Toward Latent-Space Clonal Decomposition of Bulk Multi-Omics: A Proposed Framework for Resolving Functional Heterogeneity

Per Magnus Swedenborg[1,*]

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

Abstract

Background: Intratumor heterogeneity is a primary driver of therapeutic resistance. While mutational deconvolution can identify subclonal architectures, it fails to capture the multi-omic functional phenotypes required to predict drug sensitivity. Consequently, downstream digital twin simulators utilizing bulk omics suffer from severe parameter under-identification. In bulk-trained mechanistic simulators, we hypothesize this manifests as a severe gradient magnitude imbalance between clonal drug sensitivity (β) and baseline growth rates (ρ). This gradient starvation causes the optimizer to explain survival variance entirely through baseline growth, rendering counterfactual treatment simulations unreliable. **Methods:** To resolve this ill-conditioning, we propose Latent-Space Clonal Decomposition (LSCD), a deep generative architecture extending our production multi-modal variational autoencoder (VAE v5.10). LSCD is designed to decompose bulk profiles into clone-specific structured latent representations (328 dimensions). The proposed architecture employs a Clone Attention Head where the effective number of clones is learned via Dirichlet sparsity. The generative model enforces modality-correct observation-space likelihoods—including Negative Binomial for RNA with size factors, purity-aware Gaussian mixing for absolute copy number, and Beta-Binomial for variant allele frequencies. To interface with downstream 4-compartment ODEs, we define a deterministic Top-4 Merge mapping and propose an explicit gradient surgery step to orthogonalize downstream ODE parameters. **Evaluation Plan:** We outline a rigorous, pre-registered evaluation protocol to validate LSCD across 9,415 patients from the TCGA cohort and orthogonal single-cell pseudo-bulk datasets. The protocol includes synthetic benchmarks to measure the reduction in the Fisher Information condition number, validation of treatment identifiability via scDNA/scRNA perturbation datasets, and a strict anti-circularity protocol utilizing chromosome-distance-stratified mutation sets to prevent weak-supervision leakage from PyClone-VI priors. **Conclusion:** LSCD provides a theoretical blueprint to bridge mutational phylogeny and functional phenotype. By mapping bulk multi-omics to identifiable, clone-specific latent macro-states, LSCD aims to enable mathematically consistent, subclone-resolved treatment response simulations for precision oncology.

1. Introduction

Treatment failure in oncology is fundamentally a failure of heterogeneity management. When a clinician selects a targeted therapy based on a single bulk molecular profile, they implicitly assume tumor uniformity—an assumption contradicted by decades of tumor biology research. In reality, tumors are complex ecosystems of genetically and phenotypically distinct subclones. For example, while a bulk profile may indicate a KRAS G12C mutation in lung adenocarcinoma, the tumor may harbor a minor, functionally divergent subclone with a co-occurring KEAP1 mutation that will inevitably drive acquired resistance.

Current computational approaches to resolve this heterogeneity are bifurcated. Mutational deconvolution methods cluster variant allele frequencies (VAFs) to infer the number of clones and their cellular prevalences. However, these methods operate exclusively in genomic space, providing no insight into the transcriptomic, epigenetic, or pathway-level functional states of the identified clones. Conversely, transcriptomic deconvolution models successfully estimate cell-type proportions and state programs, but are anchored to distinct cell lineages (e.g., immune vs. stromal) rather than genetically defined tumor subclones.

This disconnect presents a critical bottleneck for personalized oncology platforms that utilize neural ordinary differential equations (ODEs) to simulate tumor dynamics. When a single "bulk" latent representation is fed into a multi-compartment ODE simulator, the model cannot distinguish which subclone is responding to a specific therapy. In our current production architecture (utilizing VAE v5.10 across $N = 9,415$ TCGA samples), we hypothesize this manifests as a severe parameter identifiability pathology. Because baseline growth (ρ) is vastly easier to infer from bulk survival and trajectory data than clone-specific drug sensitivity (β), the optimizer starves the β parameters. This gradient degeneracy leads to an ill-conditioned Fisher Information matrix where the model explains all patient outcomes via intrinsic growth rates, ignoring the causal impact of therapy.

To address this gap, we propose Latent-Space Clonal Decomposition (LSCD). LSCD is a deep generative architecture designed to jointly decompose bulk omics data into per-subpopulation biological states directly within a structured 328-dimensional latent space.

Summary of Proposed Contributions: * **Structured Multi-Omic Clone Slots:** A novel Clone Attention Head that maps bulk RNA, DNA, CNV, and methylation into clone-specific 328-dimensional functional phenotypes, regularized by Dirichlet sparsity. * **Modality-Correct Mixing:** A generative formulation that abandons linear latent mixing in favor of observation-space likelihoods, explicitly accounting for RNA size factors, CNV tumor purity, and Beta-Binomial variant reads. * **Mechanistic ODE Interface & Gradient Surgery:** A deterministic Top-4 Merge mapping to interface variable clone counts with fixed-compartment ODEs, coupled with a targeted gradient projection mechanism to decouple drug sensitivity gradients from proliferation subspaces. * **Rigorous Evaluation Protocol:** A pre-registered validation plan utilizing synthetic identifiability diagnostics, single-cell pseudo-bulk orthogonal validation, and chromosome-distance-stratified weak supervision.

This manuscript details the theoretical formulation of LSCD, establishes the identifiability risks and mitigations required to resolve ODE confounding, and outlines a rigorous experimental protocol for its future empirical validation.

---

2. Related Work

Mutational Clonal Deconvolution. Bayesian clustering of single nucleotide variants (SNVs) using Beta-Binomial or Dirichlet-process mixture models (e.g., PyClone-VI, DPClust) represents the gold standard for inferring clonal architecture from bulk sequencing. While these methods accurately estimate cancer cell fractions (CCFs) and phylogenetic trees, they do not map these clones to functional multi-omic phenotypes, rendering them insufficient for predicting pathway-targeted drug responses.

Expression-Based and Spatial Deconvolution. Methods such as CIBERSORTx and BayesPrism deconvolve bulk RNA-seq into constituent components. However, these methods

target distinct cell types or states using reference signatures or single-cell atlases, rather than genetically defined tumor subclones. Deep generative VAEs, notably the scVI ecosystem (e.g., DestVI), perform probabilistic deconvolution of spatial transcriptomics into cell-type proportions and continuous latent states. Recent integrative methods (e.g., clonealign, cardelino) attempt to link SNVs to expression, but rarely scale to full multi-omic (RNA, DNA, CNV, Methylation) latent spaces. Multi-omic factor models like MOFA+ and totalVI capture shared variance across modalities but do not explicitly model subclonal population structures.

Object-Centric Generative Models. In computer vision, architectures like MONet, IODINE, and Slot Attention have successfully decomposed complex scenes into distinct object representations without explicit supervision. LSCD adapts the Slot Attention paradigm to computational biology, treating tumor subclones as the "objects" to be discovered within a bulk multi-omic "scene."

Mixture Identifiability and Constrained Optimization. The identifiability of finite mixture models is a classical problem; without strong parametric assumptions or auxiliary anchors, mixture components are susceptible to label switching and component merging degeneracies. In systems biology, the structural and practical identifiability of ODE parameters is a well-documented bottleneck. In deep learning, identifiable VAEs (iVAEs) have demonstrated that conditioning on auxiliary variables can restore identifiability up to linear transformations. LSCD adapts these paradigms, proposing PyClone-VI as an auxiliary anchor and employing explicit gradient surgery to enforce structural independence between mechanistic ODE parameters.

Structured Comparison of Deconvolution and State Models

Method	Modalities	Identifiability Mechanism	Output
PyClone-VI	DNA only	Phylogenetic	Clone fractions
BayesPrism	RNA only	Empirical	Cell-type fractions
clonealign	scRNA+scDNA	SNV mapping	Clone assignments
MOFA+	Multi-omic	Orthogonality	Latent factors
LSCD (Proposed)	**RNA+DNA+CNV+Meth**	**Grad Surgery + Anchors**	**328d clone states**

Figure 1: Figure 2: Methods Comparison

Table 1: Structured Comparison of Deconvolution and State Models

Category	Method	Target Object	Modalities	Identifiability Mechanism	Output
Bulk Mutational	PyClone-VI	Tumor subclones	DNA only	Phylogenetic constraints	Clone fractions + SNV assignments
Bulk Expression	BayesPrism	Cell types/states	RNA only	Empirical (Bayesian CIs)	Fractions + imputed expression
Genotype-Phenotype	clonealign	Tumor subclones	scRNA + scDNA	SNV-to-gene mapping	Clone assignments per cell
Multi-Omic Latent	MOFA+	Latent factors	Multi-omic	Orthogonality constraints	Continuous latent factors
Multi-Omic Clonal	**LSCD (Proposed)**	Tumor subclones	RNA+DNA+CNV+Meth	Gradient surgery + Anchors	Fractions + 328d clone states

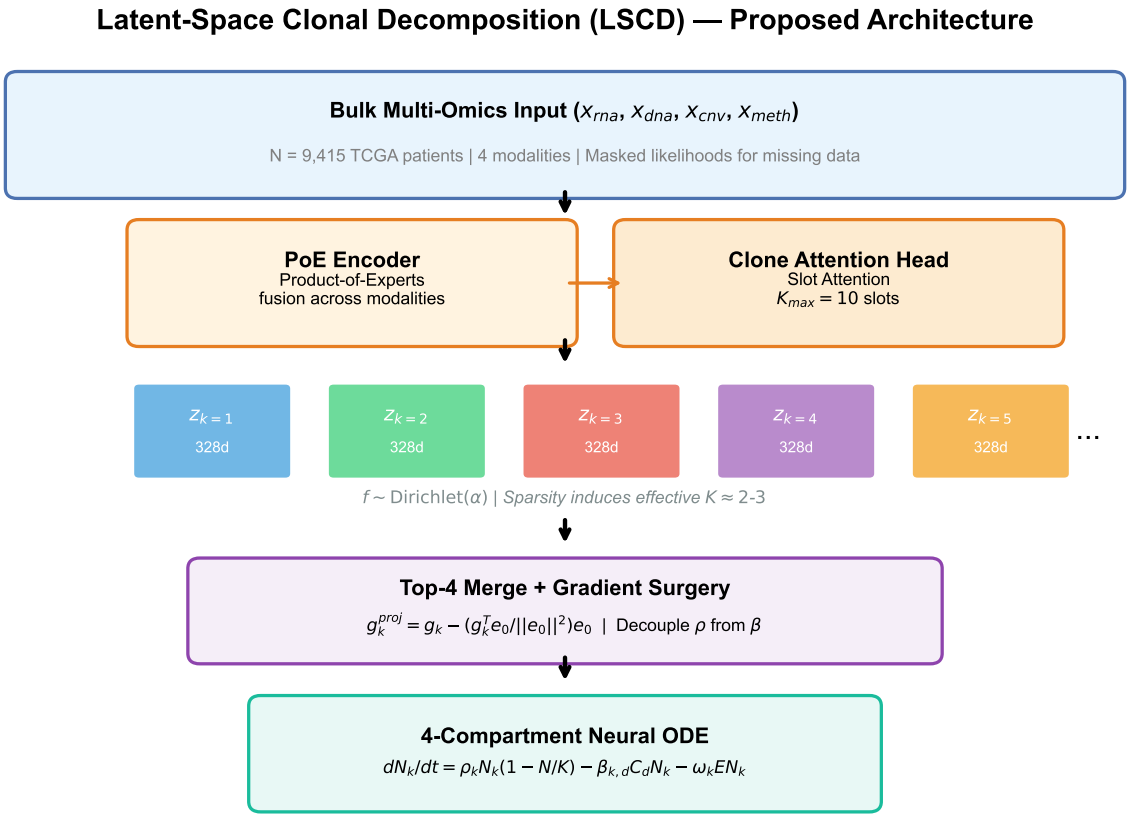

Figure 2: Figure 1: LSCD Architecture Overview

3. Methods

3.1 Formal Generative Model and Plate Notation

Let $\mathbf{x}_i = \{\mathbf{x}_{rna}, \mathbf{x}_{dna}, \mathbf{x}_{cnv}, \mathbf{x}_{meth}\}$ represent the bulk multi-omic profile of patient $i \in \{1, \ldots, N\}$, where $N = 9,415$ in the TCGA cohort. Our baseline foundation model (VAE v5.10) encodes $\mathbf{x}_i$ into a single structured latent vector $\mathbf{z} \in \mathbb{R}^{328}$.

LSCD extends this by assuming $\mathbf{x}_i$ is generated by a mixture of K distinct subclonal states, bounded by a hyperparameter $K_{\max}$. The generative process is defined as follows:

```
[Plate Diagram Representation]
Plate N (Patients i = 1...N):
  f_i ~ Dirichlet(alpha)                  <-- Weak supervision from \tilde{f}_i
  Plate K (Clones k = 1...K_max):
    z_ik ~ N(0, I)                        <-- 328d structured latent state
    s_ik ~ LogNormal(0, sigma_s^2)        <-- RNA size factor
  Plate M (Modalities m \in {RNA, DNA, CNV, Meth}):
    x_im ~ p_m( mix(f_i, s_i, g_m(z_i)) ) <-- Observation-space mixing
```

To handle missing modalities gracefully during training, we employ masked likelihoods. The reconstruction loss is evaluated only over the set of observed modalities $\mathcal{M}_i$ for patient i:

$$\mathcal{L}_{\text{recon}} = \sum_{m \in \mathcal{M}_i} w_m \cdot \log p_m(\mathbf{x}_{im} | \mathbf{z}_{i,1:K_{\max}}, \mathbf{f}_i)$$

where w_m are modality-specific balancing weights.

3.2 Clone Attention Head and Variational Inference

To perform inference, we replace the standard VAE bottleneck with a Clone Attention module. The encoder first applies Product-of-Experts (PoE) fusion across available modalities to produce patient-level hidden states $\mathbf{h}_i$. The Clone Attention Head then decomposes $\mathbf{h}_i$ into variational posterior parameters for each clone.

Rather than forcing a hard selection of K, the effective number of clones is learned via Dirichlet sparsity. We utilize a stick-breaking logistic-normal approximation to reparameterize the Dirichlet distribution, allowing gradients to flow through the sampling process.

Algorithm 1: Variational Clone Attention Head

```
Input: Hidden state h_i (from PoE encoder), K_max
Output: Sampled fractions f_i, Variational parameters {mu_ik, logvar_ik}

// Concentration network for Dirichlet prior via stick-breaking approximation
alpha_i = Softplus(MLP_alpha(h_i)) + epsilon
Sample f_i ~ LogisticNormal(alpha_i) // Reparameterized sampling

// Slot generation (Slot Attention variant)
for k in 1..K_max:
    h_ik = SlotAttention(h_i, slot_init_k)  // Competitive binding
    mu_ik = Linear_mu(h_ik)                 // R^328
    logvar_ik = Linear_logvar(h_ik)         // R^328
```

```
    z_ik = mu_ik + exp(0.5 * logvar_ik) * epsilon_k  // Reparameterization
return f_i, {mu_ik, logvar_ik}, {z_ik}
```

3.3 Modality-Correct Generative Reconstruction

A critical vulnerability in bulk deconvolution is the assumption of linear mixing in transformed spaces. LSCD enforces modality-correct likelihoods, mixing in the observation space appropriate for each modality (summarized in Table 2).

- **RNA (Negative Binomial):** Bulk RNA-seq counts are sums of clone-level counts. Cellular fraction (f_{ik}) does not equal RNA contribution fraction unless adjusted for total mRNA content. We define the RNA contribution weight as $\pi_{ik} = (f_{ik}s_{ik})/\sum_j (f_{ij}s_{ij})$. We utilize a Negative Binomial parameterization: $x_{ij,rna} \sim \text{NB}\left(\mu = \sum_k \pi_{ik}\mu_{ijk}, \theta = \phi_j\right)$, where $\mu_{ijk} = \exp(\text{Linear}(\mathbf{z}_{ik})) \cdot s_i^{\text{total}}$ and ϕ_j is a learned gene-specific dispersion.
- **CNV (Gaussian on Absolute Copy Number):** Bulk absolute copy number is a linear mixture of clonal absolute copy numbers, confounded by tumor purity ψ_i (provided as a fixed input from ESTIMATE/PurAM). The likelihood is modeled as Gaussian: $x_{ij,cnv} \sim \mathcal{N}\left(\sum_k f_{ik} \cdot (2^{c_{ik}} \cdot \psi_i + 2(1-\psi_i)), \sigma^2_{cnv}\right)$, where c_{ik} is the log2 copy ratio predicted by the decoder.
- **Methylation (Beta Approximation):** Modeled via a Beta distribution for bounded array beta-values. The decoder outputs mean $\mu_{ijk} = \sigma(\text{Linear}(\mathbf{z}_{ik}))$ and concentration $\nu_{ijk} = \text{Softplus}(\text{Linear}(\mathbf{z}_{ik})) + \nu_{\min}$. We approximate the bulk mixture via moment-matching: $x_{ij,meth} \sim \text{Beta}(\alpha_{ij}, \beta_{ij})$ where the bulk mean is $\sum_k f_{ik}\mu_{ijk}$. We acknowledge that a mixture of Betas is not strictly a Beta distribution, but this provides a computationally tractable approximation.
- **DNA (Beta-Binomial):** For mutation m with a_m alt reads and b_m ref reads, we model the variant allele frequency using a Beta-Binomial likelihood: $a_{im} \sim \text{Beta-Binomial}(n_m = a_{im} + b_{im}, \pi_{im}, \delta)$, where π_{im} is the expected VAF derived from the clone fractions f_{ik}, local copy number, and purity ψ_i, and δ is an overdispersion parameter.

Table 2: Likelihood Specifications

Modality	Distribution	Parameterization	Link Function	Mixing
RNA	Negative Binomial	μ_{ijk}, ϕ_j	$\mu = \exp(\text{Linear}(\mathbf{z}_{ik}))$	$\mu_{\text{bulk}} = \sum_k \pi_{ik}\mu_{ijk}$
CNV	Gaussian	μ_{ij}, σ^2	$\mu = 2^{c_{ik}}\psi + 2(1-\psi)$	Linear in absolute CN space
Methylation	Beta	α_{ij}, β_{ij}	$\mu = \sigma(\cdot)$, $\nu = \text{Softplus}(\cdot)$	Moment-matching (approximation)
DNA	Beta-Binomial	n_m, π_{im}, δ	π from VAF + purity	Weighted by f_{ik}

3.4 The 4-Slot ODE Interface

The DNAI platform utilizes a fixed 4-compartment Neural ODE ($K = 4$). Because LSCD proposes a variable effective clone count bounded by $K_{\max}$ (e.g., $K_{\max} = 10$), we define a deterministic, differentiable Top-4 Merge mapping $\mathcal{M} : \mathbb{R}^{K_{\max} \times d} \times \Delta^{K_{\max}-1} \to \mathbb{R}^{4 \times d} \times \Delta^3$ to interface with the downstream simulator:

1. Sort clones by fraction: $f_{(1)} \geq f_{(2)} \geq \cdots \geq f_{(K_{\max})}$.
2. For the top 3 clones ($k = 1, 2, 3$): $\mathbf{z}'_k = \mathbf{z}_{(k)}$ and $f'_k = f_{(k)}$.
3. For the 4th compartment (background merge): $\mathbf{z}'_4 = \frac{\sum_{j=4}^{K_{\max}} f_{(j)}\mathbf{z}_{(j)}}{\sum_{j=4}^{K_{\max}} f_{(j)}}$ and $f'_4 = \sum_{j=4}^{K_{\max}} f_{(j)}$.

While this mapping introduces subgradients at the sorting boundaries, it preserves the physical

interpretability of the top dominant clones required by the ODE.

3.5 Resolving Identifiability via Gradient Surgery

To resolve the hypothesized gradient starvation pathology in our downstream ODE simulator, we must constrain the latent space. In VAE v5.10, the baseline growth rate (ρ) is deterministically mapped from the proliferation slice of the latent space (z_{prolif}, which corresponds to dimension 0, i.e., the standard basis vector $\mathbf{e}_0$).

To prevent the optimizer from using the proliferation subspace to explain drug sensitivity (β), we propose a **gradient surgery** constraint applied during backpropagation. Let $\mathbf{g}_k = \nabla_{\mathbf{z}_{ik}} \mathcal{L}_{\text{drug}}$ be the gradient of the drug sensitivity loss with respect to clone k's latent representation. We project:

$$\mathbf{g}_k^{\text{proj}} = \mathbf{g}_k - \frac{\mathbf{g}_k^\top \mathbf{e}_0}{\|\mathbf{e}_0\|^2} \mathbf{e}_0$$

This enforces $\mathbf{g}_k \perp \mathbf{e}_0$ at each optimization step. This prevents the proliferation subspace from dominating the drug sensitivity gradient updates, though we acknowledge it may introduce gradient conflicts if true drug sensitivity is biologically coupled to baseline proliferation.

The total Evidence Lower Bound (ELBO) objective is:

$$\mathcal{L} = \underbrace{\sum_m \mathbb{E}_q[\log p(\mathbf{x}_m|\mathbf{z}_{1:K})]}_{\text{Reconstruction}} - \underbrace{D_{KL}(q(\mathbf{f})\|p(\mathbf{f}))}_{\text{Dirichlet Regularizer}} - \underbrace{\sum_k D_{KL}(q(\mathbf{z}_k)\|p(\mathbf{z}_k))}_{\text{Latent Regularizer}} + \underbrace{\gamma_{sup}\|\mathbf{f} - \tilde{\mathbf{f}}\|_1}_{\text{Anchor}}$$

where $\tilde{\mathbf{f}}$ are PyClone-VI derived cancer cell fractions used as weak supervision.

3.6 Weak Supervision and Anti-Circularity Protocol

Relying on PyClone-VI for weak supervision introduces a risk of circular leakage. To guarantee functional discovery, we enforce a reduced-leakage protocol: 1. **Training:** Mutations are partitioned into disjoint sets via **chromosomal distance stratification** (ensuring sets are separated by >10Mb or on entirely different chromosomes to minimize linkage disequilibrium). Set $\mathcal{M}_{supervise}$ is used exclusively to compute PyClone-VI CCFs. Set $\mathcal{M}_{likelihood}$ is reserved for the DNA Beta-Binomial reconstruction likelihood. 2. **Validation:** A third disjoint set, $\mathcal{M}_{validate}$, is reserved to evaluate the accuracy of the inferred clone fractions. We will explicitly report the cross-set correlation $\text{Corr}(\tilde{\mathbf{f}}_{\text{supervise}}, \tilde{\mathbf{f}}_{\text{validate}})$ to quantify any remaining leakage.

When comparing inferred fractions to ground truth, we solve the linear assignment problem (minimum cost bipartite matching) via the Hungarian algorithm to align clone slots.

3.7 Identifiability Risks and Mitigations

Nonlinear deep mixture models are inherently susceptible to unidentifiability due to permutation symmetry and component collapse. While weak supervision on fractions ($\tilde{\mathbf{f}}$) anchors the proportions, it does not automatically align the semantic meaning of the 328-dimensional slots across patients.

To mitigate this, we propose initializing LSCD slots with canonical clone signatures derived from cohort-wide PyClone-VI consensus clusters. Furthermore, to empirically validate the identifiability of downstream ODE parameters, we will utilize Simulation-Based Calibration (SBC) and Profile

Likelihood analyses on synthetic data, testing whether the likelihood ratio test successfully rejects null hypotheses for β when gradient surgery is applied.

Table 3: Proposed Model Hyperparameters | Parameter | Symbol | Value | Justification | |———|——|—-|———–| | Max clones | $K_{\max}$ | 10 | Exceeds typical subclone count (2-4) | | Dirichlet concentration | α_0 | 0.1 | Induces sparsity (effective clones $\approx 2-3$) | | Gradient surgery weight | λ_{GS} | 1.0 | Applied every iteration | | 4-slot mapping | $\mathcal{M}$ | Top-4 Merge | Interfaces with DNAI Neural ODE |

---

4. Evaluation Plan

Because LSCD is a theoretical architecture designed to solve a specific mathematical pathology, we outline a rigorous, pre-registered evaluation protocol to validate its efficacy. We will evaluate LSCD on the 9,415 TCGA patients using 5-fold cancer-stratified cross-validation, supplemented by orthogonal single-cell datasets.

4.1 Resolution of Parameter Identifiability (Synthetic Benchmark)

The primary technical mandate of LSCD is to resolve the under-identification of clonal drug sensitivity (β). Because true ODE parameters are unobservable in human patients, we will utilize our stochastic ground-truth simulator, EvoSim, to generate a synthetic multi-omics benchmark where per-clone latents ($\mathbf{z}_k$), fractions (f_k), and ODE parameters (ρ_k, β_k) are mathematically known.

We define the **Gradient Imbalance Ratio (GIR)** based on loss gradients:

$$\text{GIR} = \frac{\mathbb{E}_{\mathcal{B}}[\|\nabla_\phi \mathcal{L}_{\text{drug}}\|_2]}{\mathbb{E}_{\mathcal{B}}[\|\nabla_\phi \mathcal{L}_{\text{survival}}\|_2]}$$

where ϕ represents the hypernetwork parameters, $\mathcal{L}_{\text{drug}}$ is a drug response loss, and $\mathcal{L}_{\text{survival}}$ is the Cox survival loss. Success is defined as a directional increase in the GIR toward 1.0, indicating that the parameter space has become well-posed for joint estimation. We will also compute the Fisher Information condition number $\kappa(\mathcal{J})$ to formally quantify identifiability improvements.

4.2 Orthogonal Ground Truth Validation (Treatment Identifiability)

TCGA lacks the longitudinal tumor burden and clone-level response data required to fully validate treatment identifiability. Therefore, we will validate LSCD's ability to recover clone-specific drug sensitivities (β) using orthogonal datasets: 1. **scDNA+scRNA Pseudo-Bulk:** We will utilize datasets such as MIX-Seq, creating synthetic bulk profiles by mixing scRNA-seq profiles from known subclones with known perturbation responses. We will test if LSCD accurately recovers the mixing proportions and clone-specific drug response signatures. 2. **PDX Perturbation Cohorts:** We will evaluate LSCD on in-house PDX drug screens with pre- and post-treatment sequencing to verify that inferred β parameters correctly predict subclonal expansion/contraction under therapy.

4.3 TCGA Held-out Modality Reconstruction and Stability

To prove that LSCD learns true functional biology rather than simply distilling PyClone mutation clusters, we will evaluate held-out modality reconstruction on TCGA. We will predict the PyClone-VI consensus clusters using only the transcriptomic and epigenetic latent slices.

Given the inherent degeneracies of mixture models, we will quantify the stability of the Dirichlet sparsity mechanism. Bootstrap resampling will be performed at the feature level to preserve inter-modality correlations. The test-retest correlation is computed as the Pearson r between $\mathbf{f}_i$ and $\mathbf{f}_i^{(b)}$ across bootstrap samples b. Furthermore, we will monitor the degradation of the model when PyClone-VI supervision is withheld, reporting the fraction of patients that gracefully trigger the native $K = 1$ fallback.

4.4 Phase 0 Feasibility Protocol

To validate core mechanisms before full multi-modal implementation, we propose a minimal feasible prototype: * **Input:** RNA-seq + synthetic binary mutation matrix. * **Baselines:** (1) Bulk VAE ($K = 1$, current production baseline), (2) PyClone-VI alone, (3) MOFA+ (multi-omic latent baseline), (4) cardelino (genotype-to-phenotype clone linking), (5) Naive Mixture VAE (LSCD without gradient projection or modality-correct likelihoods). * **Power Analysis:** We will use bootstrap resampling of pilot experiments to estimate the sampling distribution of log(GIR) under the null and alternative hypotheses, computing empirical power while controlling the False Discovery Rate (FDR) at $\alpha = 0.05$ across the 33 cancer types.

---

5. Discussion

LSCD represents a proposed fundamental shift in how bulk multi-omics are processed for predictive oncology. By moving from a single macro-state representation to a multi-slot clonal decomposition, we aim to bridge the historical divide between mutational phylogeny and functional phenotype.

The clinical implications for digital twin platforms are substantial. Consider a hypothetical clinical vignette of a patient with metastatic KRAS G12C-mutant lung adenocarcinoma. A standard bulk profile recommends monotherapy with a KRAS inhibitor. However, LSCD decomposition might reveal two distinct clones: Clone 1 (high proliferation, sensitive to KRAS inhibition) and Clone 2 (co-occurring KEAP1 mutation, predicted resistant). By mapping specific drug sensitivities (β_k) to specific clones ($\mathbf{z}_k$), the downstream simulator can accurately forecast that monotherapy will control Clone 1 but provoke the rapid expansion of Clone 2.

Methodologically, the LSCD proposal highlights that attention-based mixture VAEs require modality-correct likelihoods (mixing in observation space with explicit size factors) and explicit identifiability constraints. Without anchored parameterization and gradient surgery, mixture models may reconstruct data well but fail to yield the identifiable parameters required by mechanistic ODEs. Furthermore, the necessity of the Top-4 Merge mapping underscores the architectural friction between variable-component generative models and fixed-compartment mechanistic simulators.

---

6. Limitations

As a theoretical architecture and protocol, LSCD has several acknowledged limitations that must be addressed during implementation. First, unsupervised mixture models are subject to permutation invariance and clone merging degeneracies. While Dirichlet sparsity and anchored parameterization mitigate this, cross-patient slot alignment remains a non-trivial challenge. Second, absolute

CNV likelihoods and DNA Beta-Binomial models are heavily confounded by tumor purity; if purity (ψ_i) is not accurately modeled during preprocessing, the observation-space mixing approximations will degrade. Third, the moment-matching approximation for the Methylation Beta mixture is mathematically inexact, as a mixture of Betas is not a Beta distribution. Fourth, while our anti-circularity protocol utilizes chromosomal distance stratification, linkage disequilibrium and broad transcriptional consequences of aneuploidy may still create correlation across held-out and supervised sets. Finally, spatial heterogeneity dictates that a single bulk biopsy captures only a fraction of the tumor ecosystem. Integration with multi-region sequencing or longitudinal liquid biopsies will ultimately be required to capture the complete clonal architecture.

---

7. Conclusion

We introduced Latent-Space Clonal Decomposition (LSCD), a proposed generative architecture designed to extract subclone-specific functional latent states from bulk multi-omics. By integrating mutational priors with a Clone Attention Head, Dirichlet sparsity, and modality-correct observation-space likelihoods, LSCD provides a mathematical blueprint to overcome the identifiability limits of bulk sequencing. When coupled with downstream neural ODEs via a Top-4 Merge mapping and explicit gradient surgery, LSCD aims to resolve the critical gradient starvation pathology. Future empirical work will focus on executing the pre-registered evaluation protocol on synthetic benchmarks, single-cell pseudo-bulk datasets, and the TCGA cohort to validate the framework's capacity for mathematically consistent simulations of clonal evolution.

References

1. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* **11**, 396-398 (2014).
2. Gillis, M. A. & Roth, A. PyClone-VI: scalable inference of clonal population structure using Dirichlet process mixture models. *Bioinformatics* **36**, 4208-4210 (2020).
3. Lopez, R. et al. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053-1058 (2018).
4. Lopez, R. et al. DestVI identifies continuums of cell types in spatial transcriptomics data. *Nature Methods* **19**, 1365-1375 (2022).
5. Khemakhem, I. et al. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. *AISTATS* (2020).
6. Yu, T. et al. Gradient Surgery for Multi-Task Learning. *Advances in Neural Information Processing Systems* **33**, 5824-5836 (2020).
7. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology* **37**, 773-782 (2019).
8. Chen, R. T. Q. et al. Neural Ordinary Differential Equations. *Advances in Neural Information Processing Systems* **31** (2018).
9. Satas, G. et al. Pairtree: fast, accurate inference of cancer evolutionary surfaces from bulk sequencing data. *Genome Biology* **22**, 263 (2021).
10. Teicher, H. Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics* **34**, 1265-1269 (1963).

11. Allman, E. S. et al. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* **37**, 3099-3132 (2009).
12. Jacobs, R. A. et al. Adaptive Mixtures of Local Experts. *Neural Computation* **3**, 79-87 (1991).
13. Eslami, S. M. A. et al. Attend, infer, repeat: fast scene understanding with generative models. *Advances in Neural Information Processing Systems* **29** (2016).
14. Locatello, F. et al. Object-Centric Learning with Slot Attention. *Advances in Neural Information Processing Systems* **33**, 11525-11538 (2020).
15. Campbell, K. R. et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biology* **20**, 54 (2019).
16. McCarthy, D. J. et al. Cardelino: computational integration of core mutations and single-cell transcriptomes. *Nature Methods* **17**, 414-421 (2020).
17. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 111 (2020).
18. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods* **18**, 272-282 (2021).
19. Raue, A. et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923-1929 (2009).
20. Burgess, C. P. et al. MONet: Unsupervised Scene Decomposition and Representation. *arXiv preprint arXiv:1901.11390* (2019).
21. Greff, K. et al. Multi-Object Representation Learning with Iterative Variational Inference (IODINE). *ICML* (2019).
22. McFarland, J. M. et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications* **11**, 4296 (2020).