The Sim-to-Real Scaling Paradox: Biological Heterogeneity Reverses Transfer Learning Gains in Cancer Digital Twins

Per Magnus Swedenborg[1],

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

Abstract

Background: Digital twins for precision oncology require learning complex temporal dynamics from sparse human data. While preclinical models like Patient-Derived Xenografts (PDX) offer abundant temporal data, they introduce severe domain shifts. A common assumption in machine learning is that scaling the diversity of the source domain improves transfer performance.

Methods: We investigated this scaling hypothesis in a physics-constrained neuro-symbolic digital twin (DNAI v3.1). We compared transfer performance from a homogeneous source cohort ($N = 128$ prostate PDX) versus a heterogeneous scaled cohort ($N = 573$ PDX across 23 cancer types) to a target human cohort ($N = 9{,}415$ TCGA). We employed a Validated Domain Separation Network (DSN) for latent alignment and a Neural ODE hypernetwork for downstream survival forecasting.

Results: Contrary to naive scaling expectations, we observed **heterogeneity-induced negative transfer**: expanding the source dataset degraded the downstream human Stratified C-index from 0.654 to 0.640 (1.4 percentage points). We demonstrate that this degradation is associated with the domain discriminator exploiting cancer-type prevalence imbalances as a proxy for domain labels (accuracy 0.832), rather than learning true invariant features. While stratified sampling and Conditional Domain Adversarial Networks (CDAN) recovered latent dimensionality (Effective Rank 201/201) and biological utility (subtype accuracy 0.919), they failed to recover downstream survival performance. Furthermore, we identify **Frozen Batch Normalization** as a critical stabilizer for joint sim-to-real training, preventing gate collapse and recovering optimal performance (Global C-index 0.691).

Conclusion: In biological transfer learning, source domain heterogeneity can be detrimental if it introduces label-distribution shifts that the discriminator exploits. We provide evidence that domain purity and rigorous normalization freezing often outweigh raw data scale in sim-to-real oncology pipelines.

1. Introduction

The development of medical digital twinscomputational models capable of forecasting individual patient trajectoriesis hindered by a fundamental data scarcity. While static genomic profiles are abundant (e.g., TCGA), longitudinal time-series data capturing tumor dynamics under treatment are rare in humans [1]. Conversely, preclinical models like Patient-Derived Xenografts (PDX) provide dense, controlled temporal data but lack fidelity to human biology due to species-specific confounding (murine stroma) and immune deficiencies [2].

Transfer learning, specifically Domain Adaptation (DA), offers a theoretical bridge. By aligning the latent representations of preclinical (sim) and clinical (real) data, one aims to transfer the laws of tumor dynamics learned in mice to human patients [3]. A common expectation in machine learning is that increasing the volume and diversity of source domain data should monotonically improve target domain performance [4].

In this work, we report a failure of this scaling hypothesis in the context of cancer digital twins. We observe that expanding the preclinical source dataset from a small, homogeneous cohort ($N = 128$) to a larger, heterogeneous cohort ($N = 573$) degrades the models ability to rank patients *within* cancer types. We term this phenomenon **heterogeneity-induced negative transfer**. Understanding this failure mode is clinically important: as PDX-derived digital twins move toward deployment for treatment selection and virtual trials, naive data scaling strategies could silently degrade the clinical utility of patient-specific predictions.

We demonstrate that this failure stems from the interaction between biological heterogeneity and adversarial alignment. Standard Domain Adversarial Neural Networks (DANN) [5] utilize a marginal discriminator to enforce feature invariance. In a multi-cancer setting, we show that this discriminator learns to distinguish domains by exploiting cancer-type prevalence shifts (e.g., associating specific squamous histology exclusively with human samples due to sampling bias), rather than removing species-specific artifacts. This results in **negative transfer** [12], where the alignment process strips biologically relevant signals.

Our contributions are: 1. **Quantification of Negative Transfer:** We show that scaling PDX data from 128 to 573 samples degrades downstream Stratified C-index from 0.654 to 0.640 (1.4 pp), despite improved DSN diagnostics with stratified sampling. 2. **Diagnostic Decomposition:** We characterize three failure modes latent collapse (Effective Rank 3), prevalence-driven domain discrimination (accuracy > 0.80), and downstream survival degradation showing that fixing upstream DSN metrics does not guarantee downstream improvement. 3. **Stabilization via FrozenBN:** We demonstrate that **Frozen Batch Normalization** is a critical stabilizer for joint sim-to-real training, preventing gate collapse (from 19:81 to 22:78 omics:WSI) during multitask fine-tuning. 4. **Negative result for trajectory supervision:** Across 4 joint training variants, we confirm that PDX trajectory loss does not meaningfully improve human survival prediction, establishing a principled negative result for this specific transfer task.

---

2. Related Work

Sim-to-Real Transfer: In robotics, domain randomization bridges the reality gap [7]. In biomedicine, the simulation (PDX) is a biological approximation. Previous work has focused on batch correction (ComBat) or statistical alignment (CORAL) [8]. However, these methods often fail to preserve the non-linear dynamics required for trajectory forecasting. Guan & Liu (2021) highlighted similar domain shifts in medical imaging, where site-specific artifacts confound biological signals [13].

Domain Adversarial Learning & Negative Transfer: Ganin et al. [5] introduced DANN to learn domain-invariant features. However, Wang et al. (2019) characterized **negative transfer**, showing that when source and target label distributions are disjoint, forced alignment harms performance [12]. Ben-David et al. [16] provided the theoretical bounds for this, showing that adaptation error is lower-bounded by the divergence of the labeling functions. Our work provides empirical confirmation of these theories in a high-dimensional biological setting, specifically identifying *mixture heterogeneity* as the driver of label shift.

Cancer Digital Twins: Existing approaches largely rely on pure mechanistic models (ODEs) calibrated to individuals [9] or pure deep learning (DeepSurv) [10]. The DNAI platform [11] integrates these via a neuro-symbolic architecture. This study specifically addresses the data-scaling properties of such hybrid systems.

---

3. Methods

3.1 The DNAI Architecture

The system consists of three coupled components: 1. **Foundation Model (VAE v5.10):** A hierarchical Variational Autoencoder pre-trained on 9,415 TCGA samples. It encodes multi-omics data (RNA, DNA, CNV, Methylation) into a 328-dimensional biologically disentangled latent space. 2. **Domain Separation Network (DSN):** A transformation module that maps the VAE latent space (z_{rna}, 201d) to a domain-invariant subspace (z_{shared}), removing species-specific signals ($z_{private}$). 3. **Hypernetwork & Neural ODE:** A hypernetwork takes z_{shared} (plus imputed methylation/CNV, total 281d) and Whole Slide Image (WSI) embeddings to predict parameters (, ,) for a pharmacodynamic Neural ODE governing tumor evolution.

3.2 Domain Separation Network (DSN)

To enable transfer, we employ a DSN that decomposes the input latent into shared and private components. All DSN results reported utilize the `ValidatedDSN` architecture implemented in `models/VAE/scripts/train_dsn_validated.py`.

Architecture: The shared encoder is a 3-layer MLP (201 128 128 201) with LayerNorm and ReLU activations. The private encoder is a high-capacity sponge (201 256 256 256 128). The composite loss minimizes:

$$L_{total} = L_{recon} + 5L_{coral} + 10L_{mean} + 2L_{mmd} + {}_{grl}(L_{adv} + 2L_{linear}) + L_{impute} + {}_{kl}L_{kl}$$

where ${}_{grl}$ ramps from 0 to 2.0 over 30 epochs via gradient reversal. The reconstruction loss enforces information preservation; CORAL, mean-matching, and MMD align domain statistics; adversarial losses confuse domain discriminators; and the imputation loss trains a conditional prior to predict methylation (48d) and CNV (32d) from shared representations.

Conditional Domain Discriminator (CDAN): To address heterogeneity, we implement a conditional discriminator (`ConditionalDomainDiscriminator`). Unlike standard CDAN [6] which uses multilinear conditioning, we use element-wise modulation via learned class embeddings to handle high dimensionality. Let y_{type} be the cancer type index:

$$c = \mathrm{softmax}(W_c \;\; E_{emb}(y_{type}))$$

$$z_{conditioned} = z_{shared} \;\; c$$

This forces the discriminator to align distributions *within* the context of the cancer type.

3.3 Evaluation Protocol

Data Cohorts: * **Target (Human):** TCGA Pan-Cancer (N = 9,415 total; split into training N = 7,056 and validation N = 2,337). * **Source A (Homogeneous):** PDX Prostate (N = 128, Champions Oncology/GEO GSE184427). * **Source B (Heterogeneous):** PDX Multi-Cancer ($N = 573$, 23 types, including Source A).

Evaluation Metrics: * **Stratified C-index:** Concordance index calculated only between pairs of patients within the same cancer type, then averaged across types. This is our primary utility metric. * **Global C-index:** Concordance across all pairs, ignoring cancer type. * **Effective Dimensionality (Eff Dims):** Defined as the number of Principal Components required to explain 99% of the variance in z_{shared}. A value near 0 indicates collapse to a mean vector. * **Subtype Accuracy:** Classification accuracy of a linear probe trained on z_{shared} to predict the 33 TCGA cancer types (biological utility). * **Domain Accuracy:** Accuracy of the domain discriminator (Target: < 0.60 for successful confusion).

Pre-specified Validation Criterion: We define a successful DSN training run as one achieving Domain Accuracy < 0.60 (indicating confusion) while maintaining Subtype Accuracy > 0.70 (indicating biological preservation).

3.4 Joint Training & FrozenBN

During joint training, the Hypernetwork (initialized on human data) processes PDX batches to minimize trajectory error. We observed that Batch Normalization (BN) statistics drifted rapidly when exposed to PDX data. **Correction:** We implemented `FrozenBN`, where BN layers are locked in evaluation mode (using pre-computed running stats from the standalone human training) during the joint fine-tuning phase. This applies to all BN layers in the Hypernetwork.

---

4. Experiments & Results

4.1 The Scaling Failure

We evaluated the impact of scaling PDX data on downstream human survival prediction. All DSN variants were trained, then used as the encoder for a downstream DSN-Hypernetwork (Path B) evaluated on the TCGA validation set ($N = 2{,}337$).

Table 1: Impact of PDX Scaling on Human Survival Prediction (TCGA Validation Set)

DSN Configuration	PDX N	DSN Domain Acc	Eff Dims	Subtype Acc	Downstream Strat C	Downstream Global C
128-PDX Baseline	128	0.584	0	0.800	**0.654**	**0.687**
573-PDX (Marginal)	573	0.832	3	0.746	0.640	0.678
573-PDX (CDAN + Strat)	573	0.807	**201**	**0.919**	0.641	**0.698**

Analysis: The 128-PDX baseline (single cancer type, prostate) achieves the best downstream Stratified C-index (0.654). Scaling to 573 multi-cancer PDX using a marginal discriminator results in high domain accuracy (0.832), indicating a failure to align domains, and degrades Stratified C to 0.640. Even with CDAN conditioning and stratified sampling, domain accuracy remains high (0.807), and while Global C improves (0.698), the within-type ranking (Stratified C 0.641) does not recover.

4.2 DSN Diagnostic Decomposition

To understand why the 573-PDX DSN fails downstream despite improved metrics with stratified sampling, we examine three diagnostic dimensions.

Table 2: DSN Diagnostic Decomposition

Diagnostic	128-PDX	573-PDX (Marginal)	573-PDX (CDAN+Strat)	Interpretation
Domain Accuracy	0.584	0.832	0.807	573 fails Validation Criterion (< 0.60)
Eff Dims (99% var)	0/201	3/201	**201/201**	Stratified fixes collapse
Subtype Acc	0.800	0.746	**0.919**	Stratified recovers biology
Meth Imputation r	0.862	0.790	0.837	Imputation degrades at scale
CNV Imputation r	0.967	0.825	0.926	Imputation degrades at scale

The 573-PDX (Marginal) variant collapses the latent space to 3 effective dimensions (Effective Rank 3) while achieving 0.832 domain accuracy. This suggests the discriminator exploits cancer-type prevalence as a shortcut. Stratified sampling fully recovers dimensionality (201/201) and biological utility (subtype accuracy 0.919), but the downstream survival predictor (Hypernetwork) cannot leverage these improvements. The 128-PDX baselines 0 effective dimensions (indicating a spectrum dominated by a mean vector with very low variance components) surprisingly does not impair survival prediction, suggesting the Hypernetwork extracts survival-relevant features from this highly compressed representation.

4.3 Mechanism: Prevalence-Driven Domain Discrimination

The 573-PDX pool spans 23 cancer types with highly imbalanced representation (e.g., PRAD constitutes 22.3% of PDX samples but only ~5.3% of TCGA). The marginal domain discriminator achieves 0.832 accuracy on the 573-PDX data far above the 0.60 threshold required for domain confusion. Even with CDAN conditioning and stratified sampling (balancing cancer types in the batch), domain accuracy remains at 0.807.

This implies that multi-cancer prevalence differences provide a strong domain-identification shortcut. The discriminator can distinguish domains primarily from cancer-type composition rather than species-specific stroma signal. The result is **semantic misalignment**: the shared encoder suppresses cancer-type-correlated biological signal to confuse the discriminator. While stratified balanced sampling recovers dimensionality, the discriminators domain accuracy remains too high (0.807), indicating that even cancer-type-conditioned discrimination finds exploitable domain proxies in the multi-cancer setting.

4.4 Stabilization via FrozenBN

Finally, we evaluated the `FrozenBN` technique during joint training (simultaneous survival + trajectory optimization). We tested 5 variants to isolate the effect.

Table 3: Joint Training Variants (128-PDX DSN, Warm-Started Hypernetwork)

Variant	FrozenBN	Gate (Omics:WSI)	Global C	Stratified C
Path B Baseline (no joint)	N/A	21 : 79	0.687	**0.654**
Joint v2 (warm-start)	No	19 : 81	0.685	0.642
Joint v3 (LayerNorm fix)	No	19 : 81	0.665	0.653
Joint v4 (FrozenBN)	**Yes**	**22 : 78**	**0.691**	**0.654**
Joint v5 (Multi-curve aug)	Yes	31 : 69	0.679	0.651

Without FrozenBN, the model shifts toward WSI dominance (gate 19:81), and Stratified C degrades (1.2 pp). The LayerNorm variant (v3) partially recovers Stratified C (0.653) but degrades Global C (0.665). Only FrozenBN (v4) simultaneously recovers the baseline Stratified C-index (0.654) and improves Global C-index (0.691, +0.4 pp). Joint v5, which added multi-curve augmentation, improved the gate balance (31:69) but failed to improve survival metrics, reinforcing the negative result for trajectory supervision.

---

5. Discussion

Our results challenge the assumption that more data is better in biological transfer learning. We identify **heterogeneity-induced negative transfer** as a critical failure mode where the domain discriminator aligns prevalence distributions rather than biological features.

Disconnect between DSN diagnostics and downstream performance: A key finding is that improving DSN-level metrics (effective dimensionality, subtype accuracy) does not guarantee downstream survival prediction improvement. The 573-PDX (CDAN+Strat) variant recovers full dimensionality (201/201) and superior subtype accuracy (0.919) but yields lower Stratified C-index (0.641) than the 128-PDX baseline (0.654). This suggests the Hypernetwork learns compensatory representations that work with rather than despite the compressed DSN output of the baseline.

Recommendations for Practitioners: Based on these findings, we propose three heuristics for sim-to-real transfer in oncology:

1. **Diagnosis (The Prevalence Test):** Before training, train a simple classifier $P(Domain|Label)$. If this accuracy is high (>60%), standard adversarial alignment will likely fail due to the prevalence shortcut.
2. **Control (Homogeneous Alignment):** Do not pool multi-cancer PDX data naively. Train separate DSN adapters for each cancer type, or strictly balance the source batches to match target prevalence.
3. **Stabilization (FrozenBN):** When fine-tuning a human-initialized model on preclinical data, **freeze Batch Normalization layers**. Allowing BN statistics to update on the source domain causes drift in the shared feature space, as evidenced by our gate collapse results.

---

6. Limitations

1. **Immunological Gap:** PDX models are immunodeficient. Our model sets $= 0$ for PDX, limiting transfer of immune dynamics.
2. **Endpoint Mismatch:** Joint training optimizes PDX trajectory fidelity (tumor volume) but evaluates human survival (OS). The correlation between these objectives is imperfect.
3. **Single Architecture:** All results use a single DSN architecture (ValidatedDSN).
4. **Single Seed:** Each training variant was run with a single random seed. While the effect sizes are consistent across variants, multi-seed evaluation would strengthen the statistical claims.
5. **Limited PDX Diversity:** The 128-PDX baseline is prostate-only; the 573-PDX spans 23 cancer types but with imbalanced representation.
6. **Allometric Scaling:** PDX-to-human trajectory transfer relies on a fixed allometric time dilation ($= 0.29$) and volume normalization, which may not generalize across cancer types.

---

7. Conclusion

We identified a structural failure mode in cancer digital twins where scaling source data heterogeneity degrades within-cancer target performance, even when DSN-level diagnostics improve. The 128-PDX homogeneous baseline consistently outperforms 573-PDX multi-cancer variants on Stratified C-index (0.654 vs 0.6400.641), while FrozenBN during joint training recovers the optimal combined performance (Global C 0.691, Stratified C 0.654). Across 4 joint training variants, we further establish that PDX trajectory loss does not meaningfully improve human survival prediction. These findings suggest that in biological transfer learning, domain purity and downstream-anchored training often outweigh source data scale and upstream diagnostic quality. For clinical deployment of PDX-to-human digital twins, practitioners should prioritize disease-matched preclinical cohorts over pooled multi-cancer datasets, and freeze batch normalization statistics when fine-tuning human-trained models on preclinical data.

---

8. Resource Availability

Lead Contact Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Per Magnus Swedenborg (per.swedenborg@dnai.bio).

Materials Availability This study did not generate new unique reagents.

Data and Code Availability * **Data:** Human cancer data is available from the TCGA Research Network (https://portal.gdc.cancer.gov). PDX RNA-seq data is available from GEO (Accession GSE184427) and the Novartis PDXE portal (accessible via academic license). * **Code:** Code and model weights are available from the corresponding author upon reasonable request.

---

9. Declarations

Ethics Statement This study exclusively utilized de-identified, publicly available retrospective datasets (TCGA, PDX via GEO GSE184427, Novartis PDXE). All datasets were accessed in accordance with their respective data use agreements. No patient contact, intervention, or collection of new human biological material was performed. As all data were previously collected, de-identified, and publicly released under institutional review, additional IRB approval was not required per the Common Rule (45 CFR 46.104(d)(4)).

Author Contributions (CRediT) P.M.S.: Conceptualization, Methodology, Software, Validation, Formal Analysis, Writing Original Draft, Visualization.

Competing Interests P.M.S. is the founder of DNAI Biotech and has a financial interest in the commercialization of the DNAI platform. P.M.S. is the inventor on provisional patent applications related to the DNAI platform, including U.S. Provisional Application Nos. 63/967,576 (Physics-Constrained Sim-to-Real Transfer Learning), 63/974,083 (Preventing Metabolic Scaling-Induced Collapse), and 63/974,099 (Uncertainty-Calibrated Missing Modality Imputation).

---

10. References

1. Yuldashev, E., et al. (2024). Data scarcity in precision oncology. *Nature Medicine.*
2. Byrne, A.T., et al. (2017). Interrogating open issues in cancer precision medicine with patient-derived xenografts. *Nature Reviews Cancer.*
3. Weiss, K., et al. (2016). A survey of transfer learning. *Journal of Big Data.*
4. Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv.*
5. Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *ICML.*
6. Long, M., et al. (2018). Conditional adversarial domain adaptation. *NeurIPS.*
7. Tobin, J., et al. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. *IROS.*
8. Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. *ECCV.*
9. Brady, R., & Enderling, H. (2019). Mathematical models of cancer: when to predict and how to predict. *Bulletin of Mathematical Biology.*
10. Katzman, J.L., et al. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology.*
11. Swedenborg, P.M. (2025). DNAI: Digital Twin Platform for Precision Oncology. *Internal Technical Report v3.1.*
12. Wang, Z., et al. (2019). Characterizing and avoiding negative transfer. *CVPR.*
13. Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering.*

14. Zhao, H., et al. (2018). Adversarial multiple source domain adaptation. *NeurIPS*.
15. Bousmalis, K., et al. (2016). Domain Separation Networks. *NeurIPS*.
16. Ben-David, S., et al. (2010). A theory of learning from different domains. *Machine Learning*.