Resistance Forecasting via Clonal Digital Twins: A Pre-Registered Retrospective Validation Protocol

Per Magnus Swedenborg[1,*]

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

Abstract

Background: Mechanistic digital twins promise to forecast tumor evolution and optimize treatment sequencing. However, existing models are predominantly validated against static survival metrics or bulk tumor volume, failing to evaluate whether they accurately predict the identity of emerging resistant subclones. To bridge this gap, we propose a rigorous, pre-registered protocol for the retrospective validation of clone-aware digital twins against longitudinal patient data. **This manuscript is a validation protocol; no completed cohort validation results are reported here. Methods:** We will execute this validation protocol using the DNAI digital twin platform on longitudinal paired biopsies from the GLASS (glioma, target $n = 222$), TRACERx (NSCLC), and GENIE BPC cohorts. To prevent information leakage, baseline (T_1) bulk sequencing will be deconvolved via PyClone-VI in strict isolation from recurrence (T_2) data. To bridge variable inferred mutation-cluster prevalences with fixed simulation compartments, we formalize a "Resistance Sentinel" normalized prevalence-to-burden mapping that protects minor subclones harboring resistance drivers. Forward trajectories will be generated using a 100-member Itô stochastic differential equation (SDE) ensemble to propagate posterior deconvolution uncertainty into probabilistic compartment fate predictions. **Results (Planned Analyses):** The primary endpoint is Clone Fate Concordance (CFC) at progression, evaluated against a patient-specific frequency-proportional null distribution using a stratified Poisson-binomial test. Secondary endpoints include Expected Dominant Compartment Concordance (scored via the Brier score) and multiclass calibration. The underlying DNAI platform has been strictly calibrated prior to this protocol, achieving C-indices of 0.704 (internal), 0.718 (CPTAC external, DRO), 0.839 (OpenPedCan pediatric fusion), 0.860 (sarcoma Green-tier), and 0.708 (CGGA glioma cross-ethnic), validated across 7 external cohorts totaling 236,000+ patients (note: GENIE contributes 228,000 patients for driver validation only, not survival prediction). **Conclusion:** By shifting the validation target from static risk to longitudinal compartment dominance, this protocol provides a reproducible blueprint for evaluating mechanistic tumor models, establishing a framework to assess proactive, evolution-informed treatment sequencing.

1. Introduction

Precision oncology is undergoing a paradigm shift from static prognostic stratification to dynamic, counterfactual treatment optimization. While current multi-omics machine learning models excel at predicting baseline risk, they fundamentally answer *what* is happening rather than *when* and *how* progression will occur. Mechanistic digital twins—computational models that simulate tumor dynamics using physics-constrained differential equations—promise to fill this gap by forecasting evolutionary trajectories under selective therapeutic pressure.

Despite rapid architectural advancements, the clinical translation of digital twins is bottlenecked

by a lack of rigorous, clone-aware validation frameworks. Historically, predictive oncology models are validated using static survival endpoints (e.g., overall survival concordance index) or bulk radiographic volume changes. However, clinical progression is driven by the competitive release of specific resistant subclones. A model that correctly predicts the time to progression but hallucinates the identity of the dominant resistant lineage could misinform subsequent lines of therapy. Validating the identity of emerging compartments requires longitudinal molecular data, which has only recently become available at scale through cohorts like TRACERx, GLASS, and GENIE BPC.

In this paper, we propose a comprehensive, pre-registered protocol for the retrospective validation of clone-aware digital twin predictions against longitudinal patient data. We operationalize this framework using the DNAI platform (v2.1), a production digital twin that maps multi-omics and whole-slide imaging (WSI) into a structured latent space to parameterize a differentiable stochastic Neural ODE (SDE). To overcome the structural mismatch between variable-resolution bulk sequencing and fixed-compartment mathematical models, we introduce the "Resistance Sentinel"—a knowledge-grounded mapping algorithm that ensures clinically actionable minor mutation clusters are preserved during simulation initialization. Crucially, we address the fundamental non-identifiability of bulk sequencing by propagating posterior deconvolution uncertainty through a 100-member stochastic ensemble.

This manuscript details the computational protocol, establishes the baseline calibration of the underlying platform, and defines the statistical analysis plan for retrospective validation on matched longitudinal biopsies. By formalizing how to score compartment-aware predictions against serial molecular measurements—while strictly prohibiting multi-timepoint information leakage—we provide a blueprint for moving digital twins from prognostic risk engines to actionable clinical decision support systems, adhering to emerging standards for AI validation in healthcare.

---

2. Related Work

Our framework sits at the intersection of longitudinal clonal evolution measurement, clonal deconvolution, and mechanistic tumor modeling, guided by rigorous reporting standards for predictive models.

Longitudinal Clonal Evolution Measurement The availability of serial tumor sequencing has transformed our understanding of cancer as an evolutionary process. The TRACERx consortium established the clinical relevance of intratumor heterogeneity in non-small-cell lung cancer (NSCLC), demonstrating that phylogeny-informed tracking can detect relapse and characterize emerging subclones. Similarly, the GLASS consortium provides pooled longitudinal glioma sequencing, mapping trajectories between initial and recurrent disease. While these resources provide the necessary ground truth for evolutionary dynamics, they are primarily observational. Our framework leverages these datasets as a benchmark for *forward prediction*, testing whether a computational model can forecast the trajectories these consortia have documented.

Clonal Deconvolution from Bulk Sequencing Inferring clonal population structure from bulk sequencing is a well-studied inverse problem. Foundational Bayesian clustering methods, such as PyClone and its variational inference successor PyClone-VI, utilize beta-binomial likelihoods to group single nucleotide variants (SNVs) into mutation clusters. It is critical to note that these methods infer the cellular prevalence (cancer cell fraction, CCF) of mutation clusters, which are not strictly disjoint clone mixture weights. Adjacent methods incorporate copy number alterations (CNAs)

and multi-sample constraints to build true phylogenies. However, joint multi-sample deconvolution (using both primary and relapse biopsies) introduces catastrophic information leakage if used to initialize a forward prediction model. Our framework uses PyClone-VI strictly as a single-timepoint (T_1) observation module, explicitly mapping cluster prevalences to simulation compartments while modeling the observation noise to feed a forward-simulating SDE.

Mechanistic Tumor Dynamics and Digital Twins Mathematical oncology has long employed mechanistic models to study evolutionary dynamics. However, these models are rarely initialized directly from patient-specific, high-dimensional multi-omics data. Conversely, modern deep learning survival models handle high-dimensional omics but lack identifiable mechanistic states. The DNAI platform bridges this gap via a dual-paradigm architecture: a Variational Autoencoder (VAE) compresses omics into a biological latent space, which a Hypernetwork uses to emit parameters for a physics-constrained ODE. Previous validations of such systems have relied on bulk metrics. Our contribution is defining a protocol that forces the ODE to be evaluated on its internal compartmental accuracy.

Validation Standards and Probabilistic Forecasting To ensure clinical relevance, this protocol aligns with the TRIPOD+AI and PROBAST-AI guidelines for the transparent reporting and risk-of-bias assessment of AI models in healthcare, as well as the SPIRIT-AI and CONSORT-AI extensions for clinical trial protocols. Furthermore, because deterministic predictions are insufficient for highly stochastic evolutionary processes, we draw on the literature of proper scoring rules—specifically the Brier score and multiclass calibration—for probabilistic forecast evaluation in biomedical settings.

---

3. Methods

3.1 Observation Model and Strict Temporal Isolation

Let V_{ij} denote the variant read count for mutation i in sample j from total depth n_{ij}. To account for sequencing noise and tumor purity, the generative model underlying our deconvolution assumes the variant allele frequency (VAF) follows an overdispersed Beta-Binomial distribution, as formalized in the PyClone-VI framework. The expected VAF is a function of the cancer cell fraction (CCF) of the mutation cluster, adjusted for local copy number (CN) and sample tumor purity.

Operationally, we will use **PyClone-VI v0.1.1** with locked hyperparameters (alpha0=1.0, precision=100, max_clusters=10) to infer the posterior CCF distributions $p(\mathbf{f}_j, \mathbf{z}_j | \mathbf{V}_j)$ via variational inference. Tumor purity and allele-specific copy number profiles will be sourced from cohort-specific harmonized clinical data following a strict precedence hierarchy (Cohort-provided > FACETS > ABSOLUTE).

Crucial Anti-Leakage Constraint: To prevent silent information leakage, clonal deconvolution is strictly restricted to T_1 data for initialization. T_2 data is processed in a completely isolated, sealed pipeline solely for endpoint scoring. Joint multi-sample deconvolution across T_1 and T_2 is formally prohibited.

Leakage Audit: Prior to unblinding, we will conduct a procedural leakage audit. We will attempt to reproduce a known leakage pathway (e.g., passing a T_2-exclusive variant into the T_1 deconvolution manifest) and computationally verify that the pipeline's strict temporal isolation halts execution and flags the violation.

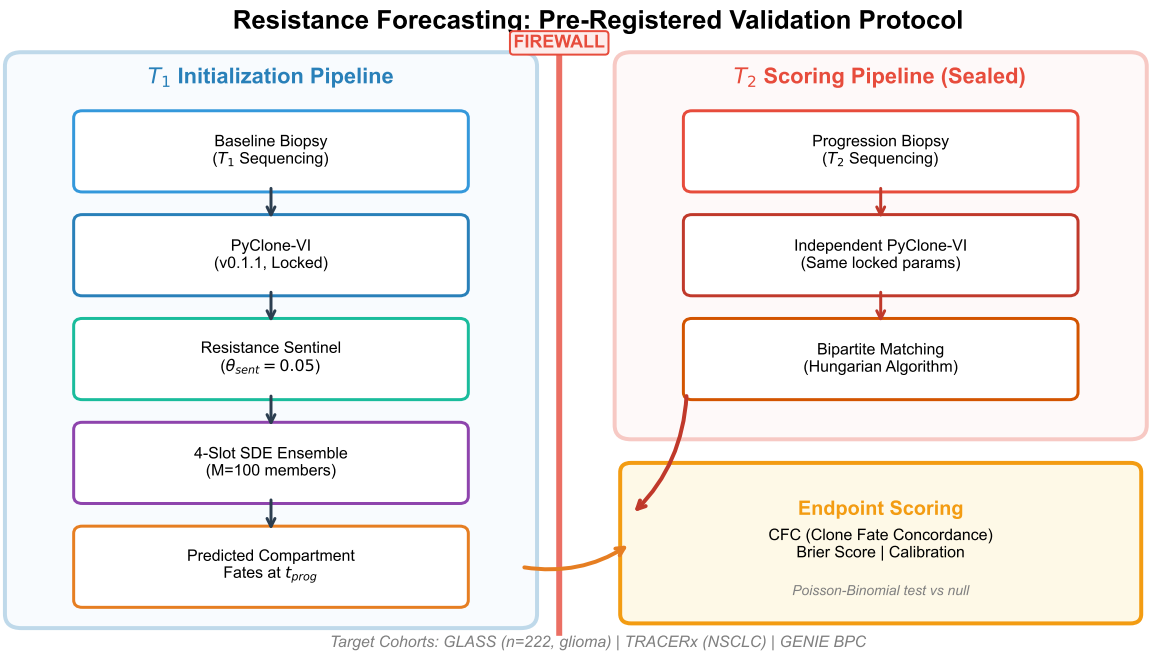

Figure 1: Figure 1: Protocol Schema

Figure 1 (Protocol Schema - Planned): A visual directed acyclic graph demonstrating the strict isolation between the T_1 initialization pipeline (PyClone-VI → Sentinel → SDE Ensemble) and the T_2 scoring pipeline (Independent PyClone-VI → Bipartite Matching → Endpoint Calculation).

3.2 Normalized Prevalence-to-Burden Mapping and the Resistance Sentinel

PyClone-VI outputs cellular prevalences (CCFs) of mutation clusters, which are not strictly compositional (they do not sum to 1, as subclonal clusters are nested within truncal ones). To ensure clinical safety and SDE stability, we map the variable inference cardinality K_{inf} to a fixed 4-slot SDE ($K_{sim} = 4$ compartments) using a normalized prevalence-to-burden mapping.

We define a resistance status indicator $\mathbf{r} \in \{0,1\}^{K_{inf}}$ where $r_k = \mathbb{I}(\exists m \in \mathcal{M}_k : m \in \mathcal{R})$, with $\mathcal{R}$ representing the OncoKB/CIViC resistance mutation set for the administered therapy, locked prior to data access.

We define the Sentinel operator $\mathcal{S} : \mathbb{R}^{K_{inf}} \times \{0,1\}^{K_{inf}} \to \Delta^3$ via the following deterministic procedure:

1. **Raw Weighting:** Define raw cluster weights $w_k = f_k \cdot \psi$, where ψ is tumor purity.
2. **Simplex Normalization:** Normalize to a proper probability vector: $\tilde{N}_0^{(raw)}[k] = \frac{w_k}{\sum_j w_j}$ for $k \in \{1...K_{inf}\}$.
3. **Sentinel Aggregation:** Partition indices into Sensitive (S) where $r_k = 0$, and Resistant (R) where $r_k = 1$. Calculate mass $M_S = \sum_{k \in S} \tilde{N}_0^{(raw)}[k]$ and $M_R = \sum_{k \in R} \tilde{N}_0^{(raw)}[k]$. Sort S descending by $\tilde{N}_0^{(raw)}[k]$ to yield indices $s_1, s_2, ...$
4. **Compartment Assignment:**
   - $N_0[1] = \tilde{N}_0^{(raw)}[s_1]$ (largest sensitive)

- $N_0[2] = \tilde{N}_0^{(raw)}[s_2]$ if $|S| \geq 2$, else 0
- $N_0[4] = \max(\theta_{sent}, M_R)$ if $\max_{k \in R} \tilde{N}_0^{(raw)}[k] \geq \theta_{sent}$, else M_R
- $N_0[3] = 1.0 - (N_0[1] + N_0[2] + N_0[4])$ (bulk remainder, guaranteed non-negative)

This ensures $N_0 \in \Delta^3$ (the 3-simplex), preserving the non-negativity and sum-to-1 constraints required for the SDE state space.

3.3 Protocol-Platform Interface Contract and Forward Model

The DNAI platform utilizes a frozen VAE (v5.10, 328d latent space) and a Conditioned Gating Hypernetwork (Path A v3.2, checkpoint frozen prior to protocol registration). This is a **predictive validation**; the Hypernetwork parameters are strictly locked, and no post-hoc refitting of SDE parameters will occur.

Interface Contract: The Hypernetwork emits *global* base tumor parameters: base growth rate $\rho_{base} \in [0, 0.3]$ day^{-1}, base immune kill $\omega_{base} > 0$, and base drug sensitivities $\beta_{base} \in [0, 1]^D$. To construct the compartment-specific parameters required by the SDE, we define a per-drug resistance mask $\mathcal{M}_{k,d} \in \{0, 1\}$ derived from the Sentinel mapping: * $\rho_k = \rho_{base}$ * $\omega_k = \omega_{base}$ * $\beta_{k,d} = \beta_{base,d} \times (1 - \mathcal{M}_{k,d})$

The deterministic baseline is an extended Lotka-Volterra system. To model evolutionary stochasticity and propagate deconvolution uncertainty, we extend this into a geometric Brownian motion per compartment with diagonal noise:

$$dN_k(t) = \left(\rho_k N_k(t) \left(1 - \frac{\sum_{j=1}^{4} N_j(t)}{K_{cap}} \right) - N_k(t) \sum_{d=1}^{D} \beta_{k,d} C_d(t) - \omega_k E(t) N_k(t) \right) dt + \sigma_{base} \rho_k N_k(t) dW_{k,t}$$

where $W_{k,t}$ are independent standard Wiener processes, $K_{cap} := 1.0$, and $E(t)$ is treated as a constant absorbed into ω_k.

Numerical integration is performed using the Euler-Maruyama method in log-space ($Y_k = \log N_k$). We utilize the log-Euler approximation which absorbs the Itô drift correction ($-\frac{1}{2}\sigma_{base}^2 \rho_k^2$) to strictly enforce non-negativity. Initial conditions $N_k(0)$ are drawn from the PyClone-VI posterior. We will generate a **100-member stochastic ensemble** to yield a compartment fate probability distribution $\hat{p}_k(t)$.

3.4 Operational Definitions and Cross-Timepoint Matching

- **Progression Time (t_{prog}):** Defined as the date of the T_2 biopsy.
- **Treatment Mapping ($C_d(t)$):** Real-world regimens are mapped to $D = 12$ drug classes using the HemOnc.org ontology. $C_d(t) \in [0, 1]$ encodes relative dose intensity. Linear interpolation is used for gaps < 14 days; interruptions > 14 days set $C_d(t) = 0$.
- **WSI Missingness:** WSI is optional. When absent, the platform utilizes a locked learned missing token $\mathbf{m} \in \mathbb{R}^{1536}$ during Hypernetwork gating. Clone fate forecasting relies strictly on molecular deconvolution; WSI informs the global prognostic risk parameters ($\rho_{base}, \omega_{base}$).

Bipartite Matching Algorithm ($T_1 \leftrightarrow T_2$): To score predictions, we must map the observed T_2 mutation clusters back to the T_1-initialized SDE compartments. We pre-register a bipartite matching $\mathcal{M} : \{1..K_{T_2}\} \times \{1..4\} \to [0, 1]$: 1. **Harmonization:** Restrict analysis to the intersection of callable mutations between T_1 and T_2 assays (critical for GENIE BPC targeted panels). If intersection size < 10 mutations, declare indeterminate. 2. **Distance Metric:** Compute Jaccard

distance $J(A,B) = |A \cap B|/|A \cup B|$ on the mutation sets. Define cost matrix $d(k,l) = 1 - J(\mathcal{M}_k^{T_1}, \mathcal{M}_l^{T_2})$. 3. **Linear Assignment:** Solve the assignment problem using the Hungarian algorithm to minimize total cost. 4. **Indeterminate Rule:** If $\min_k d(k,l) > 0.5$ for any T_2 cluster l, flag the patient as indeterminate (indicating new metastatic seeding or profound assay mismatch). 5. **Tie-Breaking:** If multiple T_2 clusters match to one SDE slot, their CCFs are merged for scoring.

3.5 Validation Endpoints

Predictions will be scored against the matched progression/recurrence biopsy (T_2). The "observed dominant compartment" g^* is defined as the SDE slot corresponding to the T_2 cluster with the highest posterior mean CCF.

Primary Endpoint (Clone Fate Concordance, CFC): CFC is defined as the fraction of patients for which the predicted dominant compartment at progression matches the observed dominant compartment at progression.

$$\text{CFC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\arg\max_k \hat{N}_{i,k}(t_{prog}) == g_i^* \right)$$

Co-Primary Endpoint (Coverage): The fraction of enrolled patients with scorable (determinate) outcomes.

Secondary Endpoints: 1. **Expected Dominant Compartment Concordance (Expected DCC):** Evaluated as the Brier score for dominant compartment classification: Expected DCC $= \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{4} \left(\hat{p}_{i,k} - y_{i,k} \right)^2$, where $\hat{p}_{i,k}$ is the ensemble probability that compartment k is dominant, and $y_{i,k}$ is the indicator of observed dominance. 2. **Multiclass Calibration:** Assessed via one-vs-rest calibration curves for each compartment and multinomial reliability diagrams.

3.6 Statistical Analysis Plan and Uncertainty Quantification

Unit of Inference: The patient. **Estimand:** Mean patient-level clone fate concordance (CFC) at T_2, evaluated per cohort.

Hypothesis Testing (Stratified Poisson-Binomial): The primary null hypothesis (H_0^1) is that the model's CFC is less than or equal to a patient-specific frequency-proportional random baseline (Null Model B). If the baseline predictor samples a compartment $\hat{g}_i \sim \text{Categorical}(\mathbf{q}_i)$ where $q_{i,k} = \tilde{N}_{0,i}^{(raw)}[k]$, the probability of correct prediction by chance is exactly the normalized frequency of the *observed* T_2 winner: $p_{0,i} = q_{i,g_i^*}$.

The total number of successes $S_c = \sum_{i \in c} \mathbb{I}(\hat{g}_i = g_i^*)$ within cohort c follows a Poisson-binomial distribution with parameters $\{p_{0,i}\}$. We will compute exact p-values per cohort using the Poisson-binomial CDF. False discovery rate (FDR) across the three cohorts will be controlled via the Benjamini-Hochberg procedure at $\alpha = 0.05$.

Uncertainty Quantification: We will compute 95% confidence intervals for CFC using the exact Clopper-Pearson interval for the point estimate. For stratified estimates, we report $\widehat{\text{CFC}}_c \pm z_{1-\alpha/2} \sqrt{\widehat{\text{CFC}}_c (1 - \widehat{\text{CFC}}_c)/n_c}$.

Power Analysis: Power was estimated via simulated Poisson-binomial draws using empirical T_1 distributions. Assuming a generative model where the baseline random CFC is $p_0 \approx 0.40$, the

GLASS cohort (target $n = 222$) provides $> 90\%$ power to detect an absolute improvement of $\Delta = 0.15$ at $\alpha = 0.05$.

3.7 Sensitivity Analyses and Abstention Policy

1. **Intent-to-Treat (ITT) Attrition Analysis:** To prevent optimistic bias on the "scorable" subset, we pre-register an ITT sensitivity analysis where all indeterminate and abstained cases are counted as CFC failures.
2. **Purity Perturbation:** We will re-run the full scoring pipeline under plausible purity shifts ($\pm 10\%$ absolute purity, bounded $[0.1, 1.0]$).
3. **Abstention Policy:** The DNAI platform utilizes an Information Sufficiency Score (ISS), defined as $1 - H(\hat{\mathbf{p}})/\log(4)$, where H is the Shannon entropy of the ensemble prediction. Predictions where ISS $< \tau_{ISS}$ (locked at 0.6) will trigger an "abstain" state.

---

4. Protocol Overview and Pre-validation Calibration

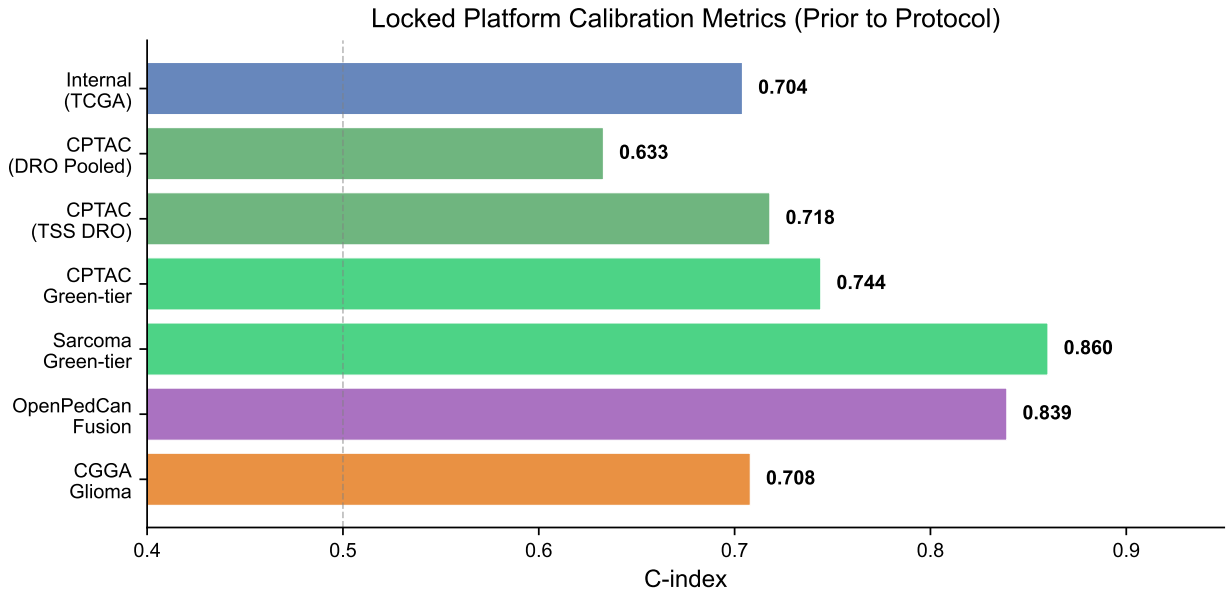

Figure 2: Figure 2: Platform Calibration Metrics

4.1 Background: Existing Locked Platform Calibration

This manuscript is a validation protocol; no cohort validation results are reported here. However, to justify the deployment of the DNAI platform for this protocol, we report its prior locked calibration metrics. These metrics were computed independently of the validation cohorts and represent the baseline state of the prognostic risk engine. Note that survival C-indices establish prognostic validity but do not inherently guarantee clone-fate forecasting validity.

Table 1: Locked Calibration Metrics (Prior to Protocol Execution)

Metric	Value	Cohort/Condition
Internal Hypernetwork C-index	**0.704**	TCGA (locked prior to protocol)
External pooled C-index (DRO)	**0.633**	4-env DRO, CPTAC LOCO, 5/5 pass
External DRO C-index (TSS)	**0.718**	Hospital-level shift robustness
Green-tier C-index	**0.744**	22.2% coverage, highest-confidence patients
Sarcoma Green-tier C-index	**0.860**	255 patients, 85.5% Green coverage
OpenPedCan pediatric fusion C-index	**0.839**	943 brain tumors, RNA + Methylation
CGGA glioma C-index	**0.708**	970 patients, cross-ethnic (Chinese cohort)

4.2 Pre-registered Protocol Locks

To eliminate researcher degrees of freedom, all critical computational parameters are locked prior to data access. All values are cryptographically hash-locked in the pre-registered manifest.

Table 2: Pre-registered Protocol Locks

Element	Locked Value/Method	Justification/Sensitivity
PyClone-VI	v0.1.1, alpha0=1.0, precision=100, max_clusters=10	Sensitivity: \$±\$2 clusters
Mutation filters	VAF $\geq$ 5%, depth $\geq$ 100×, gnomAD AF $<$ 0.1%	Exclude hypermutators >10/Mb
CN/Purity source	Cohort-provided > FACETS > ABSOLUTE	Sensitivity analysis for purity \$±\$10%
Sentinel threshold	$\theta_{sent} = 0.05$	Sensitivity: 0.01, 0.10
ODE params	$K_{cap} = 1.0$	Fixed normalized capacity
SDE ensemble	$M = 100$ members, Euler-Maruyama step $\Delta t = 0.1$ days	Convergence verified at $M = 50$
SDE volatility	$\sigma_{base} = 0.1$	Ensures CV(volatility) $<$ 20% in TCGA
Treatment encoding	HemOnc.org ontology mapped to $D = 12$ drug classes	Mapping table locked in registry
Random baseline	Patient-specific frequency-proportional sampling	Defined in Statistical Plan
Indeterminate rule	Jaccard distance of somatic mutations >0.5	Scorability threshold

4.3 Planned Cohorts and Attrition Flow

We will execute the validation across three independent cohorts. Inclusion and exclusion criteria are strictly pre-specified.

Table 3: GLASS Cohort (Glioma, target $n = 222$)

Criterion	Specification
Inclusion	Adult diffuse glioma (WHO grade II-IV); available baseline (T_1) and recurrence (T_2) WES/WGS; treatment history available; tumor purity $\geq 20\%$.
Exclusion	Hypermutator phenotype (TMZ-induced) if mutation burden >10/Mb; unmatched T_1/T_2 anatomical sites without phylogenetic link; insufficient coverage (<100× for WES, <30× for WGS).

Table 4: TRACERx Cohort (NSCLC)

Criterion	Specification
Inclusion	NSCLC with multi-region baseline sampling (≥ 2 regions); subsequent recurrence/progression biopsy or ctDNA; systemic therapy exposure between T_1 and T_2.
Exclusion	Purely regional relapse without multi-region baseline; synchronous primary tumors.

Table 5: GENIE BPC Cohort

Criterion	Specification
Inclusion	Targeted panel sequencing at baseline and progression; documented treatment line; RECIST-measured progression.
Exclusion	Panel lacks coverage for key resistance mutations; ctDNA-only at T_2 without tissue confirmation.

Planned Attrition Flow Template:

```
Enrolled (e.g., target n=222 GLASS)
  Excluded (n=?): purity <20%, insufficient coverage
  Indeterminate (n=?): new metastatic seeding, assay mismatch
  WSI-absent (n=?): sensitivity stratum
  Scorable Primary Analysis (n=?)
      CFC Success (n=?)
      CFC Failure (n=?)
```

4.4 Comparators and Ablations

To ensure fair comparisons, all baselines are restricted to T_1 data only and are provided the exact same treatment history $C_d(t)$ as the DNAI model. No leakage of T_2 data is permitted during comparator fitting.

Table 6: Baseline Comparators and Ablations

Comparator	Description	Purpose
Random Compartment	Patient-specific frequency-proportional sampling	Tests biological specificity of mapping (H_0^1)
Sentinel-Only	Predicts resistant slot if $M_R > M_S$ at T_1, else largest sensitive	Isolates heuristic value from mechanistic SDE
Supervised Classifier	Multinomial logistic on $\mathbf{f}$ and treatment	Tests non-mechanistic ML baseline
Deconvolution-only	Static persistence of T_1 posterior mean CCFs	Tests added value of forward simulation
Non-mechanistic Extrapolation	Exponential CCF extrapolation fit strictly on T_1 multi-region data	Tests added value of mechanistic ODE
Bulk-only (K=1)	Single compartment, no clonal structure	Tests added value of clonal modeling
No Sentinel	$K = 4$ but $\theta_{sent} = 0$ (no forced preservation)	Tests Sentinel necessity (Ablation)

4.5 Ethics, Governance, and Reproducibility

Ethics and IRB Approval: This protocol utilizes retrospective, de-identified patient data. Access to controlled datasets (e.g., TRACERx via EGA, GENIE BPC via Sage Bionetworks) will be governed by institutional Data Use Agreements (DUAs). Because the study involves secondary analysis of existing de-identified data, it operates under a waiver of informed consent, subject to local Institutional Review Board (IRB) exemption protocols.

Data and Code Availability: To guarantee the integrity of this pre-registered protocol, we implement a strict audit trail: * **Immutable Manifests:** Exact input file manifests per cohort (including clinical metadata and raw sequencing files) will be locked via SHA256 checksums prior to the unblinding of T_2 data. * **Model Registry:** The exact model weights, mapping tables (HemOnc.org to drug classes, OncoKB/CIViC resistance sets), and inference code are locked under a cryptographic hash in a Docker container (`dnai-validation:v2.1-protocol`, SHA256 hash pre-registered on Zenodo/OSF). We pre-register 5 random seeds for the SDE ensemble [42, 123, 456, 789, 101112] to report variance. * **Deviation Log:** Any necessary deviations from this protocol (e.g., due to unforeseen data formatting issues in GENIE BPC) will be recorded in a pre-registered deviation log schema, detailing the rationale and confirming that the change was made while blinded to T_2 outcomes. * **Public vs. Controlled Access:** The core validation framework, scoring scripts, and synthetic test data will be made publicly available on GitHub. Access to the DNAI model weights and patient-derived latent representations will be provided to credentialed researchers under a standard DUA to prevent patient re-identification.

Validation Timeline: * **Month 0:** Formal protocol registration and cryptographic locking of the model container. * **Month 3:** Completion of data access approvals and ingestion of raw cohort data. * **Month 6:** Finalization of the T_1 deconvolution pipeline and freezing of all baseline inputs. * **Month 9:** Unblinding of T_2 data, execution of the scoring pipeline, and manuscript preparation.

---

5. Discussion

The proposed framework addresses a critical vulnerability in the development of oncology digital twins: the reliance on bulk, static validation metrics. By forcing the model to predict which compartment will dominate at progression, we align the computational validation with the clinical reality of acquired resistance.

If the DNAI platform successfully rejects the null hypotheses outlined in this protocol, it will demonstrate that digital twins can significantly outperform clinical heuristics and non-mechanistic extrapolations. The introduction of the Resistance Sentinel is a pragmatic solution to the identifiability limits of bulk sequencing. Purely unsupervised deconvolution often fails to isolate clinically critical but numerically minor subclones. By injecting knowledge-base priors into the initialization phase, we constrain the 4-slot SDE to biologically plausible starting states, significantly reducing degrees of freedom.

Crucially, the inclusion of the Sentinel-Only baseline ensures we test whether the mechanistic SDE adds predictive value beyond simply hard-coding a "the resistant clone wins" heuristic. We have explicitly pre-registered likely failure modes—such as high aneuploidy, extreme copy number complexity, and low tumor purity—to ensure transparent reporting of where the digital twin safely operates and where it abstains.

---

6. Limitations

This protocol and the underlying platform have several inherent limitations. First, bulk sequencing deconvolution is fundamentally non-identifiable under complex copy number alterations and varying tumor purity. The mapping from PyClone-VI mutation clusters to SDE compartments is a necessary mathematical abstraction, not a perfect phylogenetic reconstruction. Second, the fixed 4-slot SDE compartment limit may under-represent highly heterogeneous tumors with complex branching relapse trajectories. Third, setting $\rho_k = \rho_{base}$ across compartments masks intrinsic fitness differences, making the model heavily reliant on the treatment mask to drive evolutionary competition. Fourth, the GENIE BPC cohort presents harmonization challenges due to differing targeted panels and assay versions; our protocol requires minimum metadata and coverage of key resistance mutations, which may induce selection bias. Finally, spatial heterogeneity remains a confounder; a T_2 biopsy may sample a different spatial region than T_1, which we attempt to mitigate by stratifying same-lesion versus distant recurrences and utilizing the bipartite matching algorithm.

---

7. Conclusion

We have established a rigorous, reproducible, and pre-registered protocol for the retrospective validation of clone-aware digital twins using longitudinal tumor sequencing. By combining biologically

disentangled foundation models, knowledge-grounded compartmental mapping, and stochastic Neural ODEs, this framework provides a stringent test of whether mechanistic models can accurately forecast the identity of emerging resistant lineages. Shifting the benchmark from static survival prediction to dynamic compartment forecasting is an essential step toward realizing the promise of counterfactual decision support in precision oncology.

---

8. References

1. Jamal-Hanjani, M. et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
2. Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).
3. GLASS Consortium. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019).
4. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
5. Gillis, A. W. & Roth, A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics* **21**, 571 (2020).
6. Ghaffarizadeh, A. et al. PhysiCell: An open source physics-based cell simulator for 3-D multicellular systems. *PLoS Comput. Biol.* **14**, e1005991 (2018).
7. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–2546 (1982).
8. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
9. Cruz Rivera, S. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020).
10. Liu, X. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–1374 (2020).
11. Wolff, R. F. et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **170**, 51–58 (2019).
12. Moons, K. G. M. et al. PROBAST-AI: A tool to assess the risk of bias and applicability of artificial intelligence-based prediction models. *BMJ* **385**, e078379 (2024).
13. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
14. Hong, Y. On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.* **59**, 41–51 (2013).
15. Kuleshin, V. et al. Accurate uncertainties for deep learning using calibrated regression. *Proc. Mach. Learn. Res.* **80**, 2796–2804 (2018).
16. Chakravarty, D. et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **1**, 1–16 (2017).
17. Griffith, M. et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.* **49**, 170–174 (2017).
18. Warner, J. L. et al. HemOnc: A new standard vocabulary for hematology/oncology regimens. *J. Biomed. Inform.* **96**, 103239 (2019).
19. Kloeden, P. E. & Platen, E. *Numerical Solution of Stochastic Differential Equations.* (Springer, 1992).

20. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. *Adv. Neural Inf. Process. Syst.* **32**, 5052–5062 (2019).