CELLULAR INTELLIGENCE

Engineering Cell Fate: Towards a Foundation Model for Virtual Cell Signaling

JANUARY 7, 2026

We are building a future where biology is no longer destiny, but design.

The Challenge

Modern biology struggles to predict and control cellular behavior because cell signaling is complex and context-dependent. Current methods rely on slow, empirical trial and error.

The Cellular Intelligence Solution

Cellular Intelligence is building the first **Universal Virtual Cell-Signaling Model**: a foundation model capable of predicting how any cell in any state changes in response to external signals.

The Competitive Advantage

- **Unrivaled Data Scale:** Utilizing a proprietary capsule-based platform, Cellular Intelligence generates massive, context-rich datasets—scaling to millions of unique perturbation conditions—to solve the problem of context dependence.

- **Static vs. Dynamic States:** While others profile cells in fixed states, we use human stem cells to decode the combinatorial signaling logic that determines cellular behavior and ultimately cell fate, turning the biological mystery of how cell types are made into a tractable engineering challenge.

Core Architecture

Our framework is built on a synergistic feedback loop between massive-scale data generation and predictive modeling. This proprietary capsule data engine covers the astronomical search space of cell signaling, distilling it into the context-rich, high-fidelity datasets required to train transformer models to learn the fundamental "grammar" of cellular signaling. As our data scales, these architectures will evolve from discrete response predictions to high-resolution continuous-time models of biological behavior, culminating in a universal simulation engine that enables the engineering of cell fate.

Translational Impact

By transforming biology into a predictive engineering discipline, Cellular Intelligence enables *in silico* control of cellular behavior, with applications ranging from rational protocol design for regenerative medicine to context-specific drug effect prediction and systematic disease modeling. This fundamentally transforms the ability to discover new treatments and save patient lives.

1.

A New Era of Cellular Control: From Empirical Biology to Predictive Control

A major bottleneck in modern medicine is the inability to predict how different cells respond to signals. We are replacing slow, manual experimentation with a predictive model that handles this complexity, accelerating the path to life-saving therapies.

A fundamental challenge in modern biology is that of precise, engineered cellular control. Cells possess their own language for communicating with each other—cell signaling—which directs core biological processes like development and is frequently dysregulated in disease.

Remarkably, biology achieves this complexity using a surprisingly concise vocabulary: only around **20 fundamental molecular signaling pathways** have been identified to date. It is the combinations and orders in which they are used that underlies how such a small number of pathways can give rise to the staggering diversity of human cell types and states. In principle, because these pathways are readily manipulated by small molecules, they provide a potent mechanism through which we could control cellular decision-making.

However, despite decades of effort, we have not yet deciphered the *grammar* of this language. Today, the effects of a given signal are largely determined through an empirical, trial-and-error process. This is due to two compounding challenges:

1. **Combinatorial Complexity:** The sheer number of signal combinations limits systematic experimental dissection.
2. **Context Dependence:** The effect of a signal depends heavily on the state of the cell prior to receiving it.

The Human Cost of Technical Limitations. The failure to decode the logic of cell signaling is not just a scientific bottleneck, but a systemic barrier to progress and, consequently, a delay in saving lives. The inability to predict cellular behavior stalls progress in regenerative medicine, where scientists painstakingly test countless combinations to guide stem cells into desired tissues, and in pharmacology, where therapies fail because we cannot foresee how diseased cells will react.

Patients waiting for organ transplants, individuals with genetic disorders, and the unfulfilled promise of personalized medicine cannot afford another decade of manual optimization. Our urgency to move beyond trial-and-error experimentation is driven by patient need.

CELLULAR INTELLIGENCE'S VISION

This white paper outlines Cellular Intelligence's solution to the challenge of predicting and controlling cellular behavior: the construction of the first **Universal Virtual Cell-Signaling Model**, a platform intended to compute how any cell state will change in response to external signals.

By combining the paradigm of developmental biology—nature's own proving ground—with our proprietary capsule platform, we transform cell signaling from an empirical art into an engineering discipline built for therapeutic design. We aim to unlock high-impact applications: from guided cell therapies that replace lost tissues, to context-specific drug response prediction, to new ways of modeling disease as signaling network failures.

Cells possess their own language for communicating with each other

2.

What is a Virtual Cell-Signaling Model?

The virtual cell-signaling model acts as a computational twin, using a cell's initial state to accurately predict how it responds to signals, thereby replacing trial-and-error with precise in silico simulation.

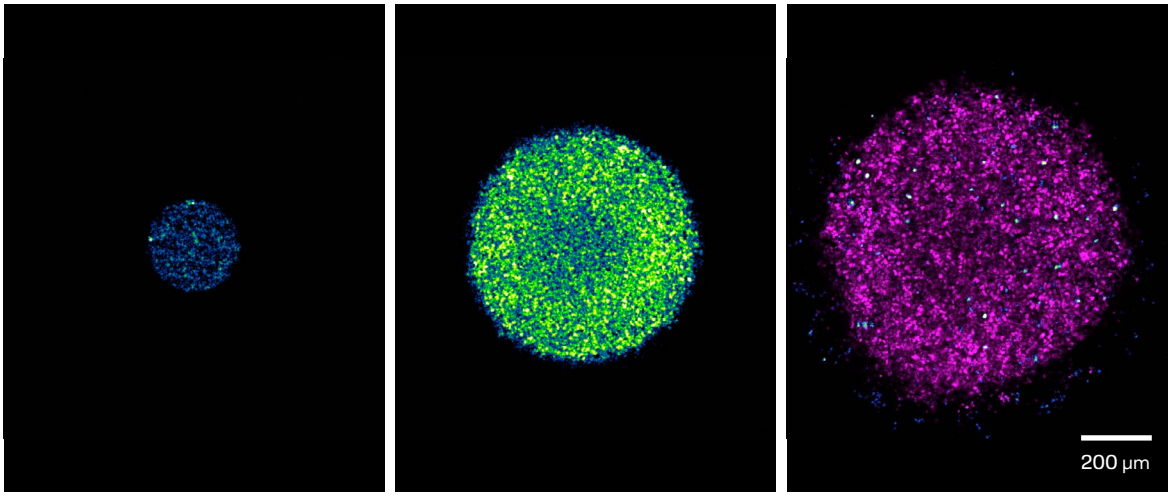
In essence, a virtual cell-signaling model is a predictive map from an initial cell state and an external signal to the cell's future state. Formally, it can be seen as a function:

f(initial cell state, signal) → future cell state

Crucially, the "cell state" encompasses the cell's molecular profile (e.g., its transcriptome and epigenetic status) and functional identity—*i.e.*, the "context" in which the signal is being applied [Wagner *et al*., 2016]. Here, "signal" refers to a perturbation, like a small molecule or growth factor, that affects a particular signaling pathway or pathways, at a specific dose. The model's output is a predicted new cell state (including gene expression changes) after the cell has been exposed to the signal.

In short, the model answers the question: "Given this type of cell (stem cell, cancer cell, neuron) and this signal at this dose, what will the cell look like and do next?"

This virtual cell-signaling model thus becomes a computational twin of living cells, allowing us to simulate how cells in new contexts would respond to signals. Our hypothesis is that our platform, based on the differentiation of all human cell types during development, will provide sufficient data for our model to generalize, enabling it to predict cell responses in as yet unseen contexts. This model will effectively solve the cell-signaling problem, enabling scientists and engineers to use *in silico* experiments to guide real-world decisions.



An example of the rich, multi-step dynamics a virtual cell-signaling model is designed to capture and predict *in silico*. Time-lapse images of human iPS cells differentiating toward a musculo-vertebral precursor fate, with green and purple fluorescent reporters marking distinct stages of maturation. (Credit: K. Zhu, Pourquié Lab)

3.

A Unique Approach to Building the "Virtual Cell"

Most AI models in biology fail because they are trained on limited data—like trying to learn a language by reading just one book. Cellular Intelligence uses stem cells to generate massive, proprietary datasets that cover the entire 'tree of life,' capturing how cells behave in every possible context. We use a 'split-and-pool' technique to run millions of unique experiments in parallel, creating the massive, high-quality training data that is required for a predictive foundation model.

Biology has run into a complexity barrier that is now blocking progress. The field requires a fundamental shift in approach, from mapping static cell states to modeling the dynamic transitions between them.

This is the gap that Cellular Intelligence is built to fill. Our approach is not just incrementally better—it is fundamentally different in how it tackles scale and complexity. Our key insight at Cellular Intelligence is that building a universal machine learning model for cellular signaling requires the interrogation of signaling across a very large number of *contexts*. Virtually all similar efforts towards cell foundation models take the approach of applying as many perturbations as possible (drug screens, genetic screens) to a limited number of cell types, often in the single digits. This limitation is inherent to their data generation methodology, which is rooted in existing screening paradigms. Thus, other datasets, while perturbation-rich, are *context-poor*, meaning that they have very limited exposure to different cellular contexts. Given that context-dependence is the biggest challenge to predicting the effects of signaling, such approaches cannot adequately train a model that will generalize across contexts.

Our approach takes full advantage of the paradigm of developmental biology, the natural process by which stem cells differentiate into all the different cell types in the human body. Stem cells have an innate ability to adopt a vast number of cellular states as guided by cell signaling, hence providing the ideal platform for generating the rich, pan-context signaling data required to train general models of cell signaling. By leveraging the process of development, we will be able to learn how signaling works across the widest possible range of human cell types, enabling us to direct cells towards particular fates and away from others. Our platform enables us to explore the order, concentration, and combination of perturbations in a way that others cannot match. Our approach has numerous advantages:

- **Exponentially Scalable Data Collection via Capsule Technology:** We recognized early that the biggest blocker to a generalizable signaling model was data. Traditional experimental platforms are context-starved—they might test many perturbations, but only on a handful of cell states, most typically derived from easy-to-use but less physiological cancer cells. Cellular Intelligence overcame this bottleneck with a proprietary capsule-based context generation system. In our platform, pluripotent stem cell colonies are grown in microscale capsules that can be split-and-pooled through multiple treatment steps, each capsule accruing a unique barcode to record its treatment history. This allows us to interrogate an exponentially expanding set of signaling factors combinations and of cell states that in principle can populate the developmental tree with only linearly increasing effort. For example, in a recent experiment we started with 30 combinations of signaling factors and applied them in 3 sequential steps, theoretically covering 30^3=27,000 unique sequences - and indeed we tested all 27K in a single multiplexed run. Recently, we scaled this to over 1 million sequential-signal combinations across potentially thousands of starting cell contexts. No other effort comes close to this scale of combinatorial perturbations. This massive, context-rich dataset is precisely aligned to the

model's learning objective, and our perturbative approach provides causal information in the non-cancer context that observational cell atlases lack. Cellular Intelligence's capsule system effectively turns data generation into a high-throughput, parallelized endeavor, creating a competitive moat via data complexity that others cannot easily match.

- **Active Learning Loop and Data Augmentation**: Building a predictive model is only half the battle—the other half is using it intelligently to accelerate learning. Cellular Intelligence's platform creates a virtuous cycle: we generate perturbation data, train our model on it, then use that model to identify the most informative next experiments to run. Rather than testing perturbations randomly or exhaustively, the model identifies gaps in its understanding—perhaps a particular signal's effect on a specific cell subtype remains poorly predicted—and prioritizes those experiments. This refinement is also guided by our deep expertise in developmental biology. This targeted approach means each experimental round maximally refines the model's capabilities, dramatically reducing the data needed to achieve broad predictive power. Over time, this self-refining cycle yields a model that not only predicts cellular responses but also efficiently guides its own improvement.

  Critically, we augment this experimental data with publicly available datasets. Cellular Intelligence has developed novel computational techniques to extract signaling information from existing transcriptomic datasets—including cell atlases, differentiation time courses, and published perturbation studies. While these public datasets weren't originally designed to study signaling in our framework, our methods can retroactively infer signal-response relationships from them, effectively multiplying our training data many times over. No other virtual cell effort systematically leverages public data in this way, giving Cellular Intelligence a unique advantage in data efficiency and model generalization.
- **Translational Relevance by Design:** From day one, Cellular Intelligence aligned its data and model to real-world therapeutic contexts. The signals we test are clinically relevant small molecules and growth factors that are GMP-compliant and used in known differentiation or treatment protocols. The timing and dosing regimens we explore mirror those that could feasibly be applied in manufacturing or in the clinic. This means the model's insights map one-to-one with actionable protocols. Competing "virtual cell" projects often use broad functional genomics data (e.g., gene knockouts or overexpression in cancer cell lines) that are valuable for discovery but may not directly translate to, say, a recipe a cell therapy company can implement. In contrast, Cellular Intelligence's foundation model is directly built to predict the effects of signaling, for which a plethora of drugs have been developed, providing a straightforward path to a variety of biomedical applications. Our data of sequential small-molecule perturbations essentially encodes the same "language" the human embryo uses to guide cell fates, giving the model a built-in translational grounding. By coupling the virtual signaling model to tangible protocols, we ensure that advances aren't just academic—they can be immediately plugged into efforts like regenerative medicine manufacturing, drug testing pipelines, or disease modeling experiments. This tight integration of wet-lab relevance is a major distinguishing factor for Cellular Intelligence in the landscape.

Taken together, these differentiators—at the levels of data generation, modeling methodology, AI-driven experimentation, and application focus—create a first-of-its-kind platform. Cellular Intelligence's approach isn't simply to build a larger cell atlas or a clever algorithm in isolation; it is to simultaneously and synergistically develop an unprecedented dataset and a specialized foundation model that together form a self-improving engine for understanding and controlling cell signaling. This is our blueprint for a true foundation model for cell biology.

Cellular Intelligence's approach isn't simply to build a larger cell atlas or a clever algorithm in isolation; it is to simultaneously and synergistically develop an unprecedented dataset and a specialized foundation model that together form a self-improving engine for understanding and controlling cell signaling.

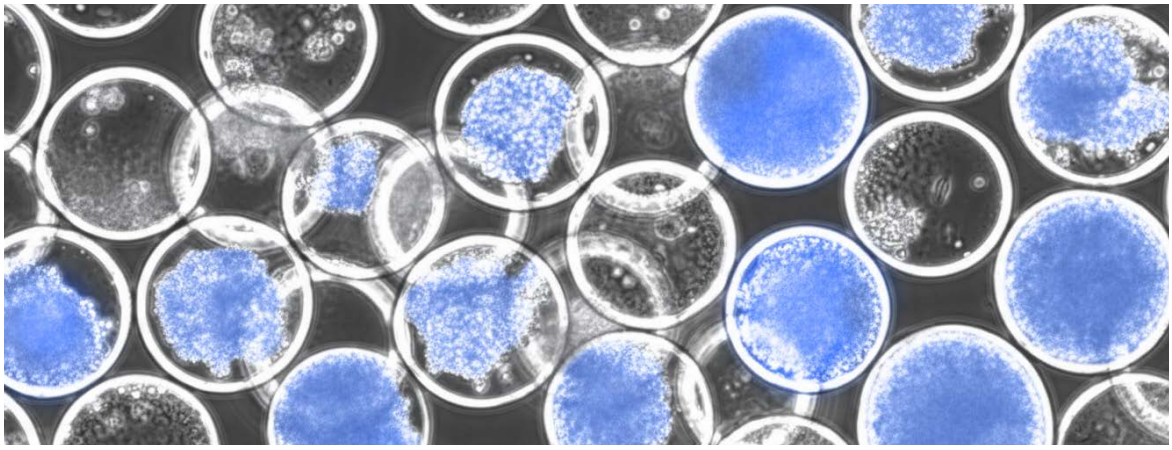
4.

Competitive Landscape: Other Approaches and How We Differ

Many players in the field are advancing complementary components of a broader "virtual cell" vision. Efforts by the Chan Zuckerberg Initiative, Genentech/Roche, Xaira Therapeutics, Tahoe Therapeutics, and the Arc Institute are focusing on mapping what a cell is (molecular profiling), where it is (spatial profiling), or how it is wired internally (genetic perturbations). Cellular Intelligence focuses on a distinct challenge: understanding the cell's built in control systems (signaling), allowing us to predictably guide cells towards therapeutic outcomes using their natural mechanisms. We use stem cells to capture how signaling works across every human tissue, not just a few artificial cancer cell lines with limited physiological value.

The vision of a "virtual cell" has attracted significant attention from non-profits, big pharma, and academic institutes [Tang, 2025; Rood *et al.*, 2024]. However, "virtual cell" is a broad term. Most current efforts are focused on mapping the static geography of cell types or understanding internal genetic circuitry, and the "purpose" for which such maps are being built is not always clear. Cellular Intelligence's distinct focus is on the *dynamic control* of these cells through external signaling.

Below, we distinguish Cellular Intelligence's approach from other notable initiatives (as of December 2025):



Semi-permeable capsules are a core part of Cellular Intelligence's technology stack, enabling sequential signaling studies at a scale competitors are unable to match. Blue staining marks cells undergoing differentiation down the ectoderm lineage.

CHAN ZUCKERBERG INITIATIVE: THE UNIVERSAL REFERENCE MAP

The Chan Zuckerberg Initiative (CZI) has invested heavily in its biomedical research organization, Biohub, to build a definitive "map" of biology. Through efforts like the CELLxGENE database and models like TranscriptFormer [Pearce *et al.*, 2025], they are creating a universal latent space—essentially a coordinate system that allows scientists to compare gene expression across species and tissues.

The Distinction: CZI's work is foundational but, at least to date, largely **observational**. Their models are excellent at answering *"What is this cell?"* or *"Where does this cell sit on the map?"* However, because they lack large-scale perturbation data, they cannot predict *"Where will this cell go if I treat it with Signal X?"* CZI provides the atlas; Cellular Intelligence provides the navigation system to drive through it. To the extent that CZI is aiming to generate perturbational data these are genetic perturbations in a limited number of contexts, like other efforts described below.

GENENTECH/ROCHE: THE SEARCH ENGINE FOR BIOLOGY

Genentech has pioneered the use of AI to query cellular data, developing models like SCIMilarity to find identifying patterns across cell types and diseases [Heimberg *et al.*, 2025]. Their goal is primarily hypothesis generation—identifying analogs of a disease state in a different context to uncover new drug targets.

The Distinction: Like CZI, this approach relies heavily on searching existing atlases rather than developing a roadmap for nimbly navigating that space. While Genentech utilizes some perturbation screens, they generally lack the massive, time-series signaling data required to learn the causal relationships required for protocol design. Their tools are designed to help scientists model disease; Cellular Intelligence's platform is designed to *engineer* cells.

XAIRA THERAPEUTICS: MAPPING THE WIRING VS. OPERATING THE CONTROLS

Xaira is a major player tackling the virtual cell from a gene-centric angle. With datasets like X-Atlas/Orion, they are performing massive Perturb-seq experiments, systematically knocking out genes to see how the cell recovers [Huang *et al.*, 2025].

The Distinction: This is a complementary but fundamentally different problem. If you view the cell as a complex machine, Xaira is mapping the internal **wiring** (asking: *"What breaks if I cut this wire?"*). Cellular Intelligence is learning to operate the **control panel** (asking: *"What happens if I turn this knob?"*). While genetic perturbation is powerful for identifying drug targets, Cellular Intelligence's focus on extrinsic signaling is the direct path to regenerative medicine and cell therapy manufacturing, where the goal is to guide cells to specific fates without genetically altering them, as well as drug target identification.

TAHOE THERAPEUTICS AND THE ARC INSTITUTE: HIGH PERTURBATION, LOW CONTEXT

The Arc Institute's recent release of the STATE model, developed using the Tahoe-100M dataset—covering 100 million cells and roughly 1,100 drugs—is the most comparable effort to date in terms of scale [Adduri *et al.*, 2025]. It represents a massive achievement in perturbation biology.

The Distinction: The critical limitation of the Tahoe dataset, and others like it, is context poverty. Standard screening paradigms maximize the number of drugs tested, but they test them on a very small number of contexts (usually cancer cell lines). However, the central challenge of signaling is that it is context-dependent—a signal acting on a liver cell does not do the same thing when acting on a neuron. Cellular Intelligence's capsule platform allows us to generate tens of thousands of distinct biological contexts (intermediate developmental states) via human pluripotent stem cells (iPSCs). While Tahoe tests many drugs on a few cell types, Cellular Intelligence tests the fundamental signals across the entire developmental tree representing all human cell lineages. This diversity of context is the mathematical prerequisite for a model that can truly generalize. Moreover, Tahoe's approach, while based on chemical perturbations, is restricted to single applications of these treatments, barely scratching the surface of the exponentially large number of orders and doses.

5.

Applications and Impact

Traditionally, discovering a new cell differentiation protocol or drug target relies on luck and brute-force labor. Cellular Intelligence replaces luck with logic. By predicting how cells respond to signals before experiments begin—whether rationally designing a protocol for regenerative medicine or searching for a drug to rescue a genetic defect—our model replaces blind experimentation with calculated prediction, de-risking the most expensive stages of drug and therapy development.

A successful foundation model for virtual cell signaling will be transformative, turning what used to be months-long lab endeavors into *in silico* design problems solvable in days. Here we highlight a few high-impact applications enabled by Cellular Intelligence's platform:

- **Rational Design of Cell Differentiation Protocols:** Perhaps the most immediate application is in regenerative medicine and cell therapy manufacturing. Today, developing a protocol to differentiate stem cells into a target cell type (e.g., pancreatic beta cells or dopaminergic neurons) often takes years of trial-and-error tweaking of growth factor cocktails and timing. With a virtual signaling model, this process becomes systematic. The model can screen countless candidate protocols *in silico*, predicting which sequence of signals will yield the highest purity of the desired cell type. This enables **rational protocol design**: instead of blind experimentation, researchers can approach protocol optimization like engineering—iterating on simulations first, then only testing the most promising conditions at the bench. We estimate we can cut protocol development time from years to months while achieving cells that more closely resemble their *in vivo* counterparts.
- **Context-Specific Drug Response Prediction:** The virtual signaling model functions as a powerful tool for predicting drug effects in specific cellular contexts. Many drugs behave differently in different cell types or disease states—a therapy might cure inflammation in one tissue but be toxic in another. Using our model, scientists can simulate how, for instance, an immune cell from a patient with an autoimmune disorder will respond to a cytokine inhibitor compared to a healthy donor. Such simulations highlight context-specific efficacy or toxicity early in the R&D process. Moreover, because our model captures sequential effects, it can predict outcomes of **drug combinations or scheduling** (e.g., finding synergistic timing). This enables virtual clinical trials on patient-derived cell models to narrow down to the most promising interventions before they reach the clinic.
- **Genetic Disease Modeling and Phenotypic Rescue:** Our platform offers a direct route to therapy for genetic disorders by treating the cell's signaling network as a correctable circuit. By generating data from iPSCs derived from patients with specific genetic mutations, we can train our model to understand exactly how a genotype alters signaling logic. Crucially, the model can then be used in "reverse engineering" mode: identifying specific **small-molecule signals that compensate for the genetic defect**. For example, if a mutation dampens a critical pathway, the model might identify a downstream signal or a parallel pathway that can be stimulated to restore the healthy phenotype. This approach allows us to discover small-molecule treatments for genetic conditions—offering a scalable, accessible alternative to complex gene therapies.

A successful foundation model for virtual cell signaling will be transformative, turning what used to be months-long lab endeavors into *in silico* design problems solvable in days.

- **Interpreting Disease as Signaling Network Failure:** Many diseases—from cancer to diabetes—can be traced to malfunctions in cellular signaling pathways. Our model provides a new lens on these pathologies: it can simulate how cells in a disease state respond abnormally to signals and help pinpoint *what* in the signaling network is broken. By comparing model predictions between healthy and diseased cell states, researchers can ask, *"Which signal does the diseased cell fail to handle correctly?"* If the model shows that a certain developmental signal no longer triggers the expected gene expression change, that suggests a specific pathway defect. In this way, the model acts as a **digital twin for diseased cells**, revealing failure points in their internal circuitry and guiding the search for therapeutic targets that restore proper signaling.
- **Uncovering Hidden Biology and New Pathways:** By systematically exploring such a vast space of conditions, our platform is naturally poised to make biological discoveries. In traditional experiments, scientists often test hypothesis-driven combinations of signals—but nature's combinatorial complexity means many interactions remain unknown. Cellular Intelligence's multiplexed approach casts a wide net, allowing us to flag serendipitous findings of "cryptic" pathways or novel cell states. For example, in our 27,000-condition screen, we unexpectedly generated **notochord-like cells** (a rare embryonic type) from pluripotent stem cells (only recently discovered in 2024 [Rito *et al.*, 2024]). The model, having seen this, can generalize and teach us the "recipe" for rare biology. Moreover, because our model incorporates literature-based priors, it serves as a hypothesis generator: if it predicts a strong outcome that isn't documented in literature, that is a cue to investigate "white spots on the map" of cell signaling.

By enabling these applications, a foundation model for cell signaling stands to accelerate progress across biomedicine. It provides a unifying platform where biologists, bioengineers, and clinicians can ask "what if" questions - What if I add this factor at hour 24? What if I inhibit this pathway in a cell with this mutation? - and get immediate, educated answers to guide their next steps. The end result will be faster development of therapies (both cell-based and traditional drugs), more precise interventions with fewer failures, and a deeper understanding of life's fundamental processes.

6.

Progress and Validation to Date

Cellular Intelligence has successfully executed some of the largest combinatorial signaling screens in history, scaling to over 1 million unique experimental conditions. We have proven that our platform can already generate a vast diversity of cell types—from muscle progenitors to neurons to blood vessel cells—in a single run. We have now reached a critical inflection point: our data engine is operating at the scale required to unlock "scaling laws" in biology, where massive datasets yield qualitatively more powerful and generalizable models.

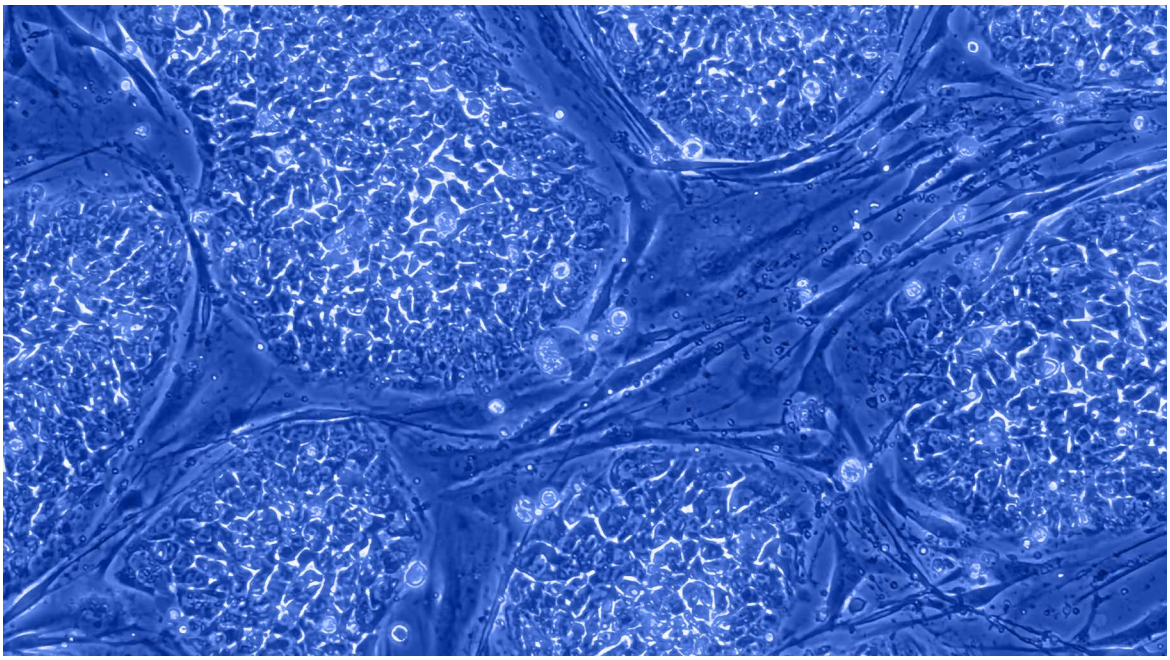
Cellular Intelligence's approach is bold, but it is grounded in concrete progress from our ongoing research. We have already achieved significant milestones that validate both the feasibility of our platform and its early predictive capabilities:

- **Unprecedented Data Scale and Diversity:** We have successfully built the world's first dataset of sequential cell-signaling responses at massive scale. In 2024, our team completed a pilot split-and-pool screen of 27,000 distinct conditions (3 sequential signal combinations applied to human iPSC-derived cells). Each condition is a unique sequence, like "Signal A (3 days) → Signal B (3 days) → Signal C (3 days)", and we profiled the cells after the first signal and the final signal via single-cell RNA sequencing. This experiment alone produced an extraordinary variety of outcomes. Remarkably, after just 3 days (one step), cells in capsules had already diverged into all major embryonic lineages (ectoderm, mesoderm, endoderm). By day 9 (at which point each capsule has seen a combination of three successive signals), we observed a rich mosaic of cell types from those lineages—including neural precursors, muscle progenitors, gut epithelial cells, blood vessel cells, and more. This comprehensive coverage is a strong validation that our capsule perturbation engine can generate the full spectrum of cell fates in a controlled way; these include the same notochordal cells we first detected in our 27K experiment. These results confirm that our method doesn't just produce "some cells"—it produces almost any cell, given the right signals, and our data captures those mappings.

- **Scaling to 1 Million Conditions:** Building on the 27K-condition success, we immediately set our sights on scaling up. In mid 2025, we executed experiments that exposed cells to on the order of 10^6 sequences of perturbations. Initial analysis of the 1M-condition run (which subsampled cells and capsules at multiple time points) confirms the trend from the 27K screen: cells exposed to this vast range of signal combinations still populate all expected major cell lineages—demonstrating scalability without loss of biological breadth. We also see finer granularity in some cases: for instance, where we had "muscle precursors" before, we now see separate clusters for skeletal vs. cardiac muscle lineage, etc. This gives us confidence that increasing data density yields increasing resolution, which will in turn power the training of increasingly predictive models. It's important to note that simply generating data is not our goal—the goal is predictive power. Our model's performance (predictive accuracy) is improving as the dataset grows, a hallmark of a true foundation model where more data directly translates into better generalization.

- **Technical Validation of Platform Components:** Alongside biological results, we have achieved numerous technical milestones that de-risk the platform. We have confirmed that our capsule barcoding method is reliable and scalable. We have shown that cells can remain healthy in capsules through multiple split-and-pool steps, maintaining viability and proliferative capacity. Our single-cell sequencing pipelines have been optimized to handle extremely large libraries, and we developed cloud-based tools to process millions of single-cell profiles per run. We have also begun public data augmentation: by pre-training on an embryo atlas (public single-cell RNA-seq from mammalian development) and fine-tuning on our perturbation data, we expect to improve model performance [Xu *et al.*, 2023]. All these pieces—data generation, data accuracy, model training, and early application—provide a solid foundation as we move from prototype to full-scale platform.

In summary, within roughly two years, Cellular Intelligence has gone from concept to demonstrating *in vitro* and *in silico* proof-of-concept results. We've generated orders-of-magnitude more relevant data than previously available, and we've shown that this data can power AI models to make non-trivial predictions in cell biology. Most importantly, we have validated that biology doesn't "break" at scale: cells in our complex experiments still yield biologically meaningful states (not just random or dead cells). This de-risks the central hypothesis that a general signaling model can be learned from these data. The progress to date sets the stage for the next phase: scaling up further and delivering on increasingly ambitious milestones of predictive power and real-world impact.



Human stem cells enable broad cell-fate coverage. Their capacity to differentiate into diverse lineages supports comprehensive mapping of cell fates.

7.

Architecture and Training Approach

We are adapting the same "Transformer" architecture that powers ChatGPT, but instead of predicting the next word in a sentence, our model predicts the next state of a human cell. By training on our massive proprietary perturbation dataset alongside public cell atlases, we teach the model to generalize across contexts. This allows us to learn the fundamental "grammar" of cellular decision-making, transforming biology into a computable problem.

Building a foundation model for cell signaling requires more than just data; it requires a computational framework capable of learning the complex, non-linear rules of biology. At Cellular Intelligence, we approach this as a machine learning problem defined by high-dimensional states and causal signaling inputs. Below, we outline our modeling philosophy, our strategy for architectural exploration, and the data infrastructure we have built to power this engine.

THE CORE LEARNING PROBLEM

Fundamentally, our model is trained to approximate a transition function for cellular state. In machine learning terms, our primary training examples take the following form:

input transcriptome + input signal → output transcriptome

The challenge in learning this function is context dependence: the same signal (e.g., Wnt activation) will produce a drastically different output depending on the input transcriptome.

Because our capsule platform generates massive variation across both the input transcriptome (thousands of intermediate developmental states) and the input signal (dosages and combinations), we provide the model with the necessary volume of examples to learn this context-specific logic. Unlike standard regression models, our goal is to learn a generalized representation of how signals perturb cell state manifolds.

ARCHITECTURAL EXPLORATION

We are currently exploring and benchmarking several state-of-the-art neural network architectures suited for this high-dimensional, causal modeling:

- **Foundation Embeddings:** A prerequisite for our model is a robust "map" of cell space. We are developing embedding models that co-embed our proprietary perturbation data alongside massive public reference atlases. This allows us to represent any given cell not as a list of 20,000 genes, but as a coordinate in a biologically meaningful latent space. By grounding our model in this universal reference frame, we ensure that our predictions obey biological constraints.

- **Transformer-Based Context Modeling:** Transformers have revolutionized language processing by handling the context of words in a sentence. We are applying similar architectures to biology, where the "context" is the current state of the cell's gene regulatory network. We are exploring transformer backbones that can attend to the specific combination of genes active in a cell to determine how to weight the incoming signal, effectively "reading" the cell's state to predict its reaction.
- **Temporal Dynamics and Neural ODEs:** Biology happens in continuous time, not just in discrete steps. To capture the dynamics of differentiation, we are investigating **Neural Ordinary Differential Equations (Neural ODEs)**. These architectures allow us to model the trajectory of a cell as a continuous flow, enabling us to predict cell states at time points we haven't explicitly sampled and to model the kinetics of how fast a cell transitions from one state to another.

DATA AUGMENTATION VIA PUBLIC KNOWLEDGE

While our proprietary data is the gold standard for causal ground truth, we amplify its power by integrating publicly available data. Cellular Intelligence's computational team has developed novel methods to "retroactively" extract signaling information from existing datasets.

By analyzing time-course data from published embryology and differentiation papers, we can infer pairs of (State A) → (State B) transitions. Even if the original authors did not explicitly structure their data for machine learning, our pipelines can ingest these trajectories to generate millions of additional synthetic training examples. This allows our model to learn from the collective history of biological research, using our high-fidelity proprietary data to fine-tune and validate the broad patterns learned from the public domain.

TECHNICAL INFRASTRUCTURE

To support this massive undertaking, we have established a cloud-native data infrastructure designed for scale.

- **Ingestion Pipelines:** We have built automated pipelines capable of ingesting and normalizing heterogeneous data sources—ranging from internal high-throughput sequencing runs to unstructured public datasets.
- **Scalable Compute:** Our analysis platform runs on distributed GPU-based cloud computing clusters, allowing us to train large-scale models on high-dimensional single-cell data without memory bottlenecks.
- **Iterative Loop:** Our infrastructure is designed for a tight loop between wet lab and dry lab. Data from the capsule platform is automatically processed, quality controlled, and fed into our modeling environment, allowing our computational biologists to rapidly iterate on model architectures as new datasets are generated.

By combining a rigorous mathematical formulation of the signaling problem with flexible, scalable architecture exploration, Cellular Intelligence is laying the groundwork for the first true simulation engine for cellular engineering.

8.

Roadmap and Milestones

Cellular Intelligence is executing on a clear roadmap to develop and deploy our foundation model for cell signaling. We are moving systematically: first predicting single signals, then mastering combinatorial complexity, and finally achieving a universal model capable of simulating cellular behavior over continuous biological time. This trajectory moves the field from "discovery by chance" to "discovery by design," creating the infrastructure needed to rapidly deliver clinical solutions.

ALPHA PHASE: ESTABLISHING THE PREDICTIVE BASELINE

CURRENT FOCUS

In the Alpha phase, our primary objective is to demonstrate the fundamental ability to predict the transcriptomic state change induced by a single signal. While simple in principle, achieving high accuracy on this task across varying cellular contexts would represent a significant leap forward for the field.

KEY GOALS

- **Single-Signal Prediction:** We aim to achieve **80% accuracy** in predicting the transcriptomic shift of cells treated with any of **10 core signaling pathways** (e.g., Wnt, BMP, FGF, Nodal).
- **Contextual Generalization:** Crucially, this accuracy must hold across a diverse set of starting iPSC-derived contexts, not just a few cell lines.
- **Fixed-Interval Training:** Initial training will focus on fixed time intervals (e.g., predicting the state at t=24 hours given state at t=0).

STATUS

We are currently generating the high-volume perturbation data required to train this initial model. We view the successful completion of Alpha as the core "proof of principle" that context-dependent signaling is a learnable function.

BETA PHASE: HIGH-FIDELITY & COMBINATORIAL EXPANSION

TARGET: SURPASSING STATE-OF-THE-ART

In the Beta phase, we aim to refine the model's precision and expand its scope to handle the complexity of real-world biological environments. At this stage, we expect the model to outperform existing observational atlases in predicting cell fate.

KEY GOALS

- **90% Accuracy Threshold:** We aim to increase our prediction accuracy to **90%** across an expanded library of signaling pathways and cellular contexts.
- **Public Data Augmentation:** We will integrate massive amounts of publicly available transcriptomic data (e.g., cell atlases, perturbation screens) to improve the model's generalization capabilities, particularly for cell states that are under-represented in our internal dataset.
- **Simultaneous Signal Integration:** Moving beyond single signals, Beta will address **combinatorial signaling**—predicting the behavior of cells when multiple signals are applied simultaneously (e.g., Wnt + FGF together), which often yields non-linear synergistic effects.
- **Spatial Context:** We will begin incorporating **spatial transcriptomics data**. This allows the model to account for cell-cell communication and local environmental factors (paracrine signaling) that influence cell fate decisions.

V1 PHASE: THE UNIVERSAL SIMULATION ENGINE

TARGET: CONTINUOUS TIME & CHEMICAL UNIVERSALITY

The V1 release represents the full realization of the virtual cell platform: a dynamic, continuous simulation engine capable of bridging the gap between protein signals and small molecule drugs.

KEY GOALS

- **Continuous Time Modeling (Neural ODEs):** We will transition from fixed-step predictions to continuous-time modeling using **Neural Ordinary Differential Equations (Neural ODEs)**. This will allow us to query the cell state at *any* time point (e.g., t=12.5 hours), enabling the precise optimization of dosing schedules and pulse durations.
- **Chemical Co-Embedding:** We will develop a **chemical co-embedding space**, mapping small molecules to the signaling pathways they modulate. This allows the model to predict how a specific chemical inhibitor or agonist affects a pathway, effectively translating "protein logic" into "drug logic." This is a critical feature for pharmaceutical applications, enabling the *in silico* replacement of expensive growth factors with easily synthesized small molecules.
- **Global Generalization:** By V1, the model will leverage a fully integrated dataset of internal and external data, aiming for high predictive accuracy across all major germ layers and standard therapeutic targets.

Beyond V1, our roadmap extends to a Vision 2.0 where the model becomes even more powerful—potentially incorporating patient-specific genetic backgrounds (to handle donor variability in cell responses), modeling cell-cell interactions (like adding immune cells into the mix for immunotherapy design), and integrating other modalities (such as signaling protein levels, epigenetic states, etc. for even richer context). The long-term vision is that Cellular Intelligence's foundation model evolves into the "operating system" for cellular engineering, supporting applications we haven't yet imagined. The milestones listed for Alpha, Beta, V1 are how we get there step by step, de-risking along the way and delivering value at each stage.

The long-term vision is that Cellular Intelligence's foundation model evolves into the "operating system" for cellular engineering, supporting applications we haven't yet imagined.

9.

Conclusion: From Observation to Engineering

Cellular Intelligence replaces decades of empirical guesswork with a predictive engine for therapeutic innovation.

The transition of biology from an empirical science to an engineering discipline is the defining opportunity of our time. At Cellular Intelligence, we are building the infrastructure to drive this transformation in the domain of cell signaling. By combining the latest advances in AI with an unprecedented experimental engine, we are systematically decoding the "language" of cell fate: the grammar of signals and contexts that determine whether a cell becomes a neuron or a muscle, whether it regenerates tissue or succumbs to disease.

The implications of mastering this language are profound. We envision a future where the question *"How do I manufacture this tissue?"* or *"How do I rescue this diseased cell?"* is not answered by years of trial-and-error at the bench, but by a query to a foundation model that returns a precise, actionable protocol. This is the future Cellular Intelligence is constructing.

THE HUMAN IMPERATIVE

Throughout this white paper, we have focused on the technical hurdles of data and modeling, but our motivation remains deeply human. The urgency to move beyond stochastic experimentation is driven by patient need. Patients waiting for organ transplants, individuals with genetic disorders, and the unfulfilled promise of personalized medicine cannot afford another decade of manual optimization. Cellular Intelligence's platform is a technological response to this biological urgency. We leverage big data and computation not for their own sake, but to create a reliable engine for therapeutic innovation.

A DISTINCT PATH FORWARD

Groundbreaking initiatives by CZI, Genentech, Xaira, Tahoe, and the Arc Institute have laid the foundation by mapping the geography of the cell. Cellular Intelligence distinguishes itself by drilling into the **dynamic, context-rich, and sequential** aspects of biology that these atlases do not address. We are not just mapping where cells are; we are building the navigation system to guide them to where they need to be.

We often draw an analogy to the revolution in protein structure: What AlphaFold achieved for protein folding—turning a physical mystery into a solvable computational problem—Cellular Intelligence aims to achieve for cellular signaling. However, we intend to go a step further. While prediction is powerful, our ultimate goal is **design**: the active intervention in cellular logic to achieve specific therapeutic outcomes.

JOIN THE REVOLUTION

"Towards a Foundation Model for Virtual Cell Signaling" is more than a white paper; it is a roadmap for a new paradigm in medicine.

- **For Strategic Investors:** This represents a shift from asset-heavy risk to platform-based scalability, offering a generator for novel therapies and high-value intellectual property.
- **For Biopharma Partners:** It offers a chance to supercharge R&D, replacing empirical guesswork with data-driven precision to de-risk drug discovery and cell therapy manufacturing.
- **For Scientists:** It promises a new generation of tools that expand the experimental horizon, allowing us to probe the "why" and "how" of cellular behavior with unprecedented clarity.

We are forging a path where data, computation, and biology converge to enable mastery over cell signaling. The data is flowing, the infrastructure is built, and the vision is clear. Cellular Intelligence invites you to be part of this revolution—a future where we do not just observe biology's complexity, but intelligently shape it for the betterment of humanity.

10.

References

Adduri, A. K., *et al.* (2025). Predicting cellular responses to perturbation across diverse contexts with STATE. *bioRxiv.* [Preprint]. (Corresponds to the Tahoe/Arc Institute STATE model). https://doi.org/10.1101/2025.06.26.661135

Tang, L. (2025). The virtual cell. *Nature Methods*, (22)2493. https://doi.org/10.1038/s41592-025-02951-5.

Heimberg, G., *et al.* (2025). A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638, 1085–1094. https://doi.org/10.1038/s41586-024-08411-y

Huang, A. C., *et al.* (2025). X-Atlas/Orion: Genome-wide Perturb-seq Datasets via a Scalable Fix-Cryopreserve Platform for Training Dose-Dependent Biological Foundation Models. *bioRxiv* 2025.06.11.659105. https://doi.org/10.1101/2025.06.11.659105

Pearce, J. D., *et al.* (2025). A Cross-Species Generative Cell Atlas Across 1.5 Billion Years of Evolution: The TranscriptFormer Single-cell Model. *bioRxiv* 2025.04.25.650731. https://doi.org/10.1101/2025.04.25.650731

Rood, J., *et al.* (2024). Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas. *Cell*, 187(17), 4520-4545. https://doi.org/10.1016/j.cell.2024.07.035

Rito, T. *et al.* (2025). Timely TGFβ signalling inhibition induces notochord. *Nature*, 637, 673–682. https://doi.org/10.1038/s41586-024-08332-w

Wagner, A., *et al.* (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11), 1145–1160. https://doi.org/10.1038/nbt.3711

Xu, Y., *et al.* (2023). A single-cell transcriptome atlas profiles early organogenesis in human embryos. *Nature Cell Biology*, 25, 604–615. https://doi.org/10.1038/s41556-023-01108-w

CELLULAR INTELLIGENCE

CELLULARINTELLIGENCE.COM | HELLO@CELLULARINTELLIGENCE.COM

Biology, Designed.