A World Model of the Virtual Cell

Eric Xing and Le Song

GenBio AI

{eric.xing, le.song}@genbio.ai

May 3, 2026

Abstract

The outlook of an AI-driven digital organism, such as a virtual cell, has recently captivated much excitement and imagination from both AI and Biology communities. With a virtual cell, one can anticipate a paradigm shift of cell biology research from trial-and-error experimental exploration with cell-culture models in a wet lab, to systematic simulation of any combinatorial interventions with a computational model in a digital lab. But what constitutes an adequate realization of virtual cell? In this paper we propose an operational definition of the virtual cell based on World Model — a modern architecture recently emerged in AI research that supports advanced capabilities such as action-conditioned simulation, counterfactual reasoning, and long-horizon planning in complex dynamic environments. A world model is an AI-driven generative software system that outputs all world-possibilities upon action-prompts for simulative reasoning. When applied to biological scenarios, a world model of the virtual cell is a generative model that simulates biological possibilities of a cell under any natural or artificial interventions of the cell, or a cell population (within a tissue type or an organ). A virtual cell world model (VCWM) contrasts predictive foundation models on specific tasks, such as gene-expression perturbation prediction, as seen in some recent definitions of the virtual cell. We present a novel architecture for such a world model that enables simulated cell as an end-to-end platform: from actionable biological prompts to anticipated outcomes at all levels — molecular, structural, interactional, and morphological, in a fully aligned, integrative, multi-modal, and multi-scale fashion. We envisage that the VCWM paradigm will not simply accelerate biological experimentation, but may transform how biological possibility is explored, shifting discovery from exhaustive experimental search to structured navigation within learned cellular worlds, bringing biology closer to an industrial age of predictive design and programmable systems.

1 Introduction

Modern biology has advanced through reductionist approaches, dissecting and studying genes, proteins, pathways, and higher levels of organization—cells, tissues, organs, and organisms—in relative isolation. This success has come with the cost of increasing fragmentation: genomics, transcriptomics, proteomics, structural biology, cell biology, histopathology, and clinical studies often evolve within separate silos, each capturing only a partial view of biological reality. This division spans both modalities and scales, with molecular, cellular, tissue, organ, and organismal biology frequently studied in distinct disciplines and research divisions. However, many biological problems—particularly those involving disease, development, and intervention—remain inherently multi-scale, requiring holistic and coordinated understanding across levels of organization. Bridging these divides calls for integrative frameworks that connect molecular measurements to system-level behavior and enable holistic reasoning about biological systems.

The outlook of an AI-driven Digital Organism (AIDO) has recently captivated much excitement and imagination across both the AI and biological communities. **At its core, AIDO is a world model of biology that can predict, simulate, and program biological systems across scales, using multi-modal and multi-scale AI foundation models to integrate heterogeneous biological data into a unified computational framework [1].** With AIDO, one can envision a new paradigm in life science in which wet labs are complemented—or partially replaced—by digital laboratories, where biological experiments can be simulated and proof-of-concept designs validated *in silico*, analogous to simulation-driven workflows in semiconductor design, aerospace engineering, or nuclear systems.

At the foundational level of AIDO lies the concept of a **Virtual Cell**: a computational system capable of capturing cellular behavior. This idea has recently attracted significant attention in the computational biology community. Initiatives such as the Chan Zuckerberg Initiative's efforts toward building predictive models of cellular systems [2], and the Arc Institute's programs on large-scale biological foundation models for gene expression prediction upon knock-outs [3], offered exploratory definitions of the virtual cell as data-driven cellular representations and perturbation prediction platforms. These efforts highlight the growing recognition that cellular behavior can be modeled computationally, but they are limited in scope and function, often emphasizing one-step prediction over narrow data modalities rather than continuous and steerable (via sequential or simultaneous interventions) long-horizon simulation of cellular dynamics over multi-modal and multi-scale cellular measurements.

A virtual cell in the sense of a holistic biological simulator is not merely about predicting the RNA counts of $N - k$ genes in a generic cell when k genes are abstractly "perturbed". What a cell biologist seeks as a surrogate for a wet-lab cell culture system—or what a pharmaceutical company seeks as an alternative to early-stage drug screening—is a simulator of actionable biological outcomes: cellular states and dynamics that reflect health, behavior, and fate under realizable perturbations such as drug exposure, genetic editing, or environmental change. In this paper, we argue that such a realization of a Virtual Cell requires a conceptual shift beyond current LLM-based or task-specific foundation models, toward a new class of AI systems: world models.

A world model is an AI-driven generative system that produces possible future states of an environment conditioned on actions, enabling simulation and reasoning over trajectories rather than static predictions [4, 5]. When applied to biology, a world model of the Virtual Cell becomes a generative system that simulates the space of biological possibilities for a given cell type under arbitrary natural or artificial interventions, at both single-cell and population levels. This contrasts with predictive foundation models that learn fixed input–output mappings for specific tasks. In a world model, interventions act as operators on a latent cellular state, and the system evolves this state coherently over time.

Recent work has begun to describe large predictive models of perturbation response as "world models," but prediction alone is insufficient. A true biological world model must support action-conditioned simulation, maintain cross-modal consistency, and remain stable over multi-step trajectories. In this context, existing approaches capture important but partial aspects of biology. RNA-seq perturbation models operate within a fixed observation space, modeling gene expression without requiring a unified latent state that coherently explains molecular, structural, functional and phenotypic outcomes over time [6, 7, 3, 8, 9]. LLM-based systems extend symbolic reasoning and experimental planning, yet their internal state remains linguistic rather than biological [10, 11, 12]. Mechanistic ODE models provide explicit dynamical descriptions where mechanisms are known, but are constrained by predefined variables and limited scalability [13, 14, 15, 16, 17, 18]. Taken together, these systems offer complementary capabilities, but none—on its own—constitutes a world model of the cell.

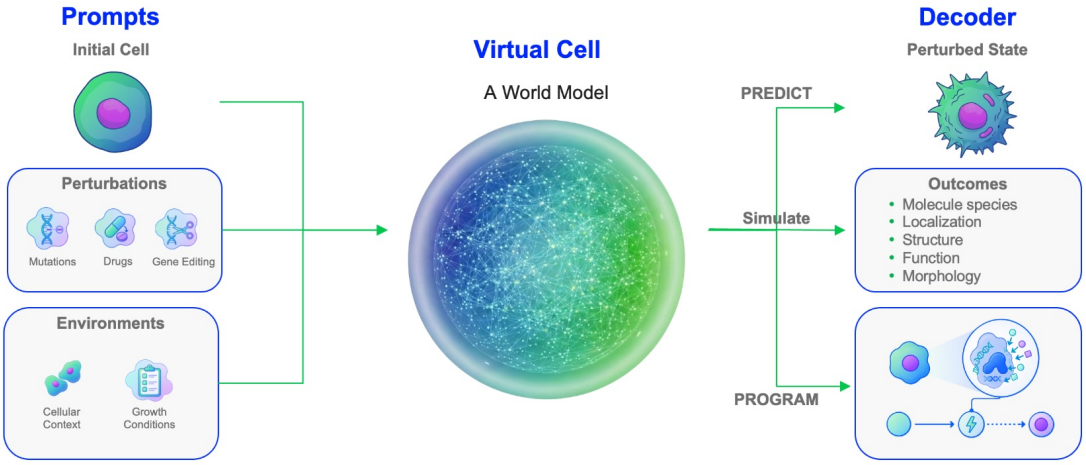

Figure 1: **From prediction to simulation.** A virtual cell is a world model of a cell, which predicts and simulates the state of a cell including molecule species abundances and regulatory relations, their structures and locations, cellular function and morphology. Interventions act as prompts in a Virtual Cell world model, which generates coherent multi-modal cellular trajectories of states rather than isolated predictions.

2 Operational Definition

We begin with an operational definition of the world model of a virtual cell anchored on multi-modal input x representing observable initial state of the cell in question, action a representing realizable exogenous intervention, perturbation, or any manipulation including change of the cellular environment applied to the cell in question; and x' representing next state of the cell conditioned on the action:

$$p(x'|x, a, e), \tag{1}$$

where e represents the environment of the cell. In the terminology of AI and world model literature, a is sometimes referred to as a "prompt", emphasizing the extrinsic, interactive, and (continuously) steerable nature of this variable versus the more intrinsic data x.

This operational definition of the Virtual Cell world model can be realized through three coupled components: a multi-modal encoder $\mathcal{E}$, an action-conditioned transition core F, and a generative decoder $\mathcal{D}$. Together, they approximate $p(x' \mid x, a, e)$ through latent cellular state evolution.

Multi-scale state reconstruction (encoder $\mathcal{E}$). The encoder maps multi-modal observations x into a latent state $z = \mathcal{E}(x, e)$ that integrates regulatory, molecular, structural, functional, phenotypic and environmental information. This representation defines a point on a cellular manifold, where RNA-seq is one projection rather than the state itself.

Action-conditioned simulation (transition core F). Given z and action a, the transition core evolves the state via $z' = F(z, a)$, inducing trajectories on the manifold. Actions act as operators on cellular state, combining symbolic reasoning over biological structure with continuous state evolution, and enabling counterfactual simulation.

Generative decoding (decoder $\mathcal{D}$). The decoder maps latent states to observables $x' = \mathcal{D}(z')$, producing coherent outputs across modalities. All generations are constrained to arise from the

same underlying state, enforcing cross-modal consistency.

Taken together, the encoder $\mathcal{E}$ defines the cellular state, the transition core F governs its evolution under action, and the decoder $\mathcal{D}$ renders observable consequences. This trinity establishes the Virtual Cell as an action-conditioned generative world model, rather than a collection of predictive mappings.

3 What can a Virtual Cell be Used for Biomedicine?

A Virtual Cell world model enables querying, simulation, and design of cellular behavior under intervention. Using K562 cell line as an example, representative use cases can be organized as in Table 1 which covers various key questions related to understanding disease mechanism, target identification, drug screening and discovery, and desired answers. These use cases illustrate a shift from querying isolated measurements to interacting with a simulated cellular system. The Virtual Cell becomes an executable model that can be interrogated, perturbed, and programmed to explore cellular possibilities.

More specifically, these capabilities map naturally onto key stages of the biomedicine pipeline. For disease understanding, a Virtual Cell provides a multi-modal representation of cellular dysfunction, linking regulatory, molecular, structural, and phenotypic alterations into a unified state. Rather than analyzing differential expression in isolation, one can trace how perturbations reshape cellular trajectories and identify critical points of failure across pathways and interactions.

For target identification, action-conditioned simulation enables systematic exploration of perturbations. By evaluating how candidate genes or proteins alter cellular states—such as viability, proliferation, or differentiation—the model can prioritize targets that most effectively drive desired phenotypic changes. Importantly, this process is inherently multi-modal, allowing targets to be assessed not only by transcriptional response but also by structural, interaction, and phenotypic consistency.

For molecule design, the Virtual Cell couples molecular generation with cellular simulation—a function we refer to as *in-context molecular design*. Candidate small molecules or proteins can be designed against targets and immediately evaluated within the same framework for their impact on cellular state, as well as properties such as binding affinity, stability, and off-target effects. This integrates molecular design with functional validation, reducing the gap between biochemical activity and cellular efficacy.

Finally, for molecular and cellular function prediction, the Virtual Cell provides a unified platform to infer how genetic variation, environmental changes, or therapeutic interventions propagate across scales. From predicting the structure and interactions of a protein to simulating its downstream effects on pathways such as localization shift and phenotype such as morphology change, the model links molecular mechanisms to system-level outcomes.

In summary, these applications suggest a shift from sequential, siloed workflows toward an integrated and iterative process. By unifying disease modeling, target discovery, and intervention design within a single computational system, the Virtual Cell has the potential to reshape how biological hypotheses are generated, tested, and translated into therapeutic strategies.

4 Architecture of a Virtual Cell World Model

A Virtual Cell world model must satisfy four architectural constraints: multi-modality, action-conditioning, cross-scale consistency, and dynamical stability. These constraints naturally give rise to three coupled components: a structured encoder $\mathcal{E}$, an action-conditioned transition core F, and a structured decoder $\mathcal{D}$. An summary of the overall architecture is illustrated in Figure 2.

Category	Representative Prompts	Representative Outputs
State Query	**1.** What is the chromatin accessibility or histone modification of gene A?	**1.** Chromatin accessibility or histone modification profiles at nucleotide resolution.
	2. What isoform of gene A is expressed, at what level, and where is it localized?	**2.** Isoform sequences, expression levels, and subcellular localization.
	3. What is the protein structure and its interaction partners?	**3.** 3D protein structure and interacting partners or complexes.
	4. What pathways are involved and what is the resulting cellular morphology?	**4.** Pathway annotations (e.g., KEGG/Reactome) and morphology feature vectors.
Intervention	**5.** What happens to viability, gene expression, and morphology if gene A is knocked out or mutated?	**5.** Predicted viability score, gene expression vector, and morphology features.
	6. How does the cell respond to RNAi, a small molecule, or a protein input?	**6.** Multi-modal response including viability, expression, and morphological changes.
Target ID	**7.** What is the likely molecular target of intervention y?	**7.** Predicted target proteins with sequences and structures.
	8. What is the structure of the molecule–target complex?	**8.** 3D structure of the molecule–target complex.
	9. Which targets most effectively drive a desired cellular state?	**9.** Ranked target candidates with predicted impact on cellular state.
Design	**10.** Can we design a molecule or protein against target B?	**10.** Designed molecules (e.g., SMILES) or protein sequences.
	11. Can we generate an intervention that induces a desired phenotype?	**11.** Candidate interventions predicted to achieve the target phenotype.
	12. What are the stability, binding affinity, and off-target effects?	**12.** Predicted stability, binding affinity, and off-target interactions.
Adaptation	**13.** Given new perturbation–response data in K562, how should the model be updated?	**13.** Updated Virtual Cell model adapted to new data and ready for querying.

Table 1: **Representative use cases of a Virtual Cell.** Prompts and outputs are independently numbered and aligned.

Structured encoder and the cellular manifold. Let $G = (V, E)$ denote the biological network of a cell, where nodes V represent entities (genes, proteins, complexes, compartments, biological pathways and processes) and edges E encode regulatory, physical, or functional relationships. Observed multi-modal data x — including sequences, structures, interaction profiles, localization, and their quantitative abundances, pathway activation — constitute measurements on G. The encoder $\mathcal{E}$ maps these heterogeneous observations together with cellular micro-environment e into a latent state

$$z = \mathcal{E}(x, G, e) \in \mathcal{M},$$

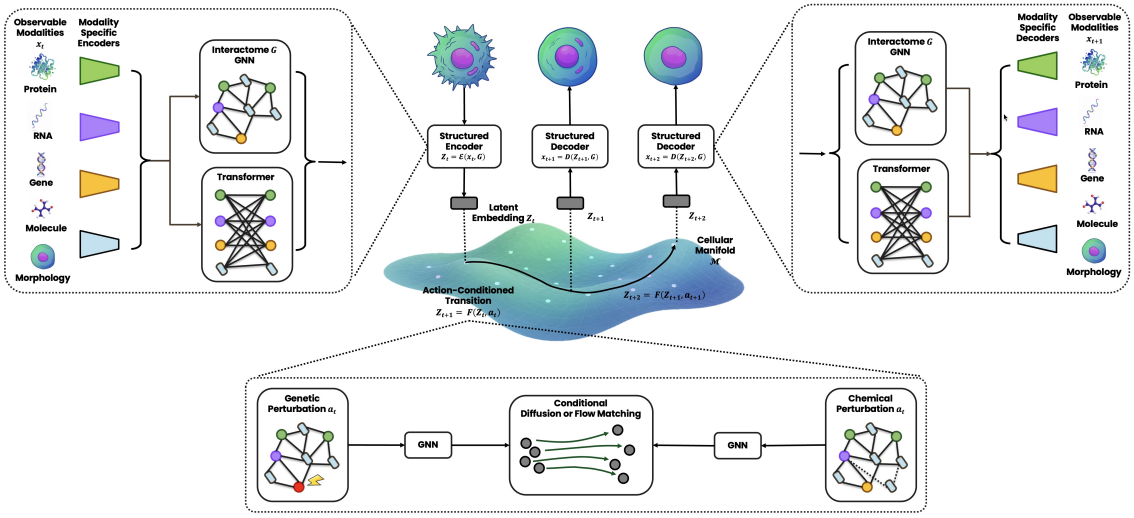

Figure 2: **Overall architecture of a Virtual Cell world model.** A structured encoder $\mathcal{E}$ embeds heterogeneous biological modalities and network structure into a unified latent manifold $\mathcal{M}$. An action-conditioned transition operator F induces trajectories $z_{t+1} = F(z_t, a_t)$ on $\mathcal{M}$, integrating symbolic reasoning and continuous dynamics. A structured decoder $\mathcal{D}$ projects latent states back to coherent molecular and phenotypic observables under cross-modal consistency constraints. The architecture of the structured encoder and decoder is a residual network architecture where a graph neural network and a transformer architecture are combined to balanced inductive bias and expressive power.

where $\mathcal{M}$ is a learned cellular manifold capturing the intrinsic geometry of viable cellular configurations.

In practice, the encoder can be implemented as a graph neural network defined over G, where node embeddings correspond to key biological entities and higher-level abstractions such as pathways and processes. Pretrained modality-specific foundation models (for sequence, structure, imaging, and quantitative omics) provide embeddings that are attached to nodes, enabling the encoder to jointly learn representations that integrate modality-specific features with relational structure.

Beyond knowledge-based graph representations, the encoder can be further enhanced through a residual network architecture. Graph neural networks can be combined with transformer-style architectures, in which latent dimensions are learned rather than pre-specified. This hybrid design balances inductive bias from biological structure with expressive capacity from data-driven representations, allowing complementary embeddings to emerge and improving robustness across incomplete or incorrect knowledge graph regimes.

In this formulation, a cell latent state z is not merely a vector of measurements, but a point on $\mathcal{M}$ constrained by the topology of G. Because the encoder is trained on diverse and partially observed data, the learned manifold also enables comparison and alignment of cellular states measured across different modalities.

Action-conditioned transition dynamics. At time t, the cellular configuration is represented by $z_t \in \mathcal{M}$. The transition core learns an action-conditioned operator

$$z_{t+1} = F(z_t, a_t),$$

where a_t denotes a realizable intervention. The operator F induces a flow on $\mathcal{M}$, approximating the dynamical system governing cellular evolution. Stability requires that repeated composition,

$$z_{t+k} = F^{(k)}(z_t, a_{t:t+k-1}),$$

remains within biologically plausible regions of $\mathcal{M}$.

Because cellular dynamics combine discrete regulatory logic with continuous quantitative evolution, F integrates two complementary components:

- A symbolic reasoning module operating over G that infers relational updates, such as pathway activation or network rewiring. Interventions are represented as structured operators on G: genetic edits modify V or E, molecular inputs augment V, and environmental changes alter contextual embeddings. In this way, action-conditioning is embedded directly within the biological representation. Such reasoning can be implemented using large language models grounded in scientific literature or biological knowledge graphs, as well as graph neural networks defined over structured biological networks.
- A continuous latent dynamics module that propagates these updates as smooth trajectories on $\mathcal{M}$, modeling gradual changes in gene expression, protein abundance, localization, and morphology. In practice, diffusion- or flow-based generative models provide natural realizations of such dynamics

Together, these components define a hybrid dynamical system coupling combinatorial structure with continuous flow.

Structured decoder and cross-modal consistency. The decoder $\mathcal{D}$ maps latent states back to observable modalities,

$$x'_t = \mathcal{D}(z_t, G),$$

producing predictions for expression profiles, protein localization, interaction probabilities, structural conformations, and phenotypic and morphological descriptors. Rather than independent prediction heads, $\mathcal{D}$ enforces cross-modal consistency through the shared geometry of $\mathcal{M}$ and the topology of G, ensuring that all outputs remain mutually compatible.

The overall architecture is organized around a single evolving cellular state z_t on $\mathcal{M}$. The encoder $\mathcal{E}$ defines the manifold, the transition operator F governs motion under intervention, and the decoder $\mathcal{D}$ renders coherent biological observables. In this formulation, the Virtual Cell is not a collection of predictive modules, but an action-conditioned generative dynamical system constrained by biological structure.

5 Key Differentiator

The defining distinction of a Virtual Cell world model versus existing and conventional approaches based on predictive models is not incremental performance over conventional benchmark tasks such as RNAseq prediction under perturbation, but a complete re-imagination of the nature of the output and user experience as reflected in the scope, information volume, and complexity of the outcome. Rather than learning a task-specific mapping (for example, $a \mapsto \Delta\text{RNA}$), it seeks to learn a cellular manifold $\mathcal{M}$, an action-conditioned transition operator F, and a generative decoder $\mathcal{D}$ defined over a structured biological network $G = (V, E)$. In this formulation, latent cellular states $z_t \in \mathcal{M}$ evolve under intervention,

$$z_{t+1} = F(z_t, a_t),$$

and all molecular and phenotypic observables are coherent projections of this evolving state through $\mathcal{D}(z_t)$. The objective is not to predict a readout, but to simulate a world.

RNA-seq perturbation models capture one projection of cellular response, often with remarkable accuracy, yet they operate within a fixed observation space. By contrast, a world model constrains transcriptomic, proteomic, structural, and phenotypic layers to arise from the same underlying state, enforcing cross-modal coherence and enabling multi-step rollouts. What is predicted is not merely an endpoint, but a trajectory.

LLM-based biological reasoning systems have expanded our capacity for symbolic synthesis and experimental planning. However, their internal state is primarily linguistic; they reason about biology without instantiating a quantitative cellular system. In a Virtual Cell, symbolic reasoning becomes embedded within F to update relational structure in G, but is inseparable from continuous latent dynamics on $\mathcal{M}$. Semantics is coupled to state evolution.

Mechanistic ODE models remain the gold standard where mechanisms are known and systems are tractable, providing explicit dynamical equations over predefined variables. A world model takes a complementary path: instead of prescribing equations, it learns the geometry of $\mathcal{M}$ and the induced flow F directly from data, using G as structural scaffolding. The aim is not to replace mechanistic insight, but to scale dynamical simulation to the full complexity of cellular organization.

In this light, the Virtual Cell World Model represents a shift from isolated prediction to integrated simulation: an action-conditioned, multi-modal dynamical system in which interventions act as operators on biological structure and induce motion on a learned manifold of cellular possibility. Rather than projecting outcomes, it aspires to model the space in which those outcomes live.

6 Data Requirements

World modeling reframes the data problem as learning three coupled components: a multi-modal state manifold (encoder–decoder learning), an action-conditioned transition operator (one-step dynamics), and stable long-horizon trajectories (multi-step evolution). Each component places distinct but complementary demands on biological data.

Learning the state manifold (encoder–decoder data). The first objective is to learn a coherent latent representation of cellular state. Let $z \in \mathcal{M}$ denote an embedding on a cellular manifold $\mathcal{M}$ that integrates genomic context, transcriptomic profiles, proteomic abundance, structural conformations, interaction networks, and spatial localization. Large-scale multi-modal datasets provide samples from this manifold. The encoder maps heterogeneous observations into z, while the decoder reconstructs observables from z, approximating a generative model of cellular configuration. Crucially, $\mathcal{M}$ is defined jointly across modalities. Pretrained foundation models for sequence, structure, imaging, and quantitative omics produce aligned embeddings in a shared latent space, such that geometric proximity reflects functional similarity. Perfect co-registration is unnecessary; scale and diversity allow the model to approximate the intrinsic geometry of cellular states and support cross-modal reconstruction.

Learning one-step action-conditioned transitions. The transition operator $z_{t+1} = F(z_t, a_t)$ requires empirical samples of how interventions move cellular state. Perturbation–response datasets, such as the CRISPR knockout screens, RNAi and overexpression studies, small-molecule treatments, and multi-omic measurements under intervention approximate action–state pairs (z_t, a_t, z_{t+1}). These data supervise the local dynamics of the manifold, teaching the model how specific edits, molecular inputs, or contextual shifts reshape regulatory configuration, abundance profiles, and phenotype. Even single-step or short-horizon measurements provide directional constraints on F, limiting the space of plausible transitions and anchoring the model's action semantics.

Learning long-horizon trajectory structure. Beyond local transitions, a world model must learn stable global flows on the cellular manifold. Direct supervision can be obtained from live-cell

imaging, which provides temporally resolved measurements of cellular dynamics and can be naturally integrated within a multimodal encoder–decoder framework. In contrast, when explicit time-series data are limited, single-cell RNA velocity offers a scalable alternative. By jointly modeling unspliced and spliced transcripts, RNA velocity estimates local directional vectors $\dot{z}$ in gene expression space, approximating a vector field over the latent manifold. Integrating these local velocities yields pseudo-temporal trajectories that capture state progression. These trajectories serve as weak but structured supervision for training the transition operator F. In this way, locally inferred vector fields constrain global flow geometry, enabling coherent multi-step evolution even in the absence of densely sampled temporal data.

Together, these data sources teach the model three things: where cellular states live, how actions move them, and how trajectories unfold. Multi-modal measurements shape the manifold, perturbations define local transitions, and temporal or pseudo-temporal signals constrain long-range flow—collectively enabling a stable, action-conditioned simulator of cellular possibility.

7 Why a World Model?

World models provide a natural framework for advancing virtual cells from predictive tools to simulators of biological systems. Most current approaches in computational biology focus on mapping inputs to outputs—for example, predicting gene expression changes under perturbation. While useful, such models remain largely reactive, lacking an explicit representation of how cellular states evolve under intervention. A world model instead seeks to learn the structure and dynamics of the cellular system itself, enabling the simulation of trajectories rather than isolated outcomes.

In this setting, the virtual cell becomes a computational environment in which interventions—genetic, chemical, or environmental—can be applied and their consequences traced across molecular, structural, and phenotypic levels. This allows the system to anticipate the effects of perturbations before they are experimentally realized, transforming biological modeling from retrospective analysis to prospective reasoning. Such capability is particularly important given the cost and complexity of biological experiments, where exhaustive empirical exploration is often infeasible.

Beyond efficiency, the key advantage of a world model lies in its ability to capture causal structure. By linking genome, expression, protein interactions, and cellular phenotype within a unified dynamical framework, it enables reasoning about mechanisms rather than correlations alone. This is essential for tasks such as target identification and therapeutic design, where understanding how interventions propagate through the system is as important as predicting their endpoints.

More broadly, world models suggest a shift toward simulation-driven biology. Just as computational simulation has become central to fields such as engineering and climate science, virtual cells grounded in world models may serve as a substrate for exploring and designing biological systems. In this view, the goal is not only to predict what a cell will do, but to understand how it can be controlled—bringing biology closer to a programmable and designable discipline.

8 Computational and Technical Hurdles

Building a Virtual Cell requires solving a representation problem before it becomes a simulation problem. Cellular state is distributed across heterogeneous modalities, including genomic and epigenomic context, gene expression, protein abundance, localization, and morphology. Yet paired multi-modal measurements remain scarce, whereas unimodal datasets are abundant. A central technical challenge is therefore to learn a unified representation of cellular state that can leverage both sparse paired data and large-scale unpaired data. This will likely require mixed pretraining and post-training strategies that combine reconstruction objectives, cross-modal reconstruction, and contrastive alignment informed by biological prior knowledge. How to balance these objectives and stabilize training remains a core active research problem.

A second challenge is scale. Both dataset size and input context can be substantial, especially when genomic sequences are included. Capturing such information may require models with billions of parameters and architectures capable of handling very long contexts. This creates significant systems-level demands: tensor parallelism to distribute model parameters across GPUs, pipeline parallelism to distribute layers across machines, and context parallelism to partition long inputs across devices. In each case, performance depends not only on algorithmic design but on careful systems engineering to minimize communication overhead and overlap communication with computation. Efficient training of Virtual Cell models is therefore not merely a modeling challenge, but a large-scale distributed computing challenge.

A third hurdle concerns the representation of temporal dynamics. Because densely sampled longitudinal measurements remain limited, current approaches often rely on pseudo-time trajectories inferred from snapshot data, for example through RNA velocity or related methods. However, these inferences are themselves technically fragile. Single-cell RNA-seq is noisy, making local velocity estimates uncertain, and reconstructing trajectories in high-dimensional state spaces is intrinsically difficult. Reliable pseudo-time therefore requires careful regularization and validation. In practice, the quality of the learned world model may depend as much on the fidelity of these inferred dynamical signals as on the architecture used to model them.

9 Evaluation

Evaluation of a Virtual Cell world model must reflect and validate its ambition: not the prediction of isolated endpoints of a small subset of the biological variables in isolation , but the simulation of a cellular world with all variables of interest. Because the system is defined by a learned manifold $\mathcal{M}$, an action-conditioned transition operator F, and a structured decoder $\mathcal{D}$ over a biological network $G = (V, E)$, its performance cannot be reduced to a single metric. It must be judged on coherence, dynamical fidelity, and experimental utility.

Existing evaluation frameworks—such as RNA-seq perturbation benchmarks and leaderboards by Arc instistute capture only a narrow projection of cellular state and are therefore not indicative of the quality of a Virtual Cell. A world model must be evaluated on its ability to generate coherent multi-modal outputs, where gene expression, protein abundance, localization, interaction networks, and regulatory states remain mutually consistent as projections of a single latent state.

This calls for a new evaluation paradigm—a *Virtual Cell Arena*—that tests action-conditioned simulation across modalities, compositional perturbations, and long-horizon trajectories. The goal is not only to match observations, but to assess whether $(\mathcal{E}, F, \mathcal{D})$ together define a coherent and stable simulator of cellular evolution under intervention.

Computational validation: coherence and flow. At the representational level, a latent state $z \in \mathcal{M}$ decoded as $x' = \mathcal{D}(z)$ should yield mutually consistent projections across modalities. Transcriptomic, proteomic, structural, and phenotypic outputs must agree as manifestations of a single evolving state. At the dynamical level, the transition operator

$$z_{t+1} = F(z_t, a_t)$$

must generalize beyond observed interventions and remain stable under composition,

$$z_{t+k} = F^{(k)}(z_t, a_{t:t+k-1}).$$

Evaluation therefore includes unseen perturbations, combinatorial actions, and multi-step rollouts, testing whether trajectories remain within biologically plausible regions of $\mathcal{M}$. The goal is not merely predictive accuracy, but preservation of coherent flow on the cellular manifold.

Biological validation: perturbation as probe. A Virtual Cell should also be evaluated as one would evaluate a surrogate experimental system. Multi-modal profiling—transcriptomics, proteomics, and high-content imaging—can calibrate whether $\mathcal{D}(z)$ captures real cross-modal coupling. Controlled perturbations (CRISPR edits, RNAi, small molecules) provide empirical tests of F, comparing predicted state transitions to observed responses. Sequential and combinatorial perturbations probe whether simulated trajectories anticipate real cellular evolution under protocol-level interventions.

Generative utility and closed-loop discovery. The most transformative benchmark lies in inverse control. Given a target region $\mathcal{M}^* \subset \mathcal{M}$ corresponding to a desired phenotype, the model proposes actions a intended to move z toward $\mathcal{M}^*$. Experimental validation of these proposals—measuring enrichment of successful outcomes relative to baseline screening—tests the model not only as a predictor but as a design instrument.

In this perspective, evaluation becomes the validation of a new kind of scientific apparatus. The central question is not whether $\mathcal{D}(z)$ matches data in isolation, but whether $(\mathcal{E}, F, \mathcal{D})$ together define a coherent, stable, and actionable simulator—one that can be interrogated, perturbed, and ultimately trusted as a computational counterpart to the living cell.

Call for community benchmarks and data. Realizing this evaluation paradigm will require coordinated community effort. We call for the development of new datasets and benchmarks tailored to Virtual Cell world models, including goal-oriented experimental data in which perturbations are designed to achieve explicit phenotypic outcomes, and long-horizon, multi-modal measurements that track cellular state continuously across time. Such datasets should integrate molecular, structural, and phenotypic readouts under controlled interventions, enabling rigorous assessment of action-conditioned simulation and dynamical consistency. Establishing a shared *Virtual Cell Arena*—with standardized tasks, datasets, and evaluation protocols—will be critical for advancing the field from isolated predictors to robust, experimentally grounded simulators of cellular systems.

10 From Integration to Holistic Modeling

We can begin the implementation of a Virtual Cell (VC) system with traditional integration which organizes function- or modality-specific modules under a unified framework. That is, modeling the effects of gene perturbations (e.g., knockout or mutation) relies on separate datasets and models to predict outcomes such as gene expression, protein abundance, localization, morphology, and even tissue-level changes, often through independent encoding–decoding pipelines.

In this setting, constructing a Virtual Cell through conventional pipelines becomes increasingly dependent on human expertise, as models must be developed or adapted for each modality and dataset. This process does not scale with the growing diversity and volume of biological data. A natural direction is therefore to develop systems that automate model construction and adaptation. Recent advances in AI coding agents suggest a paradigm in which *AI builds AI*: models are proposed, implemented, and refined through automated, iterative workflows. Applied to the Virtual Cell, this approach enables the coordinated use of multiple foundation models across modalities, while reducing the need for manual engineering. Realizing this vision requires new forms of harness engineering that orchestrate coding agents and model families into a unified system. Frameworks such as VCHarness exemplify this direction, enabling rapid construction and integration of models for Virtual Cell applications [19].

However, such integration remains largely superficial: outputs across modalities are not necessarily linked by a coherent causal structure. In reality, biological systems operate as a continuous cascade of effects across scales and modalities. A perturbation at the genome level induces changes

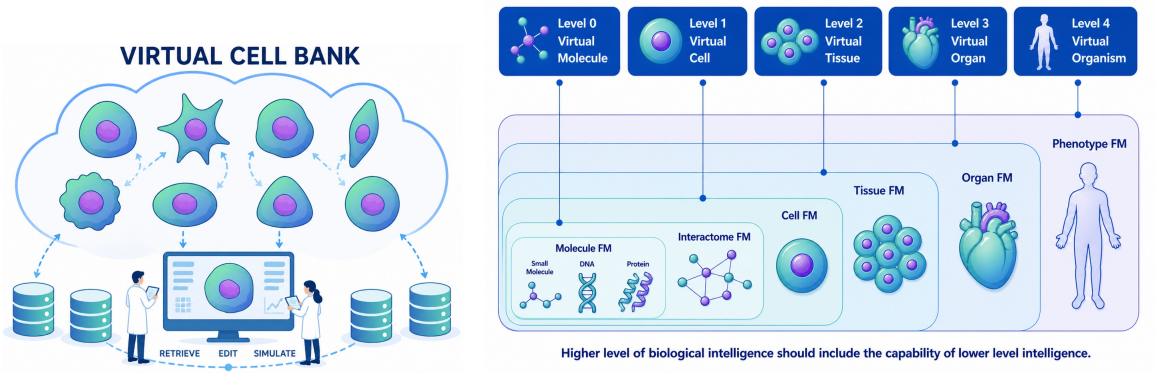

Figure 3: **Two axes of expansion.** Along the vertical axis, a Virtual Cell Bank accumulates diverse cellular states across types, genotypes, and health conditions. Along the horizontal axis, virtual cells compose into higher-order systems, such as tissue, organ and organism, forming progressively integrated biological world models.

in transcription, which propagate to protein abundance and structure, alter protein–protein interactions, and affect subcellular localization. These molecular and structural changes reshape cellular functions and morphology, and can further propagate to tissue organization and higher-order biological structures. In turn, feedback from these higher levels can influence regulatory states and genome activity. Prediction and simulation are therefore inherently coupled: to simulate a perturbation is to trace its causal propagation through this multi-scale chain, rather than to independently predict isolated endpoints.

Furthermore, instead of a stateless latent embedding of a cell for one-shot prediction given a prompt, a world model treats the cell as a stateful system. The latent representation evolves over time and accumulates the effects of sequential and combinatorial perturbations—such as gene knockout, chemical treatment, and environmental signaling. Each intervention modifies the cellular state, which then serves as the initial condition for subsequent dynamics. Moreover, cellular behavior is intrinsically temporal: internal processes such as the cell cycle and regulatory oscillations introduce an endogenous clock that governs state evolution. A Virtual Cell world model must therefore capture both externally driven transitions and intrinsic temporal dynamics, enabling simulation of trajectories rather than isolated responses as explained in the architecture in Section 4.

11 Toward Virtual Cell Banks and Digital Organisms

The future of the Virtual Cell unfolds along two complementary axes: vertical diversification of cellular states and horizontal composition into higher-order biological systems.

Vertical axis: diversity, stratification, and personalization. A Virtual Cell Bank represents the vertical scaling of cellular intelligence. Each latent state $z \in \mathcal{M}$ becomes a reproducible generative seed corresponding to a defined cellular configuration by cell type, developmental stage, disease condition, or genetic background. As data accumulate, this repository could extend beyond canonical cell lines to encompass stratified populations and, ultimately, individualized virtual cells reflecting specific genotypes, environmental exposures, or health status.

Unlike physical biobanks, virtual cells are non-depletable, forkable, and evolvable. A single state can be cloned, perturbed, and explored across alternative intervention trajectories without material constraint. Over time, the Virtual Cell Bank becomes not merely a collection of states,

but a structured atlas of cellular diversity and transition mapsi.e., a navigable landscape of how different biological contexts respond to action. In principle, interventions could be evaluated across a panel of virtual individuals before entering experimental or clinical testing, shifting discovery from one-size-fits-all to stratified exploration.

Horizontal axis: compositional biological intelligence. Parallel to diversification is compositional scaling. The Virtual Cell is the minimal viable unit of biological world modeling, but it is not the terminus. Tissues, organs, and organisms introduce emergent constraints, such as spatial organization, intercellular signaling, metabolic coupling, mechanical forces, that cannot be reduced to single-cell behavior.

In horizontal expansion, virtual cells become building blocks. A Virtual Tissue aggregates interacting states; a Virtual Organ coordinates tissue-level flows under physiological constraints; a Virtual Organism integrates organ dynamics into systemic regulation. Each layer composes previously learned manifolds and transition operators into larger dynamical systems. Organism-level simulation does not bypass the cell, but it recursively builds upon it. If realized, such composition would mark a shift from modeling fragments of biology to modeling its structured integration.

A two-dimensional biological intelligence. Vertically, the Virtual Cell Bank increases biological diversity and personalization; horizontally, compositional modeling increases systemic complexity. Together, these axes outline a trajectory from individual cellular simulators to stratified digital cell populations and, ultimately, multi-scale digital organisms. The cell remains the atomic unit, but replicated and evolved in silico, it becomes the foundation for a new layer of biological intelligence.

Ethical and data-governance considerations. The prospect of personalized virtual cells introduces non-trivial ethical and governance questions. A digital representation of an individual's cellular states, particularly when derived from genomic and longitudinal health data, raises issues of privacy, consent, ownership, and potential misuse. Unlike de-identified datasets, virtual cells may encode actionable biological information. Robust safeguards, transparent data governance, and clear consent frameworks will be essential if personalized virtual cells are to become a responsible component of biomedical infrastructure. As biological world models grow in fidelity, their governance must evolve in parallel.

12 Conclusion

A world model of the Virtual Cell reframes biological modeling from predicting measurements to simulating systems. By learning a cellular manifold $\mathcal{M}$, an action-conditioned transition operator F over biological networks $G = (V, E)$, and a coherent decoder $\mathcal{D}$ across modalities, the Virtual Cell defines a dynamical system rather than a projection. Interventions become operators on structure; trajectories become flows on a learned state space; phenotype emerges as a consistent projection of an evolving latent state.

This is a shift in scope. RNA-seq models predict endpoints; symbolic systems reason about mechanisms; mechanistic ODEs specify equations. A Virtual Cell instead seeks to integrate representation, action, and evolution within a single generative framework—an executable model of cellular reality. It need not encode every biochemical detail to be useful; coherence can emerge from scale, structure, and iterative experimental refinement.

Along one axis, Virtual Cell Banks may stratify cellular states across types, genotypes, and health conditions. Along another, virtual cells may compose into tissues, organs, and eventually digital organisms. The cell remains the atomic unit—but replicated in silico, it becomes the foundation of scalable biological intelligence.

If realized, the Virtual Cell will not simply accelerate experimentation, but may transform how biological possibility is explored, shifting discovery from exhaustive search to structured navigation within learned cellular worlds. More broadly, this paradigm points toward a transition of biology into a computational and design-driven discipline. Just as computer-aided design (CAD) facilitated architecture and mechanical design, electronic design automation (EDA) revolutionized semiconductor engineering, and large-scale simulation underpins modern weather prediction, Virtual Cell world models may become the computational substrate for biological design. In this view, simulation and computational reasoning become integral to decision-making, guiding experimentation, intervention, and engineering, and bringing biology closer to an industrial age of predictive design and programmable systems.

References

[1] Le Song, Eran Segal, and Eric P. Xing. Toward ai-driven digital organism: Multiscale foundation models for predicting, simulating and programming biology at all levels. *arXiv preprint arXiv:2412.06993*, 2024.

[2] Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric P. Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

[3] Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025.

[4] Eric P. Xing, Mingkai Deng, Jinyu Hou, and Zhiting Hu. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025.

[5] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in Neural Information Processing Systems*, 2018.

[6] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024.

[7] Haotian Cui, Chao Wang, Hritik Maan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 2024.

[8] Guohui Chuai, Xiaohan Chen, Xingbo Yang, Cheng Zhang, Kairu Qu, Yiheng Wang, Wannian Li, Jingya Yang, Duanmiao Si, Feiyang Xing, Yicheng Gao, Siqi Wu, Shaliu Fu, Bing He, and Qi Liu. Towards building a world model to simulate perturbation-induced cellular dynamics by alphacell. *bioRxiv*, 2026.

[9] Chloe Wang, Mehran Karimzadeh, Neal G. Ravindra, Lexi R. Bounds, Nader Alerasool, Ann C. Huang, Shihao Ma, Daniel R. Gulbranson, Haotian Cui, Yongju Lee, Anusuya Arjavalingam, Elliot J. MacKrell, Matthew S. Wilken, Jieming Chen, Benjamin W. Herken, Jesse A. Weber, Massimo M. Onesto, Barbara Gonzalez-Teran, Nicole F. Leung, Sally Yu Shi, Byron J. Smith, Sharon K. Lam, Adam Barner, Philip Wright, Elizabeth M. Rumsey, Soohong Kim, Rene V. Sit, Adam J. Litterman, Ci Chu, and Bo Wang. X-cell: Scaling causal perturbation prediction across diverse cellular contexts via diffusion language models. *bioRxiv*, 2026.

[10] Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, Chang Li, Emily Sun, David Jeong, Lawrence Zhao, Jennifer Kwan, David Braun, Brian Hafler, Jeffrey Ishizuka, and David van Dijk. Scaling large language models for next-generation single-cell analysis. *bioRxiv*, 2025.

[11] Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: A biological world model for virtual cell simulation. In *International Conference on Learning Representations (ICLR)*, 2026.

[12] Yicheng Gao, Weixu Wang, Yuheng Zhao, Kejing Dong, Caihua Shan, Weizhong Zheng, Till Richter, Zekai Li, Siming Chen, Fabian J. Theis, and Qi Liu. Language may be all omics needs: Harmonizing multimodal data for omics understanding with cellhermes. *bioRxiv*, 2025.

[13] Boris M. Slepchenko, James C. Schaff, Ian Macara, and Leslie M. Loew. Quantitative cell biology with the virtual cell. *Trends in Cell Biology*, 13(11):570–576, 2003.

[14] Jonathan R. Karr, Jayodita C. Sanghvi, Derek N. Macklin, Miriam V. Gutschow, Jared M. Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I. Glass, and Markus W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

[15] Arthur P. Goldberg, Balázs Szigeti, Yin Hoon Chew, Jeyashree A. Sekar, and Frederick P. Roth. Emerging whole-cell modeling principles and methods. *Current Opinion in Biotechnology*, 51:97–102, 2018.

[16] Konstantina Georgouli, Jae-Sung Yeom, Robert C. Blake, and Ali Navid. Multi-scale models of whole cells: Progress and challenges. *Frontiers in Cell and Developmental Biology*, 11:1175803, 2023.

[17] Vivien Marx. How to build a virtual embryo. *Nature Methods*, 20:1023–1027, 2023.

[18] Zane R. Thornburg, Andrew Maytin, Jiwoong Kwon, Troy A. Brier, Benjamin R. Gilbert, Enguang Fu, Yang-Le Gao, Jordan Quenneville, Tianyu Wu, Henry Li, Talia Long, Weria Pezeshkian, Lijie Sun, John I. Glass, Angad P. Mehta, Taekjip Ha, and Zaida Luthey-Schulten. Bringing the genetically minimal cell to life on a computer in 4d. *Cell*, 2026.

[19] Xingyi Cheng, Pan Li, Han Guo, Youwei Liang, Jing Gong, William de Vazelhes, Changjiang Gou, Pengtao Xie, Le Song, and Eric P. Xing. Harnessing ai to build virtual cells. *bioRxiv*, 2026.