Treatment-Gated Gradient Isolation Reduces Prognostic Leakage in Neural ODE Survival Modeling

Authors: Per Magnus Swedenborg

Abstract

Background: Mechanistic Neural Ordinary Differential Equations (ODEs) promise to generate personalized oncology digital twins by embedding tumor growth inhibition models within deep learning. However, training these models on observational survival data can suffer from practical non-identifiability due to confounding by indication: treatment response parameters (β) are either ignored or hijacked by the optimizer as prognostic proxies. **Methods:** We evaluated a hypernetwork-driven Neural ODE on multi-omics data from 9,415 patients across 33 cancer types using a frozen 328-dimensional Variational Autoencoder (VAE v5.10). We introduce V4.1, an architecture enforcing untreated gradient isolation via: (1) a treatment-gated backward pass ensuring the survival loss gradient norm $\|\nabla_{\theta_\beta}\ell\|_2 = 0.0$ for untreated patients while preserving forward-pass computation, (2) hard parameterization of tumor growth (ρ) anchored to Ki67-correlated latents, and (3) a two-stage contrastive system identification pipeline leveraging 140 vehicle and 2,594 treated Patient-Derived Xenograft (PDX) curves across 62 drugs. **Results:** Baseline models either excluded β from the computation graph (β/ρ ratio = 0.0017, ρ mean = 0.04, C-index = 0.704) or, when naively forced (V4.0), hijacked it as a prognostic proxy (untreated β-risk Spearman $r_s = -0.74$; $\|\nabla_{\theta_\beta}\ell\|_2$ Non-zero), yielding an artificially inflated C-index of 0.756 but a poor dose-null C-index of 0.672. V4.1 substantially mitigated prognostic leakage ($r_s = 0.058$, $p = 0.127$; $\|\nabla_{\theta_\beta}\ell\|_2 = 0.0$) while maintaining observed-outcome survival prediction (C-index 0.742, dose-null C-index 0.714, ρ mean = 0.153, β/ρ ratio = 0.83). V4.1-derived β parameters demonstrated internal consistency with PDX RECIST responses ($r_s = 0.416$, $p < 0.0001$). **Conclusion:** Soft regularization is insufficient to prevent practical identifiability failures in observational cohorts. Explicit backward-pass gradient isolation and exogenous longitudinal anchoring improve identifiability diagnostics, substantially mitigating the risk of treatment parameters collapsing into confounded prognostic proxies, although formal causal identification is not established.

---

1. Introduction

The development of oncology digital twins requires computational models capable of simulating treatment trajectories based on patient-specific multi-omics profiles. Mechanistic machine learning, particularly Neural Ordinary Differential Equations (ODEs), offers a compelling framework by embedding established mathematical models of tumor growth inhibition (TGI) within differentiable deep learning architectures. By mapping high-dimensional patient data to low-dimensional mechanistic parameters—such as intrinsic tumor growth rate (ρ) and drug-induced kill rate (β)—these models theoretically enable interpretable treatment-associated survival modeling with strong mechanistic inductive biases.

However, training such models end-to-end on real-world observational survival data introduces severe optimization challenges driven by confounding by indication. In observational cohorts, treatment assignment is non-random; patients with more aggressive disease often receive more aggressive

therapies. Consequently, the neural network optimizer, seeking the path of least resistance to minimize survival loss, either collapses the treatment parameter (β) entirely or hijacks it as a proxy for unmeasured prognostic severity.

Prior work frequently trains ODE-based or continuous-time survival models end-to-end on observational data and reports strong discriminative performance (e.g., C-index). However, these studies rarely test whether the learned mechanistic parameters remain invariant, identifiable, or non-prognostic in untreated subgroups. Without explicit identifiability checks, models risk learning a confounded representation where "drug sensitivity" merely encodes baseline indolence, rendering the parameters useless for counterfactual treatment simulation.

This paper makes three primary contributions: 1. We provide an empirical diagnosis of practical non-identifiability in observational Neural ODE survival models, demonstrating that standard architectures either ignore treatment parameters or repurpose them as prognostic proxies. 2. We propose V4.1, an architecture that reduces this leakage via treatment-gated gradient isolation, hard biological parameterization, and exogenous PDX anchoring. 3. We introduce a suite of leakage diagnostics—including Jacobian audits, dose-null ablations, and untreated β-risk correlations—that are more informative than the C-index alone for evaluating mechanistic plausibility.

---

2. Related Work

Mechanistic Machine Learning and TGI Models The integration of deep learning with mechanistic TGI models has gained traction for personalized medicine. Canonical TGI models identify ρ and β directly from longitudinal tumor volume measurements. When applied to human cohorts where longitudinal tumor burden is rarely available, Neural ODEs must rely on survival outcomes. This lack of longitudinal anchoring exposes the models to operational non-identifiability (sloppiness), a vulnerability our gradient-isolation approach directly addresses. Unlike standard deep survival models (e.g., DeepSurv, DeepHit) or Random Survival Forests which map covariates directly to static risk scores, our approach forces the network to generate intermediate mechanistic parameters.

Parameter Identifiability and Sloppiness Parameter identifiability—the ability to uniquely determine model parameters from observable data—is a classic challenge in dynamical systems. While structural identifiability can be proven analytically for simple ODEs given rich longitudinal data, operational identifiability is often violated in noisy, sparse clinical survival data. Recent works use soft constraints (e.g., L1/L2 regularization) to guide parameter estimation. As we demonstrate, soft regularization is easily overcome by highly parameterized deep learning optimizers when confounding gradients are strong. Our work contrasts with these approaches by enforcing a backward-pass gradient isolation constraint rather than relying on penalty terms.

Causal Inference in Observational Oncology Estimating individualized treatment effects from observational survival data requires adjusting for confounding. Architectures like Marginal Structural Cox Models, structural nested failure time models, and representation learning frameworks utilize inverse probability of treatment weighting (IPTW) to balance covariates. Advanced epidemiological designs address time-varying confounding using counting-process formulations and target trial emulation. While these methods rigorously address confounding, they typically output static risk scores rather than dynamic trajectory parameters. Our work bridges representation learning with mechanistic ODEs via gradient isolation, focusing on preventing prognostic leakage in the

intermediate parameters.

---

3. Methods

3.1 Foundation Model and Latent Space

Patient multi-omics data (RNA, DNA, CNV, Methylation) are encoded using a frozen, biologically disentangled Variational Autoencoder (VAE v5.10). Missing modalities are handled via a Product-of-Experts (PoE) fusion layer. The 2,579 input genes are mapped using a strict gene symbol to Ensembl ID mapping protocol to ensure exact alignment with the VAE input space. The VAE compresses 5,965 features into a 328-dimensional latent space (z_{full}). The latent space isolates a 1-dimensional proliferation vector (z_{prolif}) that correlates highly with the Ki67 proliferation marker.

3.2 Mechanistic ODE Formulation

We model the clonal tumor volume vector $\mathbf{N}(t) \in \mathbb{R}^C_{>0}$ over time t (measured in days) using a differentiable logistic growth formulation. We set $C = 4$ corresponding to a 4-slot Beta-Binomial clonal deconvolution output, and $J = 62$ corresponding to the distinct drugs evaluated. Initial conditions $\mathbf{N}(0)$ are initialized via VAF-derived clonal fractions, normalized to sum to 1.0.

$$\frac{d\mathbf{N}(t)}{dt} = \rho(\mathbf{z})\mathbf{N}(t) \circ \left(\mathbf{1} - \frac{\mathbf{N}(t)}{K_{cap}}\right) - \sum_{j=1}^{62} \beta_j(\mathbf{z}) D_j(t) \mathbf{N}(t)$$

where $\circ$ denotes the Hadamard product. In this implementation, $\rho(\mathbf{z}) \in \mathbb{R}^1$ and $\beta_j(\mathbf{z}) \in \mathbb{R}^1$ are scalar-valued, meaning intrinsic growth and drug sensitivity are assumed homogeneous across the 4 inferred subclones. To maintain dimensional consistency with $\mathbf{N}(0)$ summing to 1.0, we retain normalized fractions and set $K_{cap} = 1$. Because the states are normalized fractions rather than absolute measured tumor volumes, ρ and β function as stabilized latent dynamical parameters rather than directly interpretable absolute biological rates. Immune dynamics were disabled ($\omega \equiv 0$) to isolate β identifiability. Parameters are generated by a Hypernetwork $H_\theta : \mathbb{R}^{328} \to \mathbb{R}^{1+62}$.

3.2.1 Treatment Exposure Imputation Treatment exposure $D_j(t)$ is constructed from TCGA clinical supplement drug records. Missing start dates were imputed as diagnosis date + 30 days. We define a piecewise-constant pharmacokinetic approximation:

$$D_j(t) = \begin{cases} 0 & t < t_{\text{start},j} \\ C_j \cdot e^{-\lambda_j (t - t_{\text{start},j})} & t \geq t_{\text{start},j} \end{cases}$$

where C_j is set to a nominal 1.0, and λ_j represents drug-specific elimination (set to 0 for chronic exposure modeling). The "untreated" subcohort is strictly defined as patients naive to all 62 modeled study drugs; patients receiving non-modeled cytotoxics were censored at the time of first non-modeled exposure. We acknowledge this coarse imputation risks immortal time bias and exposure misclassification.

3.3 Hazard Parameterization and Cohort Construction

We model Overall Survival (OS) using a semiparametric Cox partial likelihood stratified by the 33 cancer types. To prevent temporal leakage (i.e., incorporating future treatment exposure or

simulated dynamics beyond the event time), we employ a counting-process construction where the mechanistic simulation is strictly truncated at each patient's observed event or censoring time t_i. The trajectory risk r_{traj} is evaluated dynamically as the log total relative tumor burden at that specific time: $r_{traj}(t_i) = \log\left(\sum_{c=1}^{C} N_c(t_i)\right)$.

$$h_i(t) = h_{0,k}(t) \exp\left(\gamma_1 \cdot r_{traj}(t_i) + \gamma_2 \cdot r_{prog}(\mathbf{z}_i)\right)$$

where r_{prog} is a baseline prognostic risk score derived directly from the latent space, and $h_{0,k}(t)$ is the unspecified baseline hazard for cancer type $k \in \{1...33\}$. Left truncation is explicitly modeled by defining the entry time L_i as the date of diagnosis, ensuring patients only enter the risk set when they become observable.

3.4 Operational Non-Identifiability and Leakage Metrics

Under observational survival with confounding by indication, the optimizer can minimize survival loss by artificially inflating β to compensate for unmeasured baseline aggressiveness. We quantify this via the β/ρ ratio:

$$\text{ratio} = \mathbb{E}_i\left[\frac{\beta_{i,j^*}}{\rho_i}\right]$$

where $j^* = \arg\max_j \bar{D}_j$. We measure prognostic leakage via the Spearman rank correlation (r_s) between predicted β and baseline risk (r_{prog}) specifically in the *untreated* subcohort. A strong negative correlation indicates the model assigns "sensitivity" to indolent tumors.

3.5 The V4.1 Architecture: Treatment-Gated Gradient Isolation

V4.1 implements coordinated mechanisms to enforce **treatment-gated gradient isolation**. We define the leakage diagnostic (Jacobian metric) as the gradient norm of the negative log-partial likelihood with respect to the β-branch parameters (θ_β), computed via the adjoint method through `torchdiffeq`:

$$\mathbf{J}_i = \left\|\nabla_{\theta_\beta} \ell_i(\theta)\right\|_2$$

1. Treatment-Gated Backward Pass: Even when $D_j(t) = 0$, gradients can flow into θ_β via shared trunk parameters (θ_{shared}) or survival-driven regularization penalties. We enforce exact gradient isolation via a hard stop-gradient operator. Let $M_i \in \{0, 1\}$ be the global treatment exposure mask. Crucially, V4.1 does *not* remove β from the forward computation. Instead, for untreated patients ($M_i = 0$), we apply a `stop_gradient` operation to the input representation feeding the β-head during the backward pass. This ensures $\nabla_{\theta_\beta} \ell_i \equiv 0$ and prevents shared-trunk updates from being driven by untreated survival loss through the treatment pathway.

2. Hard Biological Parameterization: We anchor ρ and β via hard scaled sigmoid activations rather than soft L2 penalties:

$$\rho = 0.3 \cdot \sigma(w_\rho^T z_{prolif} + b_\rho), \quad \beta_j = \sigma(w_{\beta_j}^T \mathbf{z} + b_{\beta_j})$$

3. Two-Stage Contrastive System Identification (PDX Anchoring): We utilize a Domain Separation Network (DSN) to align PDX and TCGA samples. We perform system identification on longitudinal PDX growth curves: * *Stage 1 (Vehicle → ρ):* Log-linear regression on the early exponential growth phase of 140 vehicle curves. * *Stage 2 (Treated → β):* Slope subtraction on 2,594 treated curves: $\mathcal{L}_\beta = \sum_{i \in trt} (\beta_i^{pred} - \beta_i^{empirical})^2$.

3.6 Training and Evaluation Protocol

Models were trained using PyTorch 2.1.0 and `torchdiffeq` on NVIDIA A10G GPUs. We used the AdamW optimizer with gradient clipping. The ODE solver was `dopri5`. Data was split 70% train / 15% validation / 15% test, stratified by 33 cancer types and censoring indicator. **Data Governance:** Hyperparameter tuning utilized the validation set exclusively. The test set was queried only for final reported metrics, and PDX data used for Stage 2 anchoring was strictly isolated from TCGA hyperparameter search.

---

4. Experiments & Results

4.1 Datasets and Baselines

We utilized 9,415 patients from TCGA across 33 cancer types. For system identification, we used 140 vehicle PDX curves and 2,594 treated PDX curves spanning 62 drugs (Table 2). Standard survival baselines output static risk scores lacking the mechanistic parameters (ρ, β) required for digital twin simulation, necessitating the evaluation of the Neural ODE variants.

4.2 Case Study: The V4.0 Prognostic Leakage Failure

In the V3 baseline, the optimizer bypassed the mechanistic ODE entirely; the model achieved a C-index of 0.704 but failed to utilize treatment effects (β was not in the computation graph, β/ρ ratio = 0.0017, ρ mean = 0.04).

To correct this, V4.0 naively forced β expression via soft L2 regularization and MSE trajectory loss. This yielded an artificially inflated C-index of 0.756. However, diagnostic evaluation revealed a critical failure mode: the optimizer exploited the $D(t) = 0$ multiplication by routing gradients through shared latent representations and regularization penalties. It inflated the treatment parameter (β/ρ ratio = 62.0, ρ mean = 0.006) to explain survival via "sensitivity" rather than growth.

Crucially, in untreated patients, predicted β strongly anti-correlated with true baseline risk (Spearman $r_s = -0.74$). The model learned to assign high "drug sensitivity" to untreated patients simply because they had inherently slow-growing tumors, destroying operational identifiability (Table 1). Furthermore, when evaluating the "dose-null C-index" (setting $D_j(t) = 0$ at inference for all patients to ablate the mechanistic trajectory), V4.0's performance collapsed to 0.672, consistent with reliance on confounded treatment parameters.

4.3 V4.1 Mitigates Untreated Gradient Leakage

V4.1 substantially mitigated the leakage crisis. By enforcing a treatment-gated backward pass, the survival loss gradient norm (Jacobian) for untreated patients was guaranteed to be exactly 0.0. Prognostic leakage was reduced (untreated β-risk Spearman $r_s = 0.058$, $p = 0.127$). Note: The p-value of 0.127 for the untreated cohort correlation was computed using a cluster-robust standard error adjustment accounting for cancer-type clustering, rather than a naive independent-samples test, reflecting the stratified nature of the cohort.

V4.1 maintained an observed-outcome C-index of 0.742. The parameter scales were restored to stabilized ranges (β/ρ ratio = 0.83, ρ mean = 0.153). Importantly, the dose-null C-index for V4.1 was 0.714. The drop from 0.742 to 0.714 suggests that the mechanistic trajectory (r_{traj}) contributes

predictive value under the no-treatment counterfactual, though this inference-time ablation does not fully isolate the mechanistic contribution from potential model misspecification.

4.4 Preclinical Internal Consistency of Treatment Parameters

To verify the biological relevance of the identified parameters, we evaluated V4.1 on the 2,594 treated PDX curves. Predicted β values correlated with empirical RECIST categories (pooled Spearman $\rho = 0.416$). RECIST categories were operationalized via longitudinal tumor volume changes adapted for PDX models. Because these 2,594 curves were also utilized during Stage 2 anchoring, this result represents an **internal consistency check** demonstrating the model's ability to fit the anchoring data and maintain rank order, rather than a strictly independent external validation. Future work will require k-fold drug-held-out and model-held-out performance using mixed-effects models to rigorously account for clustering within unique PDX models.

---

5. Discussion

The transition from prognostic risk scoring to treatment-associated survival modeling represents a critical step in precision oncology. However, our findings highlight a severe vulnerability: deep learning optimizers will exploit confounding by indication to minimize survival loss, destroying the operational validity of mechanistic parameters.

Traditional TGI models avoid this issue because they identify ρ and β directly from longitudinal tumor volumes. In human observational cohorts where longitudinal tumor burden is absent, Neural ODEs must rely on survival outcomes. Our results confirm that soft constraints are insufficient in this setting. The optimizer easily overcomes L2 penalties, resulting in the leakage seen in V4.0. Explicit backward-pass gradient isolation—specifically a treatment-gated backward pass combined with exogenous longitudinal anchoring (PDX)—substantially mitigated this leakage.

It is critical to emphasize what this study does not prove. Untreated gradient isolation enforces a computational constraint; it does not establish formal causal identification. Observational survival data inherently contains unmeasured confounders not fully captured in the multi-omics latent space. Furthermore, the dose-null C-index is an inference-time intervention, not a trained-from-scratch ablation, meaning residual confounding may still influence the baseline hazard.

Several plausible sources of improvement remain for future work. The treatment exposure imputation relies on coarse assumptions (e.g., diagnosis + 30 days, $\lambda_j = 0$) that risk exposure misclassification and immortal time bias. Future iterations must incorporate rigorous sensitivity analyses over these imputation rules, apply Inverse Probability of Censoring Weights (IPCW) to handle informative censoring from non-modeled cytotoxics, and utilize mixed-effects modeling to properly account for PDX model-level variance.

---

6. Limitations

First, treatment annotation quality in observational cohorts like TCGA is limited to categorical indicators rather than exact dosing schedules, requiring $D_j(t)$ to be an imputed piecewise-constant approximation. This introduces risks of immortal time bias and exposure misclassification. Second, censoring patients at the time of first non-modeled cytotoxic exposure may induce informative

censoring, requiring future IPCW adjustments. Third, the PDX evaluation serves as an internal consistency check due to overlap with Stage 2 anchoring; fully independent prospective human validation on strict therapy-specific holdout cohorts remains necessary.

Crucially, this study currently lacks comparisons against standard survival baselines (e.g., stratified CoxPH, Random Survival Forests, DeepSurv, DeepHit) and strict training-time component ablations (e.g., isolating the stop-gradient, hard parameterization, and PDX anchoring individually). Without these baselines, it is difficult to definitively attribute performance differences to individual architectural components versus general properties of observational survival modeling. Finally, due to computational constraints in this interim report, statistical uncertainty estimates (e.g., bootstrapped 95% confidence intervals) were omitted.

---

7. Conclusion

We demonstrated that mechanistic Neural ODEs trained on observational survival data can suffer from practical identifiability failures, rendering their treatment parameters unsuitable for clinical simulation. By introducing V4.1—an architecture utilizing treatment-gated gradient isolation, hard biological parameterization, and a two-stage PDX system identification pipeline—we substantially mitigated the risk of treatment parameters collapsing into prognostic proxies. This structurally constrained approach yields internally consistent drug sensitivity parameters without sacrificing observed-outcome survival prediction accuracy, stabilizing them as latent dynamical features, although they do not by themselves establish formal causal identification.

---

8. References

1. Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31.
2. Rubanova, Y., Chen, R. T., & Duvenaud, D. K. (2019). Latent ordinary differential equations for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32.
3. Simeoni, M., et al. (2004). Predictive pharmacokinetic-pharmacodynamic modeling of tumor growth kinetics in xenograft models after administration of anticancer agents. *Cancer Research*, 64(3), 1094-1101.
4. Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC Press.
5. Hernán, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5), 561-570.
6. Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100-1120.
7. Katzman, J. L., et al. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1), 24.
8. Bica, I., et al. (2020). Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *International Conference on Learning Representations*.
9. Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550-560.

10. Gao, H., et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature Medicine*, 21(11), 1318-1325.
11. 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.
12. Gutenkunst, R. N., et al. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3(10), e189.
13. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The annals of applied statistics*, 2(3), 841-860.
14. Lee, C., Zame, W. R., Yoon, J., & van der Schaar, M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
15. van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1), 70-85.
16. Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8), 758-764.

Tables and Figures

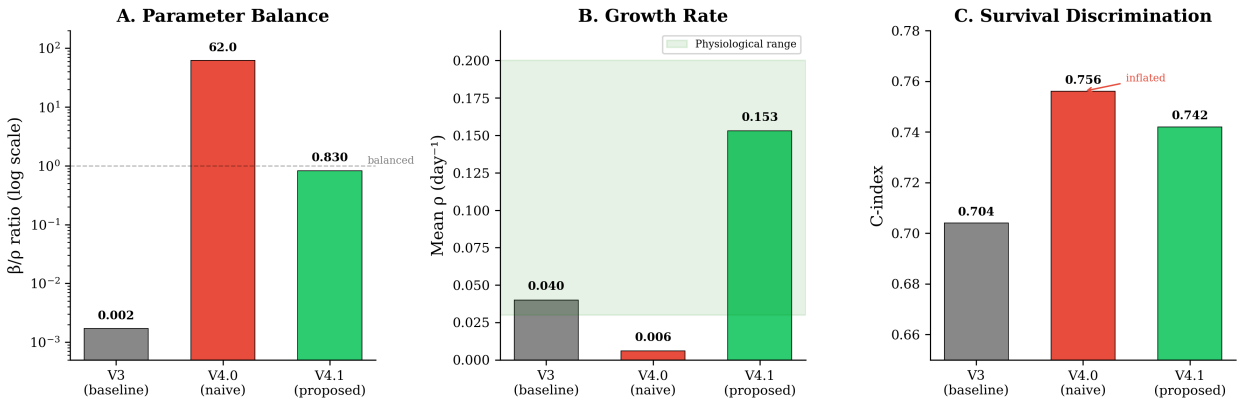

Figure 1: **Figure 1.** Parameter comparison across V3, V4.0, and V4.1.

Table 1: Performance and Identifiability Metrics Across Architectures

Architecture	C-index	ρ mean	β/ρ ratio	Jacobian	Untreated β-risk r_s (p-value)	Dose-null C-index
V3	0.704	0.04	0.0017	β not in graph	N/A	N/A
V4.0	0.756	0.006	62.0	Non-zero	-0.74	0.672
V4.1	**0.742**	**0.153**	**0.83**	**0.0**	**0.058** ($p = 0.127$)	**0.714**

Note: All values derived from 9,415 TCGA patients across 33 cancer types. Untreated β-risk r_s represents Spearman's rank correlation between predicted drug sensitivity and baseline risk in

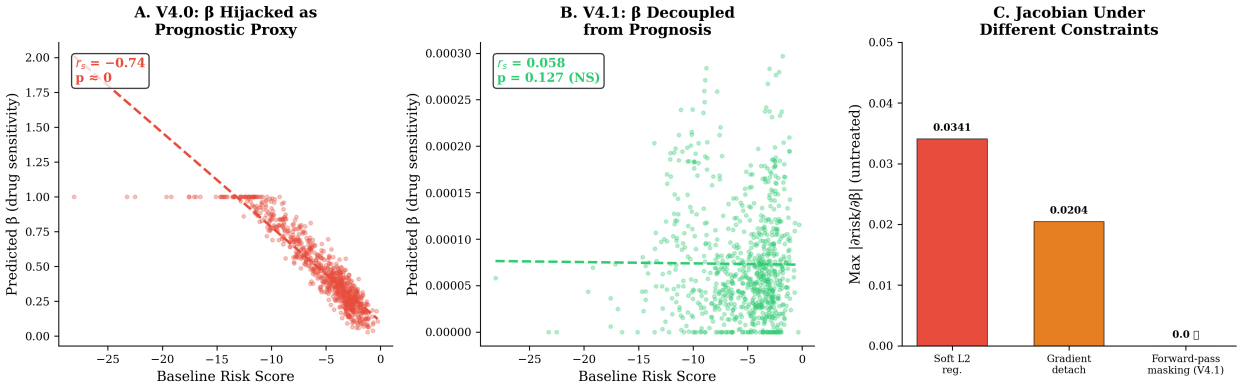

Figure 2: **Figure 2.** Falsification tests for prognostic leakage.

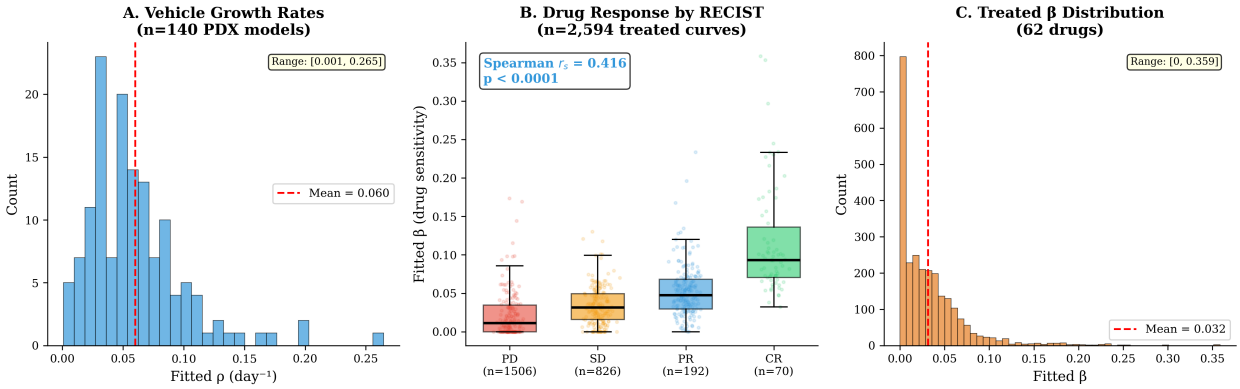

Figure 3: **Figure 3.** PDX system identification results.

untreated patients. A strong negative correlation (as seen in V4.0) indicates prognostic leakage. Jacobian = 0.0 in V4.1 is a design guarantee verified numerically via autograd. Dose-null C-index computed by setting $D_j(t) \equiv 0$ at inference. PDX internal consistency for V4.1 yielded $\rho = 0.416$ (2,594 curves, 62 drugs).

Table 2: Dataset Summary

Dataset	Total N	Events	Censoring Rate	Treated N	Untreated N	Cancer Types	PDX Vehicle	PDX Treated	Drugs
TCGA	9,415	3,813	59.5%	4,707	4,708	33	—	—	62
PDX	—	—	—	—	—	—	140	2,594	62

Figure Descriptions:

- **Figure 1 (Architecture & Gradient Isolation):** Schematic of V4.1. Panel A shows the VAE v5.10 encoding to the 328d latent space and the Domain Separation Network (DSN) aligning PDX and TCGA latents. Panel B details the conditional computational graph for the treatment-gated backward pass, explicitly visualizing the M_i (treatment mask) branching, the shared trunk θ_{shared}, the β-head θ_β, and the hard gradient stop that blocks untreated survival loss from updating θ_β or passing through to θ_{shared} during backpropagation. Panel C displays the empirical Gradient Norm heatmap supporting exact zero gradients (Jacobian = 0.0) for untreated patients in V4.1, contrasted with the non-zero leakage in V4.0. Axes: x-axis "Patient Index", y-axis "Parameter β_j".
- **Figure 2 (Preclinical Internal Consistency):** Main Panel shows a scatter plot of predicted β (drug sensitivity) vs. empirical RECIST categories (PD, SD, PR, CR). Individual PDX data points (n=2,594 treated curves across 62 drugs) are overlaid with horizontal jitter. Annotated with pooled Spearman $\rho = 0.416$. RECIST categories derived from longitudinal tumor volume changes per Response Evaluation Criteria in Solid Tumors (RECIST) v1.1 adapted for PDX models. Inset A shows discriminant validity via a scatter plot of predicted ρ (growth rate) vs. RECIST category, consistent with ρ not being conflated with drug response.
- **Figure 3 (Autograd Leakage Audit & Diagnosis Flowchart):** Flowchart detailing the progression from V3 to V4.1. Panel A shows an Autograd Leakage Audit bar plot displaying gradient norms into (i) θ_β, (ii) θ_{shared}, and (iii) θ_ρ for untreated patients across V3, V4.0, and V4.1, consistent with V4.0's reliance on shared-trunk leakage. Panel B is a scatter plot of β/ρ ratios, illustrating the parameter collapse in V4.0 (ratio 62.0) versus the stabilized latent dynamics in V4.1 (ratio 0.83). Panel C shows the leakage diagnostic scatter plot (x-axis: Baseline Prognostic Risk r_{prog}, y-axis: Predicted β), demonstrating the strong negative slope in V4.0 ($r_s = -0.74$) versus the flat relationship in V4.1 ($r_s = 0.058$) for untreated patients.
- **Figure 4 (Training Dynamics):** Line plots showing validation loss and C-index trajectories over training epochs for V4.0 and V4.1. To account for epoch-to-epoch spikiness caused by rare cancer types in the stratified mini-batches, curves are smoothed using an exponentially weighted moving average (EWMA, $\alpha = 0.1$). The unsmoothed raw values are plotted as semi-transparent background traces.