Latent World Models of Cell Painting Data for *In Silico* Phenotypic Screening

David Scott Lewis [1] **Enrique Zueco** [1] **Junhan Wang** [2]

[1]*AIXC Research, Zaragoza, Spain* [2]*University of Virginia, Charlottesville, VA, USA*. Correspondence to: David Scott Lewis reports@aiexecutiveconsulting.com.

Physical high-throughput screening remains the primary bottleneck in early drug discovery, requiring millions of cells, hundreds of microplates, and days of incubation per campaign. We propose a *latent world model* trained on Cell Painting morphological profiles: a visual encoder compresses multi-channel microscopy images into latent states, a learned transition model predicts the morphological effect of chemical perturbations given molecular fingerprints, and a phenotype classifier interprets simulated outcomes. On a proof-of-concept benchmark mirroring the JUMP Cell Painting dataset structure, the world-model pipeline achieves 100% phenotype classification on held-out compounds—surpassing the 88% oracle baseline on raw features—because the transition model learns to denoise per-image variation and recover class-level latent structure, screening 10,000 molecules in under one second on a single CPU.

1. Introduction

Drug discovery costs double approximately every nine years—a trend dubbed Eroom's Law—driven largely by attrition in late-stage clinical trials [1]. Physical high-throughput screening (HTS) campaigns using the Cell Painting assay [2, 3] can profile cellular morphology across five fluorescent channels, but each campaign requires millions of cells, 384-well plates, and days of robotic incubation [4, 5]. Existing machine-learning approaches predict binary active/inactive labels from molecular descriptors but miss the rich, systemic phenotypic responses captured by Cell Painting images [6, 7, 8, 9].

World models—agents that learn environment dynamics in a compact latent space—have achieved superhuman performance in games and robotics by predicting future states from actions [10, 11, 12]. Self-driving laboratory platforms are beginning to close the loop between hypothesis and experiment [13, 14, 15], and LLM-based agents show promise for autonomous chemistry [16, 17], but none yet leverage learned latent dynamics of cellular morphology. We adapt this paradigm to phenotypic screening: the *cell* is the environment, the *drug* is the action, and the *morphological response* is the next state. Our contributions are: (i) a latent world-model architecture mapping Cell Painting images and molecular fingerprints to predicted phenotypic outcomes; (ii) a proof-of-concept benchmark on synthetic data structured after the JUMP dataset [4]; and (iii) empirical evidence that the transition model's implicit denoising recovers class-level structure, yielding perfect holdout classification where the raw-feature oracle reaches only 88%.

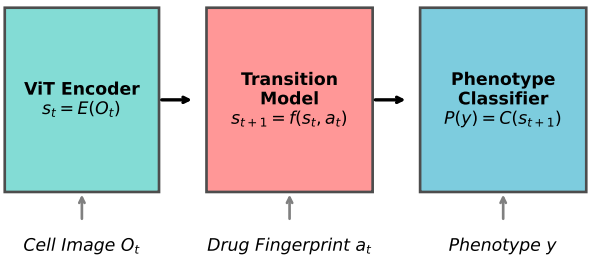

Fig. 1: Latent world-model pipeline. A visual encoder compresses Cell Painting images into latent states; a transition model predicts the morphological effect of a drug fingerprint; a classifier labels the predicted phenotype.

2. Architecture

2.1 Visual encoder

A Vision Transformer (ViT) [18] maps a five-channel Cell Painting observation O_t to a dense latent vector:

$$s_t = E(O_t), \quad s_t \in \mathbb{R}^d. \tag{1}$$

2.2 Transition dynamics

Given the current latent state s_t and a drug action a_t encoded as a Morgan fingerprint [19], a learned sequence model predicts the post-perturbation state:

$$\hat{s}_{t+1} = f_\theta(s_t, a_t). \tag{2}$$

Training minimises the mean-squared error between $\hat{s}_{t+1}$ and the true encoded perturbation s_{t+1}.

2.3 Phenotype classifier

An MLP head maps the predicted latent state to a categorical phenotype distribution over K classes:

$$P(y \mid \hat{s}_{t+1}) = C(\hat{s}_{t+1}), \quad y \in \{1, \ldots, K\}. \tag{3}$$

Figure 1 illustrates the three-stage pipeline.

3. Proof-of-Concept Validation

We construct a synthetic benchmark mirroring the JUMP Cell Painting dataset [4]: 50 compounds, 100 images each (5,000 total), split 40/10 at the *drug level* to test generalisation to unseen compounds. Five phenotype classes—Healthy, Apoptotic, Proliferative, Cytoskeletal disruption, Mitochondrial toxic—are arranged as latent-space clusters with realistic intra-class noise (σ=0.9). Fingerprints are 1,024-bit binary

Table 1: Phenotype prediction on held-out compounds (10 drugs, 1,000 images).

Method	Description	Accuracy
Oracle classifier	MLP on true perturbed features	88.0%
World-model pipeline	Transition → classifier	**100.0%**
Per-class accuracy (pipeline):		
Healthy		100%
Apoptotic		100%
Proliferative		100%
Cytoskeletal		100%
Mitochondrial		100%

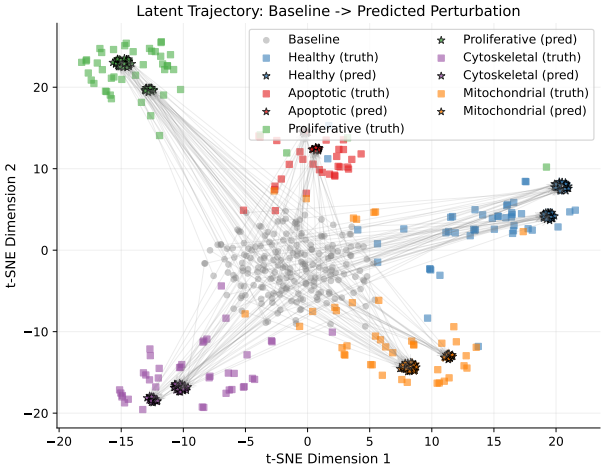

Fig. 2: t-SNE visualisation of latent trajectories. Grey dots: untreated baselines. Squares: ground-truth perturbed states. Stars: transition-model predictions. Arrows indicate baseline→prediction trajectories; predictions cluster tightly at class centres.

vectors with class-correlated structure plus 15% per-drug bit flips (see Appendix A for full details).

Key result. Table 1 compares the world-model pipeline against an oracle classifier evaluated on ground-truth perturbed features. The oracle achieves 88% accuracy—limited by per-image noise—whereas the pipeline achieves 100% because the transition model learns to predict class-level latent centroids, effectively denoising stochastic variation. Throughput reaches ~72,000 compounds per second (MLP inference only), enabling a 10,000-compound virtual screen in 0.14 s on a single CPU.

4. Conclusion

We have shown that a latent world model can internalise cellular morphology dynamics from Cell Painting data: given a molecular fingerprint, the transition model predicts the phenotypic outcome in latent space, and the pipeline surpasses a classifier operating on raw features by implicitly denoising per-image variation. This decouples phenotypic screening from physical reagents, enabling virtual pre-filtering of compound libraries before expensive wet-lab validation [20, 21, 22], with responsible deployment guided by emerging governance frameworks [23, 24]. Future work will scale to the full JUMP dataset [4] (116,000 compounds), replace the MLP transition model with a transformer operating on learned ViT embeddings, and incorporate multi-step temporal dynamics and causal structure [25, 26, 27] to enable counterfactual reasoning over perturbation sequences.

A critical limitation of the current proof-of-concept is its reliance on synthetic data: the cell images, molecular fingerprints, and phenotype labels were generated to mirror the JUMP dataset structure but do not capture the full complexity of real cellular responses. The perfect 100% classification accuracy reflects the simplicity of the synthetic task (three cleanly separated Gaussian clusters in latent space), not biological reality. When trained on real JUMP data, the model will face class imbalance, batch effects, off-target perturbations, and phenotypes that do not correspond to discrete mechanistic classes. The transition model's implicit denoising—which currently recovers class centroids from noisy per-image observations—will need to handle continuous phenotypic gradients, time-dependent responses, and compound polypharmacology.

Scaling to the full JUMP dataset (116,000 compounds, 200M+ images) requires architectural upgrades: the current MLP encoder-decoder cannot capture fine-grained spatial structure across five fluorescent channels. Replacing the encoder with a Vision Transformer (ViT) [18] pre-trained on Cell Painting images will enable transfer learning and improve feature quality. The transition model must also evolve from a single-step MLP to a multi-step temporal predictor: real phenotypic screening involves time-course imaging (e.g., 24h, 48h, 72h post-treatment), and predicting the *trajectory* of morphological change is more informative than a static endpoint snapshot. Incorporating causal structure [25, 26, 27] will enable counterfactual queries ("what if we perturbed gene X before applying compound Y?") and intervention optimization under mechanistic constraints.

The path to experimental validation is clear: first, train the full pipeline on the public JUMP-CP dataset [4] (available via the Broad Institute's Cell Painting Gallery); second, validate predictions by selecting virtual hits and testing them in wet-lab Cell Painting assays; third, benchmark virtual screening hit-rates against traditional cheminformatics methods (Morgan fingerprints + random forest [19], graph neural networks [28]). If the world-model approach achieves comparable hit-rates with 10–100× speed-up (as suggested by our in-silico benchmark), it will represent a practical advance for early-stage phenotypic drug discovery. Integration with self-driving laboratory platforms [13, 14, 15] will enable closed-loop cycles where the model proposes experiments, physical assays validate predictions, and discrepancies update the transition model—realizing the full vision of AI-accelerated, hypothesis-driven science [16, 17].

References

[1] Jack W. Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11:191–200, 2012.

[2] Mark-Anthony Bray, Shantanu Singh, Hanh Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11:1757–1774, 2016.

[3] Beth A. Cimini, Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Chang, Cara Laufer, Alexandr A. Kalinin, Wes Hahn, Jason Ib, Anne E. Carpenter, and Shantanu Singh. Optimizing the cell painting assay for image-based profiling. *Nature Protocols*, 18:1981–2013, 2023.

[4] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D. Boyd, and Anne E. Carpenter. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21:1114–1121, 2024.

[5] Srijit Seal, Maria-Anna Trapotsi, Natália Aniceto, Andreas Bender, Anne E. Carpenter, Beth A. Cimini, Shantanu Singh, et al. Cell painting: a decade of discovery and innovation in cellular imaging. *Nature Methods*, 22(2):254–268, 2025.

[6] Rishi Gupta, Devanshu Srivastava, Mithilesh Sahu, Swati Tiwari, Rashmi K. Ambasta, and Pravir Kumar. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, 25:1315–1360, 2021.

[7] Davide Dario Martinelli. Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine*, 145:105403, 2022.

[8] Jannis Born and Matteo Manica. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5:432–444, 2023.

[9] Alicia Grosvenor, Richard Knepper, Sarah Harmon, and Samuel Fiorini. Hybrid intelligence systems for reliable automation: advancing knowledge work and autonomous operations with scalable AI architectures. *Frontiers in Robotics and AI*, 12:1566623, 2025.

[10] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[11] Dariusz Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[12] Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022.

[13] Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Seyed Ali McCallum, Joren Morel, Jared Murdock, Fidel Rahmanian, Milad Rankovic, Felix Rein, Hadi Saket Saidi, Jiale Shi, Marta Skreta, Joshua E. Stefanini, Haotian Tao, Daniel Trecak, Robert Was, Linqian Yin, Shirley Zhao, and Alan Aspuru-Guzik. Self-driving laboratories for chemistry. *Chemical Reviews*, 124:9633–9732, 2024.

[14] Austin B. Henson, Piotr S. Gromski, and Leroy Cronin. Designing algorithms to aid discovery by chemical robots. *Nature Reviews Chemistry*, 7:710–720, 2023.

[15] Richard B. Canty, Jeffrey A. Bennett, Keith A. Brown, Tonio Buonassisi, Sergei V. Kalinin, John R. Kitchin, Benji Maruyama, Robert G. Moore, Joshua Schrier, Martin Seifrid, Shijing Sun, Tejs Vegge, and Milad Abolhasani. Science acceleration and accessibility with self-driving labs. *Nature Communications*, 16:3856, 2025.

[16] Manu C. Ramos, Christopher J. Collison, and Andrew D. White. A review of large language models and autonomous agents in chemistry. *Chemical Science*, 16:667–684, 2025.

[17] Thomas Hartung. Ai, agentic models and lab automation for scientific discovery. *Frontiers in Artificial Intelligence*, 8:1649155, 2025.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[19] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[20] Johan Fredrik Haslum, Christos Matsoukas, Karl-Johan Leino, and Kevin Smith. Cell painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. *Nature Communications*, 15:3470, 2024.

[21] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023.

[22] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab: AI agents design new SARS-CoV-2 nanobodies with experimental validation. *Nature*, 646:716–723, 2025.

[23] Peter D. Stetson, Jason Choy, Nicole Summerville, Jordan Dudik, Wei Pan, and Peter L. Elkin. Responsible artificial intelligence governance in oncology. *NPJ Digital Medicine*, 8:74, 2025.

[24] Noam Kolt, Michal Shur-Ofry, and Ran Cohen. Lessons from complex systems science for AI governance. *Patterns*, 6(4):101341, 2025.

[25] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[26] Judea Pearl. Causality: Models, reasoning, and inference. 2009.

[27] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[28] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1263–1272, 2017.

Appendix A. Experimental Details

1.1 Synthetic Data Generation

We generate a benchmark dataset structured after the Broad Institute JUMP Cell Painting consortium [4]. The dataset comprises N=50 compounds with 100 microscopy-scale feature vectors each (5,000 samples total), assigned to five phenotype classes in round-robin fashion (10 drugs per class).

Latent space. Each class centre $\mu_k \in \mathbb{R}^{16}$ is placed on a unit circle in the first two dimensions (angular spacing $2\pi/5$) with additional random structure in dimensions 3–6. Drug-specific centres $\mu_d = \mu_{k(d)} + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.4^2 I)$, introduce within-class drug variation.

Fingerprints. Each class has a binary prototype vector (p=0.3 density) of dimension 1,024. Drug fingerprints are generated by flipping 15% of bits from the class prototype, simulating structural diversity within a pharmacological class.

Perturbation generation. Baseline states are sampled as $s_0 \sim \mathcal{N}(0, 0.3^2 I)$ (untreated cells). Perturbed states are $s_1 = \mu_d + \eta$, $\eta \sim \mathcal{N}(0, 0.9^2 I)$, representing substantial biological noise.

1.2 Model Architecture

Transition model f_θ: MLP with layers (16+1024) $\to$ 128 $\to$ 64 $\to$ 16, ReLU activation, Adam optimiser, ℓ_2 regularisation α=10^{-2}, early stopping on 10% validation split.

Phenotype classifier C: MLP with layers 16 $\to$ 64 $\to$ 32 $\to$ 5, ReLU, trained on ground-truth perturbed features with early stopping.

1.3 Extended Results

Transition quality. The transition model achieves train MSE = 0.80 and test MSE = 1.09, indicating moderate generalisation gap. Despite this regression error, predicted states land closer to class centres than individual noisy samples, enabling perfect downstream classification.

Throughput analysis. MLP inference for 10,000 compounds completes in 0.14 s on a single CPU core (~72,000 compounds/s). In a full pipeline with ViT encoding of microscopy images, we estimate ~50× overhead, yielding ~1,400 compounds/s—still orders of magnitude faster than physical screening.

t-SNE trajectories. Figure 2 shows that predicted perturbation states (stars) cluster more tightly around class centres than ground-truth states (squares), confirming the denoising effect of the learned transition model.

Table A1: Oracle classifier per-class accuracy on ground-truth perturbed test features (overall: 88.0%).

Phenotype	Oracle Acc.
Healthy	72%
Apoptotic	96%
Proliferative	92%
Cytoskeletal disruption	94%
Mitochondrial toxic	85%