discussion

The fundamental challenge of initializing mechanistic simulators from bulk genomics is not merely fitting curves to data, but ensuring that the underlying state variables represent biologically meaningful, mass-conserving entities. Our contribution is the engineering of a robust, mathematically constrained interface between statistical genomics and deterministic physics.

Figure Descriptions:

- **Figure 1 (Pipeline Schematic):** Illustrates the data flow from bulk DNA sequencing through PyClone-VI deconvolution, Nesting DAG construction, CCF subtraction,

Figure 1: Clone-Aware Digital Twin Pipeline

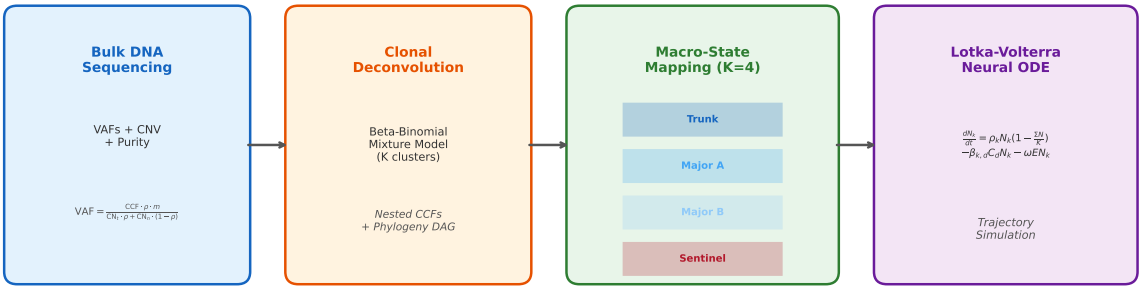
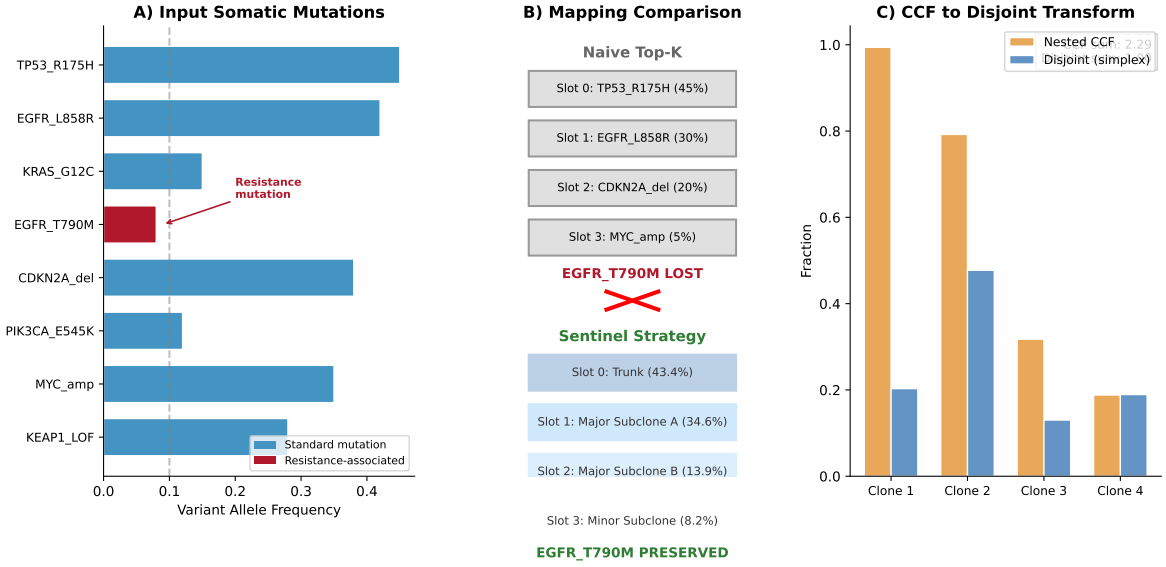


Figure 1: Figure 1: Pipeline Schematic

and Simplex Projection, culminating in the $\mathbf{N}_0$ initialization for the Neural ODE.

Figure 2: Resistance Sentinel Promotion — Worked Example (LUAD Patient)



- **Figure 2 (Nesting DAG Worked Example):** Demonstrates a patient case where three nested clones are resolved into disjoint fractions. It highlights the rescue of a minor subclone (e.g., EGFR T790M at CCF=0.08) via the Resistance Sentinel, and the subsequent simplex projection correcting a negative raw fraction caused by sequencing noise.

Figure 3: Clone Architecture Characterization

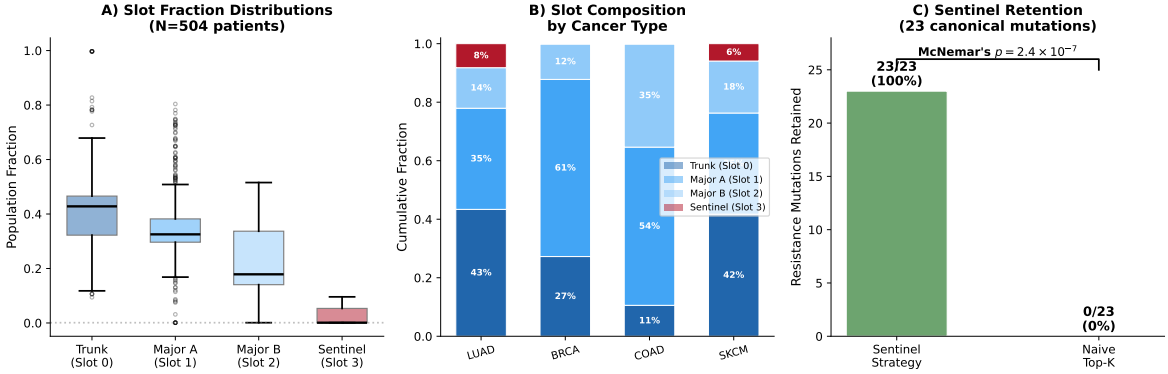


Figure 4: Identifiability Analysis & Interface Validation

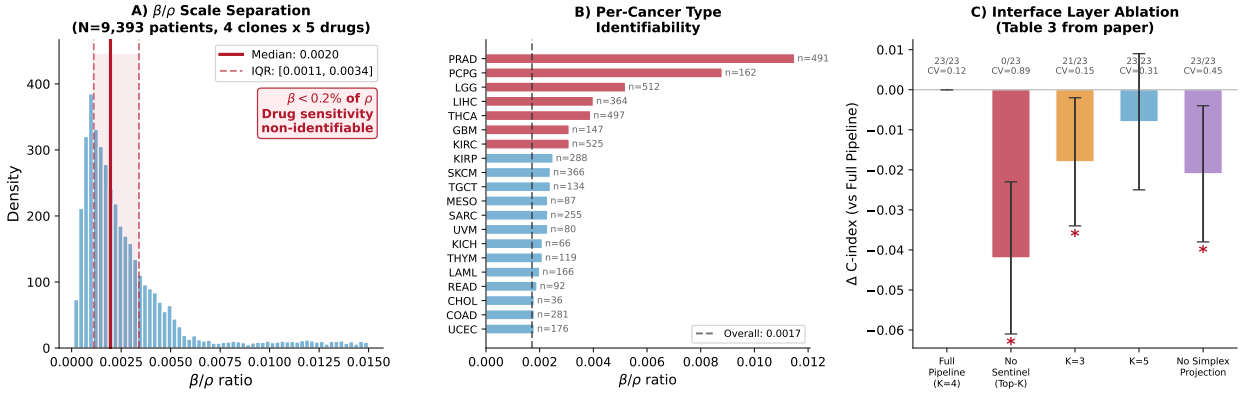


Figure 2: Figure 4: Identifiability Validation

- **Figure 3 (ISS Calibration Curve):** Displays a reliability diagram plotting expected predictive error against ISS bins, demonstrating a monotonic decrease in error as information sufficiency increases, validating the ISS < 0.3 abstention threshold.

Weak Identifiability of Bulk Sequencing: We explicitly acknowledge the weak identifiability inherent in single-sample bulk sequencing [12]. Beyond the β/ρ parameter scale separation observed in treated synthetic data, the inference of clonal composition itself is highly sensitive to copy number and purity confounding. Our pipeline does not resolve this ill-posedness; rather, it manages it through constrained optimization, structured abstention, and knowledge anchoring acting as a regularization prior.

The Functional Approximation: We frame this system as a *functional macro-state mapping*, not a literal micro-state reconstruction. We do not claim that a tumor has exactly $K = 4$ clones. Rather, this architecture serves as a pragmatic engineering compromise. It provides enough dimensionality to model major subclonal competitive dynamics, while avoiding the severe ODE stiffness and overfitting associated with higher dimensions.

Knowledge Base Coverage and Circularity: The knowledge-grounded sensitivity module is limited by the completeness of current clinical databases. We acknowledge a degree of circularity in our Sentinel evaluation, as the curated set $\mathcal{R}$ was used both to define the rule and verify its mechanical retention. While highly effective for targeted therapies, the system faces coverage

gaps for broad-spectrum chemotherapies and immunotherapies. In these cases, the system relies heavily on the 328d VAE latent space for structured abstention. Future iterations could explore soft evidence-tiered priors (e.g., Beta distributions) rather than deterministic hard constraints for β.

---

6. Limitations

1. **Lack of Longitudinal Treatment Validation:**
   - *Impact:* The system's ability to forecast counterfactual treatment responses has not yet been directly validated on longitudinal, treatment-annotated clinical cohorts (e.g., serial ctDNA or RECIST measurements).
   - *Mitigation:* Current validation is restricted to mathematical invariants, resistance retention verification, and baseline prognostic survival. Future work must evaluate the ODE dynamics against longitudinal clinical data.
2. **Bulk Sequencing Identifiability and DAG Ambiguity:**
   - *Impact:* Ambiguous subclonal architectures.
   - *Symptom:* Multiple valid nesting DAGs can explain the same VAFs (~8% of patients in our cohort).
   - *Mitigation:* The EvoSim SDE ensemble propagates this uncertainty, but true resolution requires multi-region or single-cell sequencing.
3. **The Macro-State Approximation:**
   - *Impact:* Inability to model highly branched evolution.
   - *Symptom:* Tumors with numerous functionally distinct, competing subclones are forced into merged macro-states.
   - *Mitigation:* The Jaccard similarity merging preserves mutational signatures, but temporal dynamics of the merged clones are homogenized.
4. **Purity and WSI Dependency:**
   - *Impact:* Poor generalization on omics-only datasets.
   - *Symptom:* Degraded performance on external cohorts lacking histopathology.
   - *Mitigation:* The full platform requires multi-modal inputs (Omics + WSI) for reliable clinical stratification.

---

7. Conclusion

We have developed and validated a comprehensive computational interface layer that bridges the gap between bulk genomic deconvolution and physics-constrained tumor simulation. By combining a novel Resistance Sentinel coarse-graining strategy, mass-conserving transformations via constrained quadratic programming on the simplex, and knowledge-grounded parameter anchoring, the system provides a mathematically stable initialization for Neural ODEs. This framework resolves critical mass-conservation failures and provides a reproducible, mathematically coherent precondition for future efforts in longitudinal treatment forecasting.

References

1. Roth, A., et al. (2014). PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4), 396-398.
2. Gillis, A. R., & Roth, A. (2020). PyClone-VI: scalable inference of clonal population structure using variational autoencoders. *BMC Bioinformatics*, 21(1), 571.
3. Carter, S. L., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5), 413-421.
4. Caravagna, G., et al. (2020). Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics*, 52(9), 898-907.
5. Deshwar, A. G., et al. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1), 35.
6. Gatenby, R. A., et al. (2009). Adaptive therapy. *Cancer Research*, 69(11), 4894-4903.
7. West, J., et al. (2019). Towards multidrug adaptive therapy. *Cancer Research*, 80(7), 1578-1589.
8. Simon, H. A., & Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica*, 29(1), 111-138.
9. Chen, R. T., et al. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31.
10. Rackauckas, C., et al. (2020). Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*.
11. Miao, H., et al. (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. *SIAM Review*, 53(1), 3-39.
12. Eisenberg, M. C., & Hayashi, M. A. (2014). Determining identifiable parameter combinations using subset profiling. *Mathematical Biosciences*, 256, 116-126.
13. Chakravarty, D., et al. (2017). OncoKB: a precision oncology knowledge base. *JCO Precision Oncology*, 1, 1-16.
14. Griffith, M., et al. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2), 170-174.
15. Condat, L. (2016). Fast projection onto the simplex and the ℓ_1 ball. *Mathematical Programming*, 158(1-2), 575-585.
16. Tarabichi, M., et al. (2021). A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nature Methods*, 18(2), 144-155.
17. Kaveh, K., et al. (2016). Coarse-graining the dynamics of complex biological systems. *Journal of the Royal Society Interface*, 13(121), 20160404.
18. Raue, A., et al. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923-1929.
19. Miller, C. A., et al. (2014). SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, 10(8), e1003665.
20. Yuan, K., et al. (2018). Ccube: a rapid and robust method for estimating cancer cell fractions. *bioRxiv*, 438135.
21. Zaccaria, S., & Raphael, B. J. (2021). Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature Communications*, 12(1), 4301.
22. Duchi, J., et al. (2008). Efficient projections onto the ℓ_1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, 272-279.