

# Virtual Cells: Predict, Explain, Discover

Emmanuel Noutahi\*, Jason Hartford\*, Prudencio Tossou, Shawn Whitfield, Alisandra K. Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, Emmanuel Bengio, Dominique Beaini, Christopher Gibson, Daniel Cohen, Berton Earnshaw\*

Valence Labs, Recursion

\*Correspondence: {firstname}@valencelabs.com

Drug discovery is fundamentally a process of inferring the effects of treatments on patients, and would therefore benefit immensely from computational models that can reliably simulate patient responses, enabling researchers to generate and test large numbers of therapeutic hypotheses safely and economically before initiating costly clinical trials. Even a more specific model that predicts the functional response of cells to a wide range of perturbations would be tremendously valuable for discovering safe and effective treatments that successfully translate to the clinic. Creating such *virtual cells* has long been a goal of the computational research community that unfortunately remains unachieved given the daunting complexity and scale of cellular biology. Nevertheless, recent advances in AI, computing power, lab automation, and high-throughput cellular profiling provide new opportunities for reaching this goal. In this perspective, we present a vision for developing and evaluating virtual cells that builds on our experience at Recursion. We argue that in order to be a useful tool to *discover* novel biology, virtual cells must accurately *predict* the functional response of a cell to perturbations and *explain* how the predicted response is a consequence of modifications to key biomolecular interactions. We then introduce key principles for designing therapeutically-relevant virtual cells, describe a lab-in-the-loop approach for generating novel insights with them, and advocate for biologically-grounded benchmarks to guide virtual cell development. Finally, we make the case that our approach to virtual cells provides a useful framework for building other models at higher levels of organization, including virtual patients. We hope that these directions prove useful to the research community in developing virtual models optimized for positive impact on drug discovery outcomes.

## 1 Introduction

Drug discovery is fundamentally a process of accurately inferring the effects of treatments on patients. Unfortunately, it is notoriously costly and riddled with failure (Wong et al., 2019; Jones & Wilksdon, 2018; DiMasi et al., 2016; Paul et al., 2010). Despite decades of innovations, for every ten drugs that enter clinical trials today, roughly nine of those will fail to receive approval, representing unfortunate delays in addressing patient needs, substantial losses in R&D investment, and a significant deficit in our collective understanding of human physiology and pathology. Nevertheless, the impact of each approved therapy on the lives of patients, particularly those addressing unmet need, is hard to overstate, thus any approach that meaningfully improves our ability to correctly predict the effect of treatments in patients would be of immense value to both patients and drug discoverers alike.

One such approach is the computational simulation of therapeutic interventions in *virtual patients*, or mechanistic models accounting for the physiological factors necessary to accurately infer patient-level response to treatments. Virtual patients could revolutionize drug discovery by enabling researchers to generate and test large numbers of therapeutic hypotheses safely and economically before initiating costly clinical trials. However, though simulation has already revolutionized a number of industries (Winsberg, 2019; Singh et al., 2022), examples of practical and effective simulation in drug discovery are rare due to the challenges inherent in modeling the scale and complexity of biological systems (Ideker et al., 2001; Goldberg et al., 2018; Georgouli et al., 2023). Even the simulation of a single prokaryotic cell is daunting (Karr et al., 2012), and simulating the full complexity of a eukaryotic cell lies beyond current capabilities (Georgouli et al., 2023). Nevertheless, the ability to faithfully simulate the effect of therapeutic interventions at any level of biological organization—cell, tissue, organ, patient—in a corresponding *virtual model* has the potential to significantly improve drug discovery outcomes.

Box 1: Virtual Cells: Predict, Explain, Discover

**Predict** the functional response of cells to perturbations across diverse biological contexts, timepoints and modalities. This includes modeling gene expression, morphology, protein activity, and other phenotypic changes under genetic or chemical interventions.

**Explain** these responses by identifying key biomolecular interactions, causal pathways, and context-dependent regulatory mechanisms. Correct explanations support predictions by enabling both generalization beyond the training data and reasoning about counterfactuals and the response of biological systems at higher levels of organization.

**Discover** new biological insights and actionable therapeutic hypotheses through lab-in-the-loop experimentation, using the virtual cell as a world model for systematic hypothesis generation, testing, and refinement.

### 1.1 The Predict-Explain-Discover capabilities of virtual models

What exactly makes virtual models so potentially valuable for drug discovery? They offer the ability to accurately:

1. *predict* the effects of interventions on the model system,
2. *explain* the predicted response in terms of one or more changes to supporting mechanisms, and
3. *discover* novel insights by generating and testing therapeutic hypotheses.

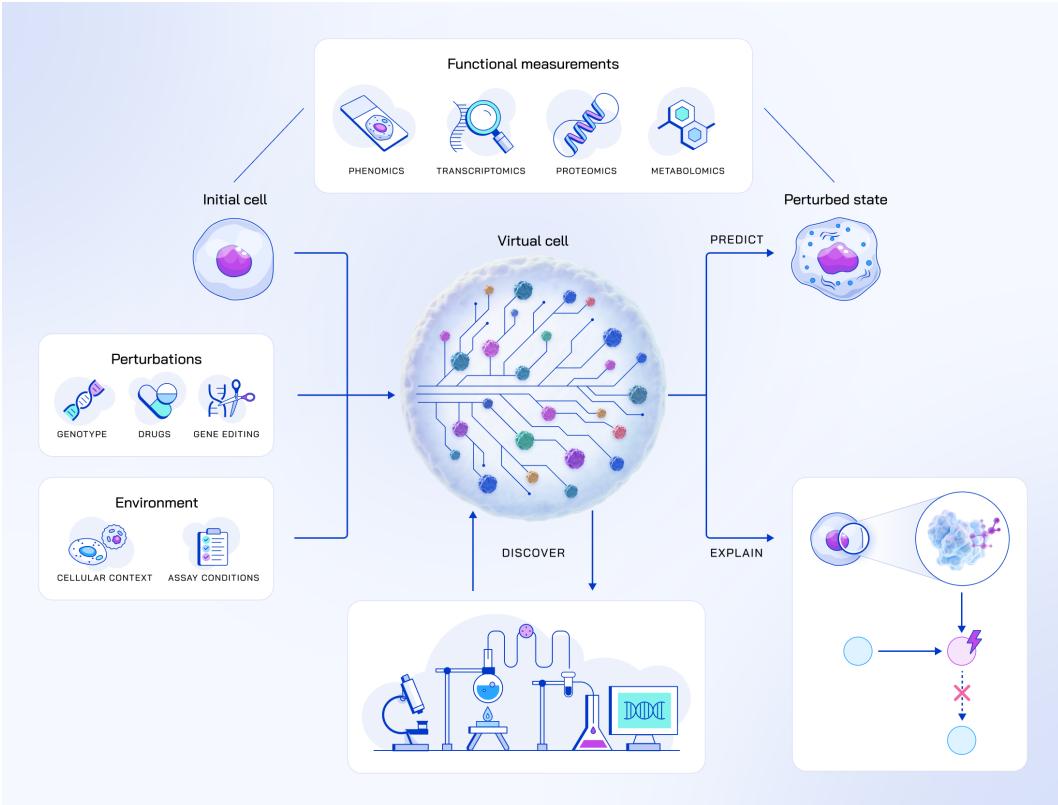
To better understand what we mean, we give two examples using well-known cancer treatments; these are only meant to motivate the concepts and not to imply that these would represent novel discoveries if made today:

**Vorinostat.** A virtual cell could *predict* that treatment with Vorinostat up-regulates tumor suppressor genes, and *explain* this effect by identifying inhibition of histone deacetylase (HDAC) enzymes as the underlying mechanism. Based on this understanding, the model could then *discover* biomarkers of response, or synergistic combinations such as with DNA-damaging agents like cisplatin, by revealing how chromatin decondensation increases sensitivity to genotoxic stress.

**Pembrolizumab.** For a virtual *patient* model that accounts for mechanisms across different cells, tissues and organs, it could *predict* that Pembrolizumab reduces tumor burden in various cancers, and *explain* this effect by simulating how the drug blocks the Programmed Cell Death Protein 1 (PD-1) immune checkpoint to restore T-cell activity. Building on this understanding, the model could then *discover* combination strategies to enhance efficacy across patient subgroups, as well as approaches to overcome emerging resistance mechanisms.

Throughout this paper, we will refer to these three capabilities as the *Predict-Explain-Discover*, or *P-E-D*, capabilities of virtual models, and claim that it is precisely the ability of virtual models to accurately *predict* outcomes and *explain* them mechanistically that would make them powerful tools to *discover* novel therapeutic insights. As our prototypical example of a virtual model, we illustrate the *P-E-D* capabilities for virtual cells in Figure 1 and describe these capabilities in more detail in Box 1.

We wish to further clarify what we mean by these terms. By *predict* we refer to the task of a virtual model to predict the effect of a perturbation on a biological system, typically a cell but applicable at all levels of organization (see Section 1.1). In causal terms, this is equivalent to estimating (the distribution of) outcomes under intervention (Pearl, 2009). We use *explain* in a broader, biologically-grounded sense than is common in machine learning, where explainability usually refers to post-hoc techniques (e.g., saliency maps, attention scores, feature attributions) used to interpret model behavior. In contrast, *explanation* in our context refers to structured, testable accounts of how a given perturbation leads to a specific biological response. These accounts must be meaningful in the language of biology and useful for hypothesis generation and falsification. See Box 2 for more on biological explanations.



**Figure 1: Virtual cells: predict, explain, discover.** In order to address critical issues holding back progress in drug discovery, virtual cells should *predict* the functional response of cells (as measured in phenomics, transcriptomics, proteomics, metabolomics, etc.) to perturbations across diverse cellular contexts and assay conditions, and *explain* these responses as modifications to key biomolecular interactions, using techniques like causal learning on interventional data, physics-informed structural predictions, and targeted molecular dynamics simulations. By offering a focused mechanistic understanding of their behavior, such virtual cells can *discover* novel biology by efficiently generating and testing large numbers of promising hypotheses before initiating costly clinical trials, offering a modern approach to rational drug design that holistically optimizes cell state rather than biomolecular interactions only.

## 1.2 Virtual models without fully mechanistic simulation

Given the current difficulties building fully mechanistic virtual model simulators, the question naturally arises: can we build these models without resorting to fully mechanistic simulation? We argue that four recent advances put this objective within reach today, particularly for virtual cells:

1. modern AI and machine learning (**AI/ML**),
2. modern **compute infrastructure**,
3. **automated labs** for high-throughput cellular data generation, and
4. the proliferation of **cellular omics datasets**.

We briefly describe each of these advances below:

**AI/ML.** While traditional computational drug discovery techniques have struggled to deal with biological complexity (Sams-Dodd, 2005; Swinney & Anthony, 2011; Waring et al., 2015), recent advancements in AI/ML have enabled the training of powerful models capable of extracting patterns from high-dimensional biological datasets and predicting complex biological phenomena (Zhang et al., 2025b; Wang et al., 2023; Sadybekov & Katritch, 2023; Vamathevan et al., 2019). Modern AI/ML also provides techniques for integrating multimodal measurements that are essential for capturing

## Box 2: Explanations in biology

Following traditions in the philosophy of biology (Mitchell, 2003; Green, 2016), we adopt an *explanatory pluralism* that recognizes multiple, complementary forms of explanation:

**Mechanistic explanations** describe how organized entities and activities (e.g., proteins, interactions, molecular pathways) give rise to observed phenomena via spatially and temporally structured interactions (Machamer et al., 2000).

**Semantic explanations** interpret biomolecules (e.g., DNA, RNA, peptides) as symbols in a code-like system, where context-sensitive meaning emerges from evolutionary, developmental, or functional constraints (Zámečník, 2021).

**Structural-statistical explanations** draw on abstract representations (e.g., embeddings, graphs, statistical dependencies) that capture functional constraints or emergent regularities, even in the absence of complete mechanistic detail (Ross, 2021).

Each of these explanatory types can serve different scientific goals. For instance, predicting that two gene knockouts produce similar phenotypes may be explained mechanistically (e.g., their protein products form a complex), semantically (e.g., they encode functionally redundant roles), or structurally (e.g., they co-occur in latent pathway embeddings). Importantly, explanations must remain falsifiable, biologically actionable, and aligned with the epistemic norms of biology, including causal relationships, evolutionary constraints, and structural integrity.

This pluralistic stance allows virtual cells to operate not only as predictive models but as explanatory tools: bridging molecular perturbations with observable functional responses in ways that are interpretable, mechanistically plausible and useful for therapeutic discovery.

a complete view of cells (Nam et al., 2024; Stahlschmidt et al., 2022; Li et al., 2018; Picard et al., 2021). Of course, scientific progress depends not only on accurate predictions but on understanding the mechanisms that give rise to these predictions, and here we can turn to modern AI/ML again for methods that leverage interventional datasets to train models that directly predict causal effects and infer causal mechanisms (Schölkopf et al., 2021; Pearl, 2009; Hill et al., 2016; Sachs et al., 2005).

**Compute infrastructure.** Recent improvements in the capability and accessibility of computational infrastructure, driven by dedicated on-premise installations, scalable cloud platforms, specialized hardware accelerators (e.g., GPUs, TPUs), and high-bandwidth networking, supports the training of powerful AI/ML models on high-dimensional biological datasets (Lee & Amaro, 2018; Zhou et al., 2024).

**Automated labs.** Advances in automation technology, including robotics for plate and liquid handling and high-throughput microscopy, make possible the building of sophisticated automated labs that generate biological data at the scale, quality, and diversity required for training useful AI/ML models. For example, the automated phenomics lab we operate at Recursion alone is capable of obtaining microscopy readouts from ~2.2 million samples per week.

**Cellular omics datasets.** We have recently witnessed a rapid expansion in the availability of both public and private cellular omics datasets—genomic, transcriptomic, proteomic, metabolomic, phenomic—as well as improved techniques for data integration and curation, providing the raw materials necessary for training powerful AI/ML models (Fay et al., 2023; Program et al., 2024; Zhang et al., 2025a; Chandrasekaran et al., 2024).

Drawing on more than a decade of experience at Recursion building predictive, generative, and causal models to accelerate drug discovery, we believe that we can build virtual models with P-E-D capabilities without resorting to fully mechanistic simulation by training AI/ML models on massive interventional datasets using powerful compute infrastructure combined with active lab-in-the-loop data generation. In so doing, we will realize the transformative impact we believe virtual models will have on drug discovery (see Table 1 for examples of how virtual cells could have this impact).

**Table 1: Applications of virtual cells in drug discovery.** Virtual models at every level of biological organization could revolutionize the drug discovery process, from early disease modeling through clinical trial design, by aiding researchers to *predict* response to therapies, *explain* response via key mechanisms, and *discover* novel insights through lab-in-the-loop experimentation. Here we outline how the Predict-Explain-Discover capabilities of virtual cells could have this impact.

Drug discovery stage	Applications	Capabilities
Understanding Disease Mechanisms	Compare healthy vs. diseased states to identify perturbed regulatory mechanisms and disease-specific vulnerabilities	Explain, Discover
	Explain how genetic backgrounds alter disease mechanisms, variability in disease manifestation, and drug responses to identify robust, context-specific druggable entry points	Explain
Target Identification & Validation	Discover and prioritize disease-driving genes by simulating the functional consequences of mutations, loss-of-function events, splicing variants, and dysregulated expression	Explain, Discover
	Predict target essentiality (pan-cell or context-specific) and co-dependencies (e.g., synthetic lethality)	Predict
	Predict target druggability and downstream effects of modulating a specific target in disease-relevant contexts	Predict
Hit Identification & Compound Screening	Perform large-scale virtual screens of compounds, predicting activity across multiple cell lines and contexts	Predict
	Predict compound selectivity and off-target effects across cell types (e.g., toxicity versus efficacy)	Predict
	Map compound phenotypic responses to upstream molecular events and generate plausible MoA hypotheses through reasoning over structural and functional data	Explain, Discover
Mechanism of Action Studies	Explain polypharmacology using multimodal perturbation signatures	Explain
	Predict molecular and phenotypic outcomes following compound perturbation, capturing both acute (short-term) and chronic (long-term) response dynamics	Predict
	Predict and explain structure-activity relationships (SAR) to guide minimal structural modifications that enhance efficacy, optimize selectivity, or reduce liabilities	Predict, Explain
Hit-to-Lead & Lead Optimization	Predict ADMET profiles to optimize pharmacokinetic and safety properties	Predict
	Identify mechanisms and guide designs for emerging therapeutic modalities (allosteric modulators, covalent inhibitors, and glues)	Explain, Discover
	Predict and explain emergence of drug resistance through pathway rewiring, feedback loops, or network-level adaptation	Predict, Explain
Resistance Prediction & Disease Evolution	Predict clonal evolution dynamics and selection pressures in response to therapeutic interventions	Predict
	Discover rational combination therapies or synthetic lethality strategies to overcome or delay resistance	Discover
	Explain context-specific compound activity (e.g., toxicity in one tissue versus efficacy in another)	Explain
Preclinical & Translational Modeling	Predict therapeutic, immune, and inflammatory responses across patient-derived and experimental models	Predict
	Discover robust biomarkers predictive of patient-specific therapeutic responses	Discover
	Inform patient stratification strategies and biomarker-based inclusion criteria	Discover
Clinical Trial Design & Biomarker Strategy	Predict optimal human dose and combination schedules for clinical studies	Predict

In this perspective we share our vision of how to leverage these advances to build therapeutically-relevant virtual models. Here we focus on *virtual cells*, due to both the heightened interest in these models currently, as well as the wide availability of large cellular datasets. We introduce key design principles for virtual cells, and describe a lab-in-the-loop paradigm for continuously refining virtual cells, treating them as testable theories of human cellular physiology and pathology that agentic systems attempt to falsify. We also advocate for biologically-meaningful benchmarks to guide virtual cell development, and make the case that the framework presented here is appropriate for building virtual models with Predict-Explain-Discover capabilities at all levels of biological organization.

### 1.3 Prior work and current perspectives on virtual cells

The idea of building virtual models, especially virtual cells, is not new; in fact, virtual cells have been recognized as one of this century’s “grand challenges” in computational biology (Tomita, 2001). Early pioneering efforts such as E-Cell (Tomita et al., 1999; Tomita, 2001) and Virtual Cell (VCCell) (Loew & Schaff, 2001; Slepchenko et al., 2003) created rule-based or reaction-diffusion frameworks that integrated diverse biological datasets to model cellular processes. These approaches were later extended to genome-scale whole-cell models for bacteria (Karr et al., 2012; Macklin et al., 2020; Sun et al., 2021) and yeast (Ye et al., 2020) which attempted to model a significant portion of genes, gene products, and their functions, demonstrating that it is, in principle, possible to predict phenotypes from genotypes. These models have been used to validate experimental findings, predict novel mechanisms, and enable counterfactual simulations. However, they remain limited to simple organisms and face major challenges in scalability, dynamic complexity, and computational tractability, and their construction remains labor-intensive, with only a handful built to date (Georgouli et al., 2023). Importantly, these efforts have exposed a persistent gap between available biological data and the requirements for parameterizing large-scale mechanistic simulations. More recently, the field has also moved toward structural modeling approaches, including the reconstruction of whole-cell 3D structures, such as the bacterium *M. genitalium* (Maritan et al., 2022), and attempts to simulate the entire minimal cell JCVI-syn3A via coarse-grained molecular dynamics, although current MD engines were unable to complete the simulation due to computational limitations (Stevens et al., 2023).

Although not virtual cells per se, several large-scale initiatives have sought to advance our understanding of human biology by mapping the relationships among biological components. Examples include the IUPS Human Physiome Project, which is developing a multi-scale framework for the hierarchical modeling of physiological function (Hunter et al., 2002; 2024; Hunter & Borg, 2003), the Human Cell Atlas, which is producing comprehensive reference maps for all human cells (Regev et al., 2018; Rood et al., 2025), the Human Proteome Project, which aims to map all expressed proteins in the human body (Hanash & Celis, 2002; Omenn et al., 2024), and the Human Connectome Project, which aims to map all neural connections in the human brain (Van Essen et al., 2013).

We note that several new perspectives have recently proposed updated visions for virtual cells in the era of AI and large-scale biological data. A group led by the Allen Institute for Cell Science advocates for the integration of top-down phenomenological models of cell behavior with bottom-up structural models of biomolecular interactions, using knowledge graphs to connect the various spatial and temporal scales (Johnson et al., 2023). Separately, a group sponsored by the Chan-Zuckerberg Initiative envisions virtual cells as collections of embeddings, generated by specific foundation models, of the various biomolecules found in cells, with “virtual instruments” designed to simulate interventions and predict associated readouts (Bunne et al., 2024). Similarly, a recent perspective (Cui et al., 2025) proposes building multimodal foundation models that integrate omics data across modalities via unified transformer architectures, enabling applications such as *in silico* perturbation, cell state characterization, and biomarker discovery.

We share several points of agreement with these perspectives, notably the need to combine top-down and bottom-up modeling approaches, the critical role of multimodal data integration, and the transformative potential of foundation model architectures. However, our view on virtual cells is distinct in critical ways: while others emphasize static representations or predictive embeddings, we prioritize building causal, mechanistically-grounded models that not only predict but also explain the functional response of cells to perturbations. Given our primary objective of bringing new medicines to patients, we

view virtual cells not simply as descriptive models, but as systems capable of iteratively generating interpretable and testable hypotheses of cellular behavior, which can be continuously refined through experimental feedback to drive therapeutic discovery.

## 2 A Vision for Virtual Cells

Cell biology can be characterized at multiple scales. At the *molecular level*, physical equations describe how forces govern the behavior of individual atoms. In principle, atomistic simulations that integrate the effects of these forces over time would allow for the exact simulation of an entire cell. However, doing so for the roughly 100 trillion atoms estimated to make up a typical eukaryotic cell (Milo & Phillips, 2015) remains computationally intractable. Still, the cumulative behavior of these trillions of atoms in response to an external perturbation, such as treatment with a drug, can be measured at the *cellular level* via omics experiments. If these cellular-level measurements are sufficiently detailed, we then obtain a holistic view of the *functional response* of the cell to the perturbation.

For drug discovery, modeling system-wide cellular response is essential, since most therapeutic interventions do not act on a single isolated target but instead modulate networks of biomolecules, triggering feedback loops and producing off-target effects. Simply predicting the impact of a perturbation on a single protein or gene is insufficient for capturing this holistic, functional view. We therefore argue that virtual cells must first act as predictors of the functional response of cells to perturbations in order to be relevant for drug discovery—the first of the P-E-D capabilities.

Several recent efforts have demonstrated progress in this direction, proposing promising approaches to predicting transcriptomic and phenotypic changes following genetic or chemical perturbations, including GEARS (Roohani et al., 2023), CPA (Lotfollahi et al., 2023), scLAMBDA (Wang et al., 2024), CellFlow (Klein et al., 2025) and TxPert (Wenkel et al., 2025). These models leverage information from the molecular scale to better predict functional response at the cellular scale, by using either embeddings of small molecules (Sypetkowski et al., 2024) or proteins (Hayes et al., 2024), or biological knowledge such as protein-protein or pathway-based interaction networks (Roohani et al., 2023; Wang et al., 2024; Wenkel et al., 2025).

We too share the objective of predicting holistic functional response, but our vision extends further: we seek to build virtual cells that can also infer molecular-scale interactions from cellular-scale observations. Thus, a virtual cell should *explain* cellular-level observations in terms of molecular-level mechanisms, without resorting to whole-cell simulation—the second of the P-E-D capabilities. Furthermore, virtual cells must also be capable of generating testable hypotheses to *discover* novel treatments for patients—the third of the P-E-D capabilities. See Figure 2 for an overview of how the P-E-D capabilities of virtual cells work together to drive therapeutic discoveries. In the sections that follow, we expand on each core capability and introduce a corresponding *design principle* to guide virtual cell development.

### 2.1 Virtual cells should predict functional responses

A cellular functional response refers to the change in behavior, state, or activity of a cell in response to stimuli, such as compounds, genetic modifications, or environmental changes. These responses can manifest across many biological levels, including gene expression, protein activity, signaling pathways, morphology, proliferation, and secretion. In recent years, our ability to generate data capturing aspects of these functional responses has improved dramatically, due in large part to advances in high-throughput omic techniques—genomic, transcriptomic, proteomic, phenomic, metabolomic, epigenomic, etc.—at increasingly fine resolution, including at the single-cell and even single-molecule levels. These datasets are often *interventional*, meaning they capture a readout from a cell after intervening on it, usually with one or more *perturbations*—molecular tools designed to alter cellular function either temporarily or permanently. Examples of perturbations include gene knockout (Dixit et al., 2016), gene overexpression (Joung et al., 2017), treatment with small molecules (Ye et al., 2018), extracellular stimulation with soluble factors (Cuccarese et al., 2020), and infection with viral agents (Heiser et al., 2020).

Taken together, these interventional datasets provide an extensive, multimodal view of how cells respond to perturbations, and learning to predict these functional outcomes, by training AI/ML models on these data directly rather than attempting whole-cell simulation, offers a practical path for building

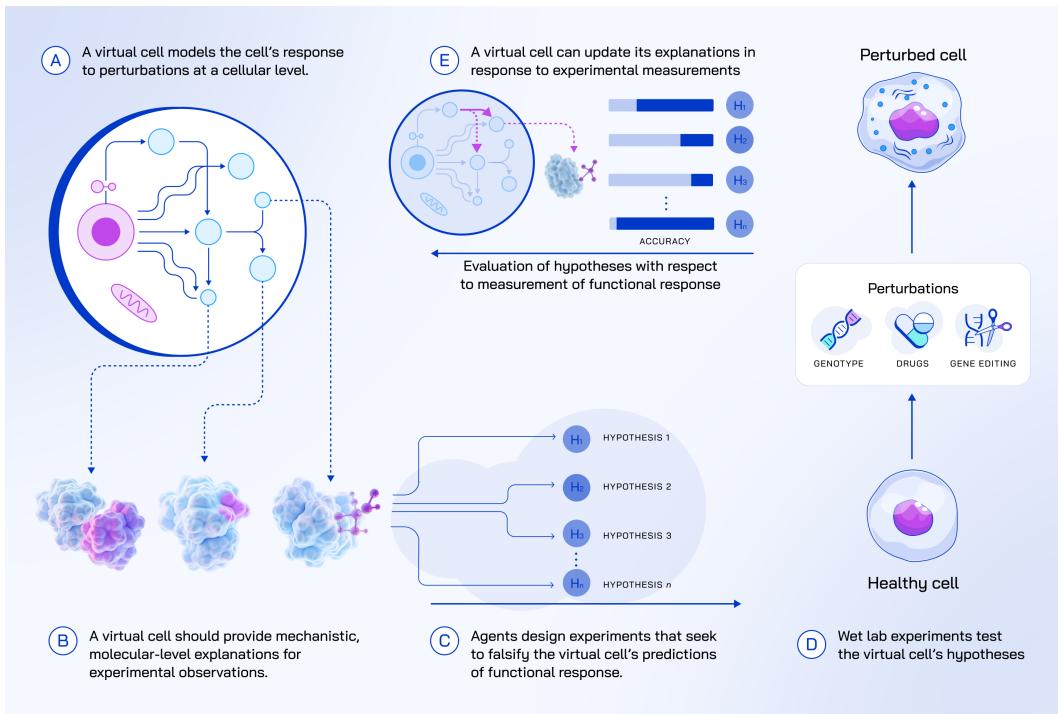


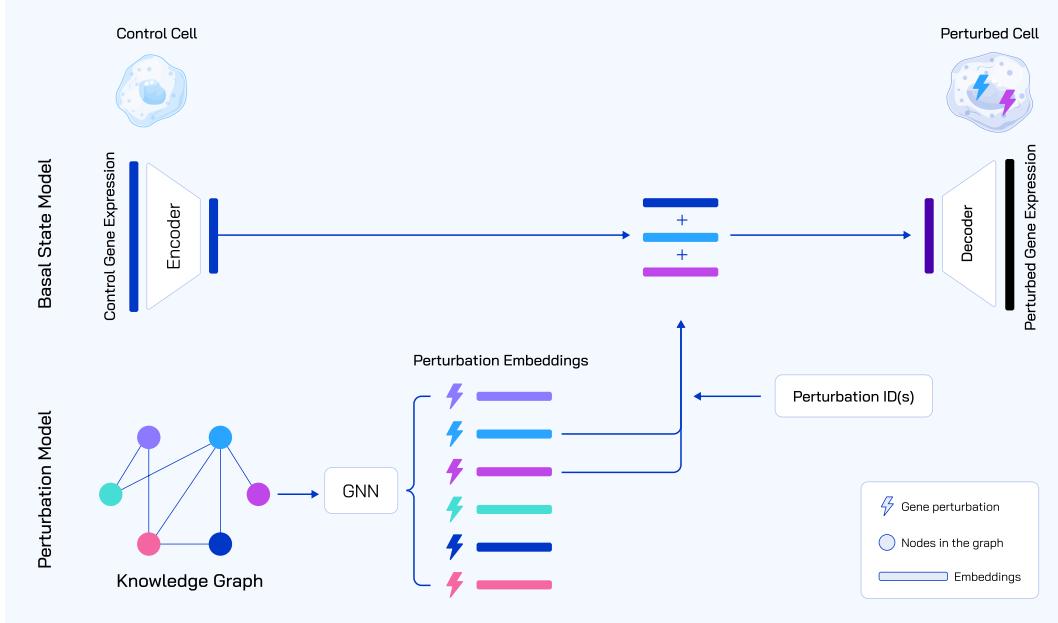
Figure 2: **A vision for virtual cells: the Predict-Explain-Discover capabilities in action.** **A.** Virtual cells *predict* the functional response of a cell to a perturbation. **B.** Virtual cells *explain* the prediction via mechanistic descriptions of key molecular interactions. **C.** Agents design experiments that seek to *discover* novel biology. **D.** These experiments are executed in real-world labs to test the generated hypotheses. **E.** Virtual cells are updated whenever a hypothesis is falsified, closing the loop between prediction and measurement. Together, these capabilities enable therapeutic discovery.

therapeutically-relevant virtual cells. Such predictive models<sup>1</sup> would be surrogates for experimental assays and enable the testing of therapeutic hypotheses *in silico*, introducing a new paradigm of rational drug design at the whole-cell level rather than at the level of individual biomolecules (Hopkins, 2008; Moffat et al., 2017; Rafelski & Theriot, 2024). Therefore, the first key capability we envision for virtual cells is to accurately *predict the functional response of cells to perturbation*.

**Design principle 1: Predict relative changes** To predict functional response, virtual cell predictions should be expressed as changes relative to the state of the cell to which the perturbation was applied; i.e., the prediction is conditioned on the state of the cell prior to perturbation (see Figure 3).

Conditioning on the initial state of the cell ensures that models focus on features directly relevant to the perturbation and its context. Ideally, virtual cells are trained on temporal trajectories from the same biological sample, measured before and after a perturbation, thereby enabling them to learn how the perturbation alters the cell's developmental path. Thus the model would learn that an effect is only expected in certain cellular contexts where the perturbation target is active, and that its magnitude depends on the state of the cell at the time of intervention.

<sup>1</sup>To be useful, such virtual cells also need to generalize well across perturbation types, cell types, and assay conditions, while maintaining robustness to batch effects and other technical noise.



**Figure 3: Virtual cells should predict relative changes.** Models like TxPert (Wenkel et al., 2025) predict the change in the readout (e.g. gene expression) of a perturbed cell given the readout of a control cell. Doing so allows the model to focus on learning the effect of the perturbation on a given control state, which provides the context for the prediction.

While such trajectories are rare<sup>2</sup>, most interventional datasets do include *negative control* samples<sup>3</sup> that approximate the unperturbed distribution. Using these controls further helps isolate the targeted effect from conserved cellular programs, such as housekeeping gene expression and homeostatic machinery.

This formulation not only improves robustness but also naturally fits the interventional nature of most available datasets. Although experimental metadata are often sparse, and small contextual differences can have significant effects on outcome<sup>4</sup>, conditioning predictions on the observed initial state allows virtual cells to absorb such variability. This reduces the need for aggressive pre-processing or batch correction and enables training across heterogeneous datasets with minimal harmonization.

## 2.2 Virtual cells should explain functional responses in terms of molecular mechanisms

While accurate prediction is essential, virtual cells must also provide *explanations*: structured, testable accounts of how perturbations give rise to observed cellular outcomes. Such explanations are critical for guiding experimental design, generating hypotheses, supporting therapeutic decisions, evaluating model trustworthiness, reasoning about biological system behavior, and identifying actionable points of intervention.

In theory, quantum mechanical (QM) approaches such as density functional theory (DFT) can accurately model biomolecular systems all the way down to the electronic level. However, they are computationally prohibitive for large systems and often fail to capture critical interactions like van der Waals forces (Grimme et al., 2016; Caldeweyher et al., 2019). Even simulating a few nanoseconds of molecular dynamics for modest systems can consume entire compute-days (Cole & Hine, 2016). Force-field-based methods improve efficiency at the cost of physical realism, limiting their reliability for

<sup>2</sup>Many common cellular assays destroy samples in order to obtain readouts (e.g., RNA-seq methods require lysing cells in order to capture RNA). Live cell assays like brightfield microscopy can preserve samples across timepoints.

<sup>3</sup>In practice, negative control samples are not simply untreated samples. Instead, they are treated with a perturbation of the same type which is expected to induce only the shared effects of that perturbation type, so that the only difference between negative controls and perturbed samples is the targeted effect of the perturbation. For example, CRISPRn controls may use guides targeting introns, which does not (intentionally) knock out any gene, but does induce the common effects of DNA cutting (e.g., DNA damage response).

<sup>4</sup>Such differences in experimental conditions are often the origin of batch effects.

modeling complex or long-timescale biochemical processes. Sparse experimental data further compound these limitations. Structural biology techniques like X-ray crystallography and cryo-EM and their deep learning counterparts such as AlphaFold and related approaches (Jumper et al., 2021; Abramson et al., 2024; Wohlwend et al., 2024; Boitreauaud et al., 2024) provide high-resolution but static snapshots of biomolecular structures. Although time-resolved methods like FRET (Sekar & Periasamy, 2003) and trEM (Amann et al., 2023) offer dynamic insights, they remain too costly and low-throughput to scale. Altogether, these computational, methodological, and empirical constraints make simulating an entire eukaryotic cell from first principles a distant goal.

Nonetheless, targeted atomistic simulations remain highly effective for modeling tractable events such as ligand–receptor binding, protein–protein interactions, allosteric regulation, and local conformational shifts. These applications are already central to drug discovery and provide essential mechanistic anchors for interpreting cellular responses (Anderson, 2003; Sliwoski et al., 2014; De Vivo et al., 2016; Durrant & McCammon, 2011).

Here too, AI/ML offer a path to scale these capabilities. ML-based interatomic potentials (MLIPs) trained on QM data can achieve near-quantum accuracy while accelerating *ab initio* simulations by several orders of magnitude (Martin-Barrios et al., 2024; Mann et al., 2025). Generative models (Jing et al., 2024; Pang et al., 2025) can simulate conformational transitions, upsample sparse MD trajectories, and design molecules under structural constraints. Incorporating these tools into virtual cells enables time-resolved, mechanistically grounded modeling of key molecular events, bridging top-down functional predictions with bottom-up molecular causes (see Figure 4).

Such integration unlocks new capabilities for perturbation modeling, causal inference, and the interpretation of dynamic cell state transitions. ML-accelerated atomistic simulation thus plays a key explanatory role: anchoring predictions in molecular reality and supporting scalable, testable hypothesis generation.

Therefore, the second key capability we envision for virtual cells is to *explain the functional response of cells to perturbation* in terms of key molecular mechanisms.

### **Design principle 2: Explain perturbations as dynamic changes to key biomolecular interactions**

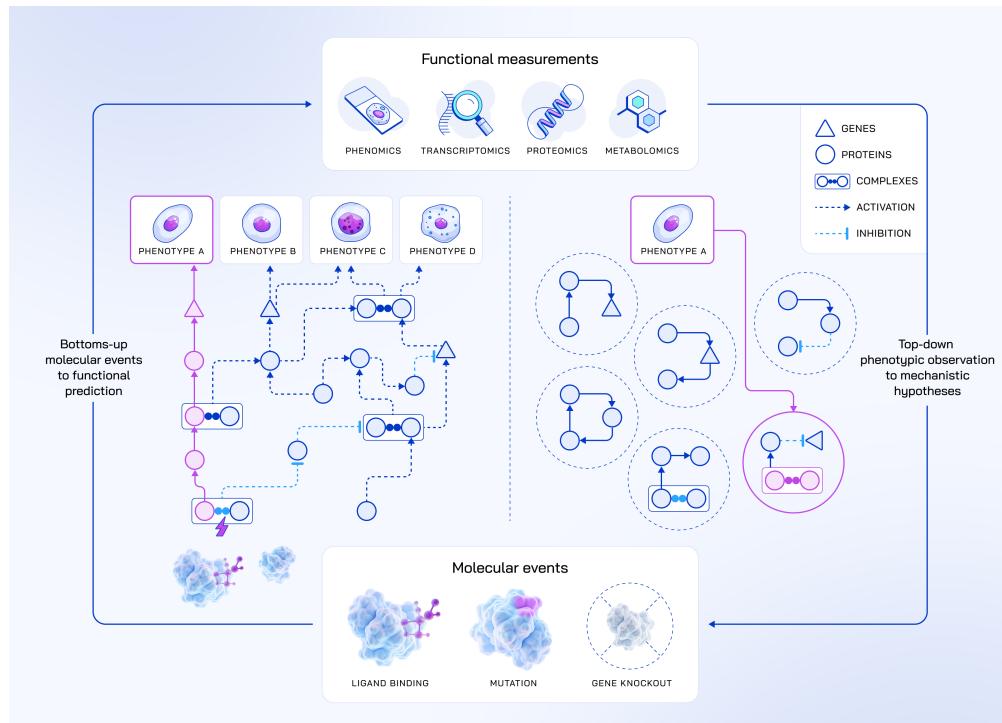
To explain functional responses, virtual cells must model how perturbations dynamically alter the structure, activity, or strength of biomolecular interactions. We adopt the systems biology view that cells function through organized, dynamic networks of interacting molecules, structured into pathways that coordinate information flow and control. Perturbations shift these networks: proteins bind differently, conformations change, and regulatory circuits rewire.

Virtual cells should frame these effects as cascades of changes to a minimal, coherent set of *key interactions*—binding affinities, post-translational modifications, transcriptional regulation, signal propagation—sufficient to explain observed outcomes. These hypotheses can be informed by ML-based structural biology tools such as AlphaFold (Abramson et al., 2024) and Boltz (Wohlwend et al., 2024), and refined through targeted atomistic simulation.

This mechanistic framing is especially valuable for therapeutic discovery. Virtual cells grounded in structural modeling can identify cryptic pockets, allosteric sites, or conformational vulnerabilities. When paired with generative molecular design tools (Winnifirth et al., 2023; Du et al., 2024; Noutahi et al., 2024; Roy et al., 2023; Cretu et al., 2025), they can propose plausible, mechanism-based interventions, thus accelerating the design–simulate–test cycle.

Identifying which interactions are *key* remains a modeling challenge. A pragmatic strategy is to prioritize candidate mechanisms that improve predictive performance or yield falsifiable hypotheses. For instance, a virtual cell might suggest that inhibiting a kinase alters transcription via a defined signaling cascade, testable through CRISPR knockouts, pathway reporters, or perturbation screens.

Importantly, we do not require virtual cells to replicate canonical biological pathways. Instead, they may uncover alternative structures that better fit the data, improve generalization, or offer more interpretable insights. Explanations should therefore be evaluated not by their agreement with textbook



**Figure 4: Virtual cells bridge top-down functional prediction with bottom-up molecular explanations.** A virtual cell’s functional predictions should be informed by omic-level observations and constrained by physical molecular interactions. Together these combine to provide a mechanistic model of a cell, propagating the effect of a perturbation from a change in molecular interactions through the affected pathways to the observed functional response.

biology, but by their ability to support causal reasoning, guide experiment design, and generate testable hypotheses.

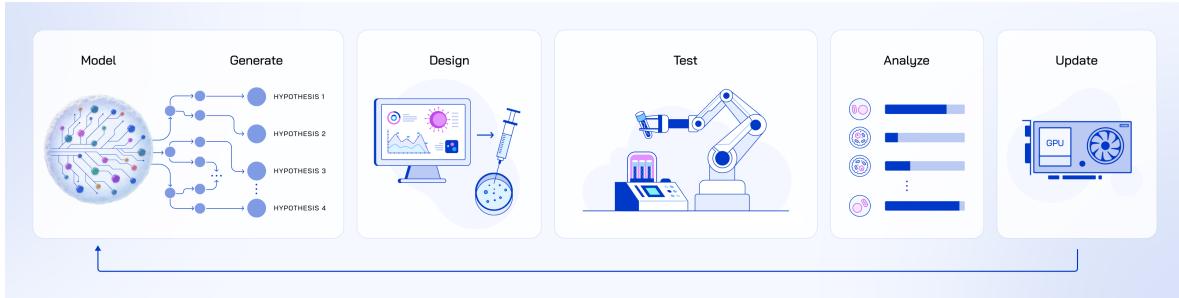
### 2.3 Virtual cells should discover therapeutically-actionable biology

Virtual cells are an essential component of a broader vision for improving drug discovery. They serve as world models of the cellular biology we observe in the “real world” via lab experiments. As such, they embody our best understanding of human cellular physiology and pathology, encapsulating the mechanistic effects of the network of biomolecular interactions that underlies cellular function.

A well-designed virtual cell should thus not only recapitulate known biology but also act as a hypothesis engine for discovering new biology. By generating responses across diverse perturbations and cellular contexts, virtual cells can propose novel mechanisms, identify promising interventions, uncover therapeutic opportunities, and prioritize experiments that yield the greatest insight.

This motivates a lab-in-the-loop paradigm, where virtual cells iteratively suggest experiments that are maximally informative. These experiments test predictions, falsify incorrect hypotheses, and update the model. Conceptually, a virtual cell becomes a testable theory of human cell physiology, continuously refined through empirical feedback (Popper, 2005; Corfield et al., 2009). Therefore, the third key capability we envision for virtual cells is to ***discover novel biology*** through iterative, hypothesis-driven experimentation.

**Design principle 3: Lab-in-the-loop falsification to align virtual cells with the real world** We envision virtual cells as being initially trained on available data, grounding them in experimentally-observed cellular behavior. Once trained, however, they could be paired with active learning and sequential model-based optimization techniques (Jain et al., 2023; Pauwels et al., 2014; Sverchkov & Craven, 2017)



**Figure 5: Virtual cells are falsifiable theories of cellular biology.** Virtual cells are world models of cellular biology, initially trained on existing experimental data, then guiding the selection of new experiments that are most informative for improving their performance. This iterative, lab-in-the-loop approach can be turned into a program of “theory falsification” to generate novel drug discoveries.

to design and prioritize experiments that are most informative for improving performance, expanding functional coverage to new cellular contexts and increasingly complex perturbation combinations.

This sets the stage for a continual lab-in-the-loop improvement of virtual cells. Conceptually, a virtual cell can be considered a theory of human cellular physiology and pathology, one that generates a diverse set of statements about the real world that we observe via experiments. Following an active learning interpretation of Popper’s theory of scientific discovery (Popper, 2005), we can imagine one or more agents, whether human or artificial, choosing hypotheses for testing, and updating the virtual cell “theory” whenever a claim is falsified (Corfield et al., 2009). Actively seeking experiments that falsify the current virtual cell could lead to surprising and interesting novel biology (Zuheng et al., 2024).

Furthermore, by conditioning the generation of hypotheses on a particular disease, the agent could steer this active improvement toward discovery of potential treatments more efficiently (Neporozhni et al., 2025). From a drug discovery perspective, such virtual cells could lead to a radical change in how programs are run, shifting from cycles of *design-make-test-measure* to *design-simulate*. Figure 5 illustrates these concepts.

### 2.3.1 A path to scientist AIs

If we imagine the agents operating within this lab-in-the-loop framework as autonomous agentic systems capable not only of the conditional generation of hypotheses using virtual cells, but also:

- prioritizing hypotheses for falsification,
- designing experiments to efficiently test those hypotheses,
- orchestrating the execution of experiments,
- analyzing the experimental outcomes, and
- integrating these into the virtual cell for iterative refinement,

then the falsification framework described in the previous section becomes a plausible path toward developing genuine *scientist AIs* capable of making novel discoveries and fundamentally transforming drug discovery workflows. Indeed, we envision a future in which such scientist AIs, with access to high-fidelity virtual models and the ability to be prompted in natural language, could accept a query such as “propose a set of mechanistically distinct compounds predicted to be effective in treating patients with the following clinical and laboratory parameters (see attached table), which may be representative of a disease that is part of a heterogeneous group of diseases often referred to as stage II non-Hodgkin’s lymphoma”. From there, they could iteratively and autonomously work toward the discovery objective, ultimately generating drug candidates with high likelihood of translating successfully to human patients.

Realizing this vision will require systematic, biologically grounded benchmarks that can track progress across the core capabilities of virtual cells.

### 3 Toward Rigorous and Biologically-Grounded Benchmarks for Virtual Cells

In designing virtual cells, it is essential to consider their downstream applications and ensure they can be systematically evaluated. Benchmarking is not an afterthought, but a fundamental constraint that must shape how virtual cells are built, trained, and validated. Driving measurable progress across different approaches will require a suite of benchmarks covering a wide range of functional responses, cellular contexts, and perturbations, with objectives spanning prediction, explanation, and discovery. To achieve this, benchmarks must capture both predictive accuracy and biological relevance, while aligning evaluation with the broader objectives of understanding and modulating cellular states for therapeutic discovery. Four major aspects must be systematically covered:

**Functional responses.** We expect functional responses to typically be measured as observed changes in omics-level readouts, like changes in gene expression captured by scRNA-seq transcriptomic assay or changes in morphological features as measured in a brightfield phenomic assay, following one or more perturbations. While neither gene expression changes nor morphological shifts alone fully capture a cell’s functional response, they provide relevant, measurable views of cellular behavior. To move closer to capturing true response, benchmarks should aim to integrate information across multiple modalities, rather than relying solely on a single type of readout that may reflect modality-specific biases or noise. We must leverage the available data and construct meaningful benchmarks from these partial but complementary measurements (Wognum et al., 2024; Tossou et al., 2024).

**Cellular contexts.** Cellular contexts can be characterized by a variety of biological states: different cell types, stages of division, differentiation, metabolism, signaling activities, spatial organization, and other physiological descriptors. In addition, the surrounding microenvironment, such as the tissue architecture, neighboring cell types, and extracellular signaling milieu, plays a critical role in shaping cellular responses. Because cellular responses to perturbations are shaped by both intrinsic state and external conditions at the time of intervention, it is essential that we evaluate our virtual cells across a wide range of physiologically-relevant settings. To support our stated therapeutic and translational goals, benchmarks should test generalization across a wide array of disease-relevant contexts.

**Perturbations.** The term *perturbation* refers to a wide array of tools designed to engage or disrupt mechanisms in and on the cell, altering its function either temporarily or permanently. CRISPR technologies knock out genes, silence translation, or increase transcription. Chemical compounds bind proteins and alter their conformations affecting downstream activities. Soluble factors such as cytokines initiate extracellular signaling cascades, while antibodies can trigger immune responses. Collecting readouts across a wide range of perturbations gives us a broad view of the inner workings of cells. Importantly, some perturbation mechanisms, particularly in specific cellular contexts, have been well characterized through decades of research. This prior knowledge can be leveraged when building benchmarks, enabling the integration of perturbations with varying levels of mechanistic certainty. Virtual cell models that accurately predict functional response across this spectrum will provide invaluable insights into how to modulate disease-associated cellular states, resulting in testable therapeutic hypotheses.

**Predict-Explain-Discover.** As described in our vision for virtual cells (Section 2), we want to build virtual cells that not only accurately predict functional responses, but explain those responses mechanistically and drive novel drug discovery. Thus our benchmarks must assess all three of these capabilities. Predictions are relatively easy to benchmark because any new experiment generates data that could be used to benchmark a virtual cell. In contrast, evaluating explanations and assessing novelty are more difficult, as it requires assessing statements about unobserved biological mechanisms that would typically demand multiple specialized assays. While we can benchmark explanations and discovery with known biology, such efforts need to be very carefully controlled to avoid data leakage, particularly in light of the growing reliance on literature-derived information as a way of constructing gene embeddings (Chen & Zou, 2024).

Following these principles, we describe a framework of capabilities we expect virtual cells to demonstrate as the field advances, and recommend that all benchmarks designed for virtual cells are intentionally associated with one or more of these capabilities. Doing so will help to better mark progress, differentiate between virtual cell models, and guide ongoing research efforts.

Table 2: **A framework of virtual cell capabilities.** Each capability unlocks specific aspects of biological understanding and enables downstream applications.

Predicts response...	Understandings unlocked	Applications
<i>Observational capabilities</i>		
For specific biology	Biomolecular features	Predicting protein localization; reconstructing higher-resolution readouts from bulk or lower-resolution assays
For core biology	Modality-invariant biological processes (e.g., transcription–translation coupling)	Predicting pathway-level effects of perturbations; replacing expensive modalities with cheaper ones (e.g., transcriptomics with imaging)
For unobserved biology	Latent molecular states (e.g. post-translational modifications, hidden regulators)	Inferring unseen regulatory nodes; grouping genes and compounds into higher-order functional modules (e.g., pathways)
<i>Contextual capabilities</i>		
Given intrinsic context	Effect of genotype, cell type, biomolecular state.	Generalizing across cell types, genotypes, and developmental states
Given extrinsic context	Influence of assay design, local microenvironment, and intercellular signaling	ADME-T prediction; modeling effects of experimental conditions and local tissue context (e.g., proximity-driven signaling, tissue-level response)
<i>Explanatory capabilities</i>		
Over time	Temporal integration of molecular and cellular dynamics	Treatment response over time
Causally	Mechanistic understanding of regulation and interaction	Mutation effect prediction; causal modeling

Table 2 summarizes this framework of capabilities and gives examples of how they could impact drug discovery and our understanding of biology. While these capabilities are organized along increasing levels of complexity, we acknowledge that they do not need to be achieved in the order presented here, nor do we claim that this list is exhaustive. We note that our framework primarily captures the Predict and Explain capabilities of virtual cells, which we believe are the core enablers of discovery. As described in earlier sections, we view Discover not as a separate axis to benchmark in isolation, but as the natural consequence of predictive and explanatory models applied to therapeutic contexts. In particular, discovery is best evaluated through a model’s ability to generate testable and actionable hypotheses, an aspect that depends on but extends beyond the capabilities presented here.

### 3.1 Observational capabilities

As performance moves beyond simple baselines, we want to assess how much a virtual cell’s predictions can leverage the different types of readouts that are provided. While we advocate for modality-agnostic evaluations where possible (see Appendix A for an extended discussion on benchmarking considerations), in practice this means measuring response predictions with respect to available modalities. These capabilities, therefore, begin with the specific biology accessible within a single modality, then expand to focus on predicting functional responses for core biology, which should be observable across all modalities, and unobserved biology, which may not be directly observed but is nonetheless known to play a role.

**Predicts response for specific biology.** The scope of this capability is relatively narrow, measuring the extent to which a virtual cell predicts functional response from single modalities. Nevertheless, since ground-truth examples used for evaluation will often come from multiple experiments, models that perform well here will likely need to predict across differences in context (see Section 3.2 for more on contexts) in order to deal with *batch effects* and other sources of variation across these datasets.

**Predicts response for core biology.** Here we broaden the scope to functional responses related to the central dogma. There are a set of core biological processes that should appear regardless of the modality in which they are observed. For example, for genes that maintain basic cellular functions—actin, GAPDH, ubiquitin, etc.—the presence of their transcripts implies the presence of their protein products, and they should be observable in both transcriptomics and proteomics. Thus this capability assesses the extent to which a virtual cell can capture fundamental biology.

**Predicts response for unobserved biology.** This capability broadens the scope even further to predicting functional responses that are not directly observed in any of our modalities, but are nonetheless expected based on our knowledge of biology. For example, predicting that an RNA but not the corresponding protein is produced (long non-coding RNAs), or that an RNA rather than a protein catalyzes a biological reaction (ribozymes) is biologically plausible. Similarly, a given readout may not explicitly capture phenomena such as post-translational modifications but should be accounted for nonetheless. The extent to which we can accurately predict into these “dark” areas of biology will help us understand which biological processes are represented in each modality, which subsets are most informative for making a particular prediction, and provide evidence that the model is capable of producing novel insights.

### 3.2 Contextual capabilities

Cells do not exist in a vacuum but within various levels of context determined by details of the cell itself, environmental factors, and the milieu of surrounding cells and biological constituents. These capabilities, therefore, measure the ability to predict functional response across a broadening set of biological contexts. Note that we could continue to describe capabilities at higher levels of biological organization, but those would be more appropriate for benchmarking correspondingly higher-levels of virtual models, and thus omit them here.

**Predicts response given intrinsic context.** By *intrinsic context*, we mean all those factors describing the biology of the cell: genotype, cell type, cell cycle phase, pathway activation, levels of key biomolecules including RNA, protein, and ATP. Virtual cells that account for these biological factors will accurately predict response across a wide range of cellular contexts. For example, cell types differ significantly in gene expression patterns, morphology, and regulatory programs. Virtual cells must account for these differences to accurately predict responses adapted to each cellular identity.

**Predicts response given extrinsic context.** By *extrinsic context*, we refer to all external factors influencing cellular behavior, including assay design (e.g., growth media, temperature, incubation time, reagent concentrations) as well as the spatial and cellular microenvironment. This includes interactions with nearby cells, such as contact inhibition, gap junction signaling, or paracrine communication, which can modulate cell state and response. Virtual cells that account for these confounding factors will generalize across different experimental conditions by modeling the complex interactions between cellular response and environment. We note that the effects of extrinsic context on a cell are often mediated through the biological factors of its intrinsic context, suggesting that this capability encompasses the previous one.

### 3.3 Explanatory capabilities

As outlined in the vision (Sections 2.1, 2.2, and 2.3), the most useful versions of virtual cells offer a mechanistic understanding of cellular biology in order to provide the evidence needed to understand therapeutic hypotheses. These capabilities, therefore, benchmark progress toward mechanistic explanations of a virtual cell’s predictions.

**Predicts response over time.** Cells are dynamical systems in a continual state of flux, converting chemical energy into work to maintain their internal homeostasis while responding to external stimuli. Although an observed response in any modality implicitly captures the notion of time, this virtual cell capability requires the time component to become explicit in predicting response. Accurately predicting the time-course of response will likely require a virtual cell to model how the interaction of key biomolecules over short timescales integrates over time to produce the response at longer timescales, potentially providing mechanistic insights that can be leveraged in drug discovery and other applications.

**Predicts response causally.** Knowing the reasons for *why* a particular response occurs is much more useful than simply predicting that it does. Identifying causal relationships between observed variables and building from them a mechanistic model allows virtual cells to more accurately generalize their predictions out-of-distribution, and move toward understanding the plethora of counterintuitive nonlinear phenomena we observe in biology, e.g., epistatic interactions like synthetic lethality. Causal learning, however, is a notoriously difficult problem in general, and particularly so in the high-dimensional omics datasets available within biology.

### 3.4 Performance levels for virtual cells

To further drive progress, we can organize the capabilities described above into performance levels that represent key milestones toward building increasingly powerful virtual cells.

We propose the following three levels as an initial framework for evaluating progress:

**VC level 1.** Predicts and explains functional response for specific biology, given intrinsic and extrinsic context, at a single timepoint.

**VC level 2.** Predicts and explains functional response for core biology, across environmental variation, and over time.

**VC level 3.** Predicts and explains functional response for unobserved biology, in spatially organized systems, and with causal interpretability.

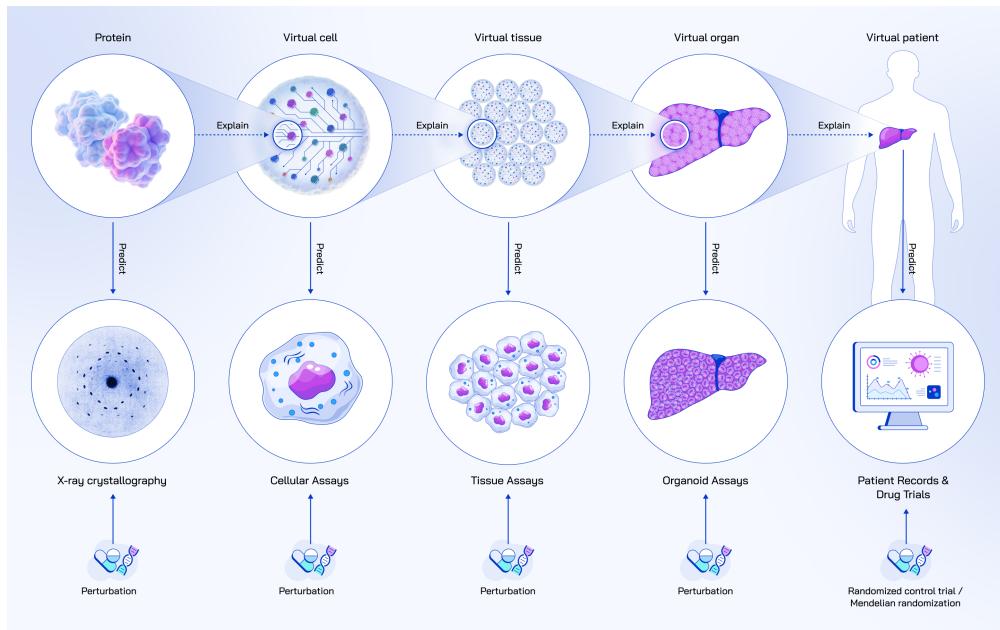
While the exact definition and composition of these levels may need to be adapted over time, we hope that adopting performance milestones for virtual cells will help organize the research community around common standards and research directions, and ultimately lead to the realization of powerful virtual cells sooner than otherwise.

## 4 A framework for building virtual models at higher levels of organization

In this perspective, we have outlined a vision for therapeutically-relevant virtual cells that predict functional responses with respect to multiple modalities, explain these in terms of molecular mechanisms, and generate novel discoveries through lab-in-the-loop experimentation. We have taken the view that such virtual cell models should combine omics data and atomistic simulation to *predict*, *explain*, and *discover*—that is, to model the functional response of cells to perturbation, reveal the underlying molecular mechanisms, and generate novel and falsifiable biological insights through iterative hypothesis testing and refinement. To support this development, we have also outlined key design principles for virtual cells and proposed a framework for the adoption of biologically-meaningful benchmarking standards to help shape the research community’s efforts.

Our primary objective in developing virtual cells as described herein is to accelerate the drug discovery process, reduce development costs, and enable more precise therapeutic intervention. While virtual cells alone will not solve every challenge, we see them as prototypical examples for virtual models at higher levels of biological organization, from virtual tissues to virtual organs and ultimately virtual patients. Importantly, we believe that virtual models at every level of organization can be built in a manner similar to what we described for virtual cells—by training models with Predict-Explain-Discover capabilities on data that suitably describes functional response at the respective level (see Figure 6 for an illustration of these ideas).

Each step toward higher-order modeling extends the Predict-Explain-Discover capabilities of virtual models into more structured and interconnected systems. While each level could build on the one below—tissue-level predictions emerge from virtual cells interacting in space and time; organ-level behavior depends on the coordination of multiple tissue types; patient-level outcomes reflect the integrated behavior of organ systems shaped by genetics, disease, and treatment history—this progression does not need to be strictly sequential. Indeed, advances in biotechnology such as organs-on-chips (Danku et al., 2022) and “*in vivo* omics” (Lim et al., 2017; Baran et al., 2021) can provide unprecedented access to structured biological responses at the tissue, organ and whole-organism levels. Combined with the increasing availability of multimodal and anonymized patient-specific profiling, these developments make it increasingly feasible to model higher-order biological organization directly.



**Figure 6: From virtual cells to virtual patients.** By focusing on predicting functional response and explaining in terms of lower-level mechanisms, we can continue to build virtual models that represent higher levels of biological organization, eventually arriving at a virtual patient.

At the tissue level, groups of interacting cells model a collective response to perturbations, predicting spatially-distributed effects such as signal propagation or local response heterogeneity, explain how intercellular communication shapes tissue phenotype, and could generate hypotheses about how spatial context modulates drug sensitivity (Yuan, 2016).

Virtual organs add another layer of complexity: they must model structured spatial organization and long-range signaling (e.g., hormonal or vascular). Predictions at this level recapitulate outputs from multiple tissue types, explanations involve anatomical dependencies and interactions between regulatory axes, while discovery may focus on falsifying hypotheses around organ-specific responses, immunogenicity, toxicity patterns, or compensatory dynamics that mask intervention effects (Sontheimer-Phelps et al., 2019; Marx et al., 2020).

At the patient level, models integrate across organs to simulate whole-body responses to interventions. These include predicting clinical outcomes such as biomarker trajectories, disease progression, or treatment efficacy over time; and explaining sources of inter-patient variability arising from differences in genetics, physiology, and environment. Virtual patient models could generate testable hypotheses for personalized treatment strategies, for example, identifying subpopulations more likely to benefit from a given therapy, or proposing individualized dosing regimens based on predicted pharmacokinetic and pharmacodynamic profiles (Minichmayr et al., 2024).

While only a vision today, it is not unreasonable to expect that we can make progress towards these objectives based on the key advances we highlighted in Section 1.2. Successfully delivering this progress would enable personalized medicine at unprecedented levels of accuracy and scale, allowing for the development of tailored therapeutic interventions based on patient-specific virtual simulations.

In the meantime, we invite the research community to engage with this perspective by refining, challenging, and extending the ideas presented herein, as we collectively progress toward a new way of doing therapeutic discovery. We hope to see the adoption of rigorous benchmarking standards as the research community working on virtual cells grows, and hope that the principles and capabilities described herein can be a useful tool in steering our collective efforts.

## 5 Acknowledgments

The perspective presented in this paper is strongly influenced by extensive discussion with colleagues at Valence Labs and Recursion, as well as with current and former interns. In particular, we would like to thank Frederik Wenkel, Cian Eastwood, Lu Zhu, Cristian Gabellini, Hatem Helal, Julien Roy, Maciej Sypetkowski, Jessica Dafflon, Wilson Tu, Austin Tripp, Kerstin Klaeser, Craig Russell, Michel Moreau, Stephan Thaler, Malika Srivastava, Thomas Rochefort-Beaudoin, Yassir El Mesbahi, Ihab Bendidi, Honoré Hounwanou, Julien St-Laurent, Cassandra Masschelein, Sébastien Giguère, Soumali Roychowdhury, Véronique Bérubé, Francesco Di Giovanni, Therence Bois, Marta Fay for the extensive discussions.

Finally, we are very grateful for the continued feedback and guidance provided by our scientific advisors Yoshua Bengio and Michael Bronstein, who helped initiate and shape many of the key ideas presented here.

## References

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.
- Amann, S. J., Keihlsler, D., Bodrug, T., Brown, N. G., and Haselbach, D. Frozen in time: analyzing molecular dynamics with time-resolved cryo-em. *Structure*, 31(1):4–19, 2023.
- Anderson, A. C. The process of structure-based drug design. *Chemistry & biology*, 10(9):787–797, 2003.
- Baran, S. W., Lim, M. A., Do, J. P., Stolyar, P., Rabe, M. D., Schaeitz, L. R., and Cadena, S. M. Digital biomarkers enable automated, longitudinal monitoring in a mouse model of aging. *The Journals of Gerontology: Series A*, 76(7):1206–1213, 2021.
- Bendidi, I., Whitfield, S., Kenyon-Dean, K., Yedder, H. B., Mesbahi, Y. E., Noutahi, E., and Denton, A. K. Benchmarking transcriptomics foundation models for perturbation analysis: one pca still rules them all. *arXiv preprint arXiv:2410.13956*, 2024.
- Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhonikov, A., and Wu, K. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pp. 2024–10, 2024.
- Buccitelli, C. and Selbach, M. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020.
- Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.
- Caldeweyher, E., Ehlert, S., Hansen, A., Neugebauer, H., Spicher, S., Bannwarth, C., and Grimme, S. A generally applicable atomic-charge dependent london dispersion correction. *The Journal of chemical physics*, 150(15), 2019.
- Chandrasekaran, S. N., Cimini, B. A., Goodale, A., Miller, L., Kost-Alimova, M., Jamali, N., Doench, J. G., Fritchman, B., Skepner, A., Melanson, M., Kalinin, A. A., Arevalo, J., Haghghi, M., Caicedo, J. C., Kuhn, D., Gustafsdottir, S. M., Rogov, P., Holbrook-Smith, D., Hasson, S. A., Wawer, M., Boland, P., Bittker, J., Subramanian, A., Singh, M., and Carpenter, A. E. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, 21(6):1114–1121, 2024. doi: 10.1038/s41592-024-02241-6. URL <https://doi.org/10.1038/s41592-024-02241-6>.
- Chen, Y. and Zou, J. Genept: a simple but effective foundation model for genes and cells built from chatgpt. *BioRxiv*, pp. 2023–10, 2024.
- Cole, D. J. and Hine, N. D. M. Applications of large-scale density functional theory in biology. *Journal of Physics: Condensed Matter*, 28(39):393001, aug 2016. ISSN 1361-648X. doi: 10.1088/0953-8984/28/39/393001. URL <http://dx.doi.org/10.1088/0953-8984/28/39/393001>.

- Corfield, D., Schölkopf, B., and Vapnik, V. Falsificationism and statistical learning theory: Comparing the popper and vapnik-chervonenkis dimensions. *Journal for General Philosophy of Science*, 40(1): 51–58, jul 2009. ISSN 1572-8587. doi: 10.1007/s10838-009-9091-3. URL <http://dx.doi.org/10.1007/s10838-009-9091-3>.
- Cretu, M., Harris, C., Igashov, I., Schneuing, A., Segler, M., Correia, B., Roy, J., Bengio, E., and Lio, P. Synflownet: Design of diverse and novel molecules with synthesis constraints. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=uvHmnahyp1>.
- Cuccarese, M. F., Earnshaw, B. A., Heiser, K., Fogelson, B., Davis, C. T., McLean, P. F., Gordon, H. B., Skelly, K.-R., Weathersby, F. L., Rodic, V., et al. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and covid-19 drug discovery. *Biorxiv*, pp. 2020–08, 2020.
- Cui, H., Tejada-Lapuerta, A., Brbić, M., Saez-Rodriguez, J., Cristea, S., Goodarzi, H., Lotfollahi, M., Theis, F. J., and Wang, B. Towards multimodal foundation models in molecular cell biology. *Nature*, 640(8059):623–633, 2025.
- Danku, A. E., Dulf, E.-H., Braicu, C., Jurj, A., and Berindan-Neagoe, I. Organ-on-a-chip: a survey of technical results and problems. *Frontiers in bioengineering and biotechnology*, 10:840674, 2022.
- De Vivo, M., Masetti, M., Bottegoni, G., and Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, 59(9):4035–4061, 2016.
- DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016.
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.
- Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P., and Blundell, T. L. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6 (6):589–604, 2024.
- Durrant, J. D. and McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC biology*, 9:1–9, 2011.
- Fay, M. M., Kraus, O., Victors, M., Arumugam, L., Vuggumudi, K., Urbanik, J., Hansen, K., Celik, S., Cernek, N., Jagannathan, G., Christensen, J., Earnshaw, B. A., Haque, I. S., and Mabey, B. Rxrx3: Phenomics map of biology. *biorXiv*, 2023. doi: 10.1101/2023.02.07.527350. URL <https://www.biorxiv.org/content/early/2023/02/08/2023.02.07.527350>.
- Georgouli, K., Yeom, J.-S., Blake, R. C., and Navid, A. Multi-scale models of whole cells: progress and challenges. *Frontiers in Cell and Developmental Biology*, 11:1260507, 2023.
- Goldberg, A. P., Szigeti, B., Chew, Y. H., Sekar, J. A., Roth, Y. D., and Karr, J. R. Emerging whole-cell modeling principles and methods. *Current opinion in biotechnology*, 51:97–102, 2018.
- Green, S. Explanatory pluralism in biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 59:154–157, 2016. ISSN 1369-8486. doi: <https://doi.org/10.1016/j.shpsc.2016.02.002>. URL <https://www.sciencedirect.com/science/article/pii/S1369848616000200>.
- Grimme, S., Hansen, A., Brandenburg, J. G., and Bannwarth, C. Dispersion-corrected mean-field electronic structure methods. *Chemical reviews*, 116(9):5105–5154, 2016.
- Hanash, S. and Celis, J. E. The human proteome organization: a mission to advance proteome knowledge. *Molecular & Cellular Proteomics*, 1(6):413–414, 2002.
- Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A.,

- Mishra, C., Kim, C., Bartie, L. J., Nemeth, M., Hsu, P. D., Sercu, T., Candido, S., and Rives, A. Simulating 500 million years of evolution with a language model. *BioRxiv*, 2024. doi: 10.1101/2024.07.01.600583. URL <https://www.biorxiv.org/content/early/2024/12/31/2024.07.01.600583>.
- Heiser, K., McLean, P. F., Davis, C. T., Fogelson, B., Gordon, H. B., Jacobson, P., Hurst, B., Miller, B., Alfa, R. W., Earnshaw, B. A., et al. Identification of potential treatments for covid-19 through artificial intelligence-enabled phenomic analysis of human cells infected with sars-cov-2. *BioRxiv*, pp. 2020–04, 2020.
- Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4):310–318, 2016.
- Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 4(11):682–690, 2008.
- Hunter, P., Bono, B. de, Brooks, D., Christie, R., Hussan, J., Lin, M., and Nickerson, D. The physiome project and digital twins. *IEEE Reviews in Biomedical Engineering*, 2024.
- Hunter, P., Robbins, P., and Noble, D. The iups human physiome project. *Pflügers Archiv*, 445:1–9, 2002.
- Hunter, P. J. and Borg, T. K. Integration from proteins to organs: the physiome project. *Nature reviews Molecular cell biology*, 4(3):237–243, 2003.
- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, 2001.
- Jain, M., Deleu, T., Hartford, J., Liu, C.-H., Hernandez-Garcia, A., and Bengio, Y. Gflownets for ai-driven scientific discovery. *Digital Discovery*, 2(3):557–577, 2023.
- Jing, B., Stärk, H., Jaakkola, T., and Berger, B. Generative modeling of molecular dynamics trajectories. *arXiv preprint arXiv:2409.17808*, 2024.
- Johnson, G. T., Agmon, E., Akamatsu, M., Lundberg, E., Lyons, B., Ouyang, W., Quintero-Carmona, O. A., Riel-Mehan, M., Rafelski, S., and Horwitz, R. Building the next generation of virtual cells to understand cellular biology. *Biophysical Journal*, 122(18):3560–3569, 2023.
- Jones, R. and Wilksdon, J. The biomedical bubble: Why uk research and innovation needs a greater diversity of priorities, politics, places and people. 2018.
- Joung, J., Konermann, S., Gootenberg, J. S., Abudayyeh, O. O., Platt, R. J., Brigham, M. D., Sanjana, N. E., and Zhang, F. Genome-scale crispr-cas9 knockout and transcriptional activation screening. *Nature protocols*, 12(4):828–863, 2017.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.
- Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Assad-Garcia, N., Glass, J. I., and Covert, M. W. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.
- Klein, D., Fleck, J. S., Bobrovskiy, D., Zimmermann, L., Becker, S., Palma, A., Dony, L., Tejada-Lapuerta, A., Huguet, G., Lin, H.-C., et al. Cellflow enables generative single-cell phenotype modeling with flow matching. *BioRxiv*, pp. 2025–04, 2025.
- Lee, C. T. and Amaro, R. E. Exascale computing: A new dawn for computational biology. *Computing in Science & Engineering*, 20(5):18–25, 2018.
- Li, Y., Wu, F.-X., and Ngom, A. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.

- Lim, M. A., Louie, B., Ford, D., Heath, K., Cha, P., Betts-Lacroix, J., Lum, P. Y., Robertson, T. L., and Schaevitz, L. Development of the digital arthritis index, a novel metric to measure disease parameters in a rat model of rheumatoid arthritis. *Frontiers in pharmacology*, 8:818, 2017.
- Liu, Y., Beyer, A., and Aebersold, R. On the dependency of cellular protein levels on mrna abundance. *Cell*, 165(3):535–550, 2016.
- Loew, L. M. and Schaff, J. C. The virtual cell: a software environment for computational cell biology. *TRENDS in Biotechnology*, 19(10):401–406, 2001.
- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.
- Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., Hetzel, L., Ji, Y., Ibarra, I. L., Srivatsan, S. R., Naghipourfar, M., Daza, R. M., Martin, B., Shendure, J., McFaline-Figueroa, J. L., Boyeau, P., Wolf, F. A., Yakubova, N., Günemann, S., Trapnell, C., Lopez-Paz, D., and Theis, F. J. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), may 2023. ISSN 1744-4292. doi: 10.15252/msb.202211517. URL <http://dx.doi.org/10.15252/msb.202211517>.
- Machamer, P., Darden, L., and Craver, C. F. Thinking about mechanisms. *Philosophy of science*, 67(1):1–25, 2000.
- Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., Sun, G., Agmon, E., DeFelice, M. M., Maayan, I., et al. Simultaneous cross-evaluation of heterogeneous e. coli datasets via mechanistic simulation. *Science*, 369(6502):eaav3751, 2020.
- Mann, E. L., Wagen, C. C., Vandezande, J. E., Wagen, A. M., and Schneider, S. C. Egret-1: Pretrained neural network potentials for efficient and accurate bioorganic simulation, 2025. URL <https://arxiv.org/abs/2504.20955>.
- Maritan, M., Autin, L., Karr, J., Covert, M. W., Olson, A. J., and Goodsell, D. S. Building structural models of a whole mycoplasma cell. *Journal of molecular biology*, 434(2):167351, 2022.
- Martin-Barrios, R., Navas-Conyedo, E., Zhang, X., Chen, Y., and Gulín-González, J. An overview about neural networks potentials in molecular dynamics simulation. *International Journal of Quantum Chemistry*, 124(11):e27389, 2024.
- Marx, U., Akabane, T., Andersson, T. B., Baker, E., Beilmann, M., Beken, S., Brendler-Schwaab, S., Cirit, M., David, R., Dehne, E.-M., et al. Biology-inspired microphysiological systems to advance patient benefit and animal welfare in drug development. *Altex*, 37(3):365, 2020.
- Milo, R. and Phillips, R. *Cell biology by the numbers*. Garland Science, 2015.
- Minichmayr, I., Dreesen, E., Centanni, M., Wang, Z., Hoffert, Y., Friberg, L. E., and Wicha, S. Model-informed precision dosing: State of the art and future perspectives. *Advanced drug delivery reviews*, pp. 115421, 2024.
- Mitchell, S. D. *Biological complexity and integrative pluralism*. Cambridge University Press, 2003.
- Moffat, J. G., Vincent, F., Lee, J. A., Eder, J., and Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nature reviews Drug discovery*, 16(8):531–543, 2017.
- Nam, Y., Kim, J., Jung, S.-H., Woerner, J., Suh, E. H., Lee, D.-g., Shivakumar, M., Lee, M. E., and Kim, D. Harnessing artificial intelligence in multimodal omics data integration: paving the path for the next frontier in precision medicine. *Annual Review of Biomedical Data Science*, 7, 2024.
- Neporozhnii, I., Roy, J., Bengio, E., and Hartford, J. Efficient biological data acquisition through inference set design. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=gVkJX9QMB03>.

- Noutahi, E., Gabellini, C., Craig, M., Lim, J. S., and Tossou, P. Gotta be safe: a new framework for molecular design. *Digital Discovery*, 3(4):796–804, 2024.
- Omenn, G. S., Lane, L., Overall, C. M., Lindskog, C., Pineau, C., Packer, N. H., Cristea, I. M., Weintraub, S. T., Orchard, S., Roehrl, M. H. A., Nice, E., Guo, T., Van Eyk, J. E., Liu, S., Bandeira, N., Aebersold, R., Moritz, R. L., and Deutsch, E. W. The 2023 report on the proteome from the hupo human proteome project. *Journal of Proteome Research*, 23(2):532–549, January 2024. ISSN 1535-3907. doi: 10.1021/acs.jproteome.3c00591. URL <http://dx.doi.org/10.1021/acs.jproteome.3c00591>.
- Pang, Y. T., Kuo, K. M., Yang, L., and Gumbart, J. C. Deeppath: Overcoming data scarcity for protein transition pathway prediction using physics-based deep learning. *bioRxiv*, pp. 2025–02, 2025.
- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., and Schacht, A. L. How to improve r&d productivity: the pharmaceutical industry’s grand challenge. *Nature reviews Drug discovery*, 9(3):203–214, 2010.
- Pauwels, E., Lajaunie, C., and Vert, J.-P. A bayesian active learning strategy for sequential experimental design in systems biology. *BMC Systems Biology*, 8:1–11, 2014.
- Pearl, J. *Causality*. Cambridge university press, 2009.
- Picard, M., Scott-Boyer, M.-P., Bodein, A., Pépin, O., and Droit, A. Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, 19: 3735–3746, 2021.
- Popper, K. *The logic of scientific discovery*. Routledge, 2005.
- Program, C. C. S., Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., Cherry, J., Chi, T., Chien, J., Dorman, L., Garcia-Nieto, P., Gloria, N., Hastie, M., Hegeman, D., Hilton, J., Huang, T., Infeld, A., Istrate, A.-M., Jelic, I., Katsuya, K., Kim, Y. J., Liang, K., Lin, M., Lombardo, M., Marshall, B., Martin, B., McDade, F., Megill, C., Patel, N., Predeus, A., Raymor, B., Robatmili, B., Rogers, D., Rutherford, E., Sadgat, D., Shin, A., Small, C., Smith, T., Sridharan, P., Tarashansky, A., Tavares, N., Thomas, H., Tolopko, A., Urisko, M., Yan, J., Yeretssian, G., Zamanian, J., Mani, A., Cool, J., and Carr, A. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1142. URL <https://doi.org/10.1093/nar/gkae1142>.
- Rafelski, S. M. and Theriot, J. A. Establishing a conceptual framework for holistic cell states and state transitions. *Cell*, 187(11):2633–2651, 2024.
- Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., et al. The human cell atlas white paper. *arXiv preprint arXiv:1810.05192*, 2018.
- Rood, J. E., Wynne, S., Robson, L., Hupalowska, A., Randell, J., Teichmann, S. A., and Regev, A. The human cell atlas from a cell census to a unified foundation model. *Nature*, 637(8048):1065–1071, 2025.
- Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, aug 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01905-6. URL <http://dx.doi.org/10.1038/s41587-023-01905-6>.
- Ross, L. N. Causal concepts in biology: How pathways differ from mechanisms and why it matters. *The British Journal for the Philosophy of Science*, 2021.
- Roy, J., Bacon, P.-L., Pal, C., and Bengio, E. Goal-conditioned gflownets for controllable multi-objective molecular design. *arXiv preprint arXiv:2306.04620*, 2023.
- Sachs, K., Perez, O., Pe’er, D., Lauffenburger, D. A., and Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

- Sadybekov, A. V. and Katritch, V. Computational approaches streamlining drug discovery. *Nature*, 616(7958):673–685, 2023.
- Sams-Dodd, F. Target-based drug discovery: is something wrong? *Drug discovery today*, 10(2):139–147, 2005.
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- Sekar, R. B. and Periasamy, A. Fluorescence resonance energy transfer (fret) microscopy imaging of live cell protein localizations. *The Journal of cell biology*, 160(5):629, 2003.
- Singh, M., Srivastava, R., Fuenmayor, E., Kuts, V., Qiao, Y., Murray, N., and Devine, D. Applications of digital twin across industries: A review. *Applied Sciences*, 12(11):5727, 2022.
- Slepchenko, B. M., Schaff, J. C., Macara, I., and Loew, L. M. Quantitative cell biology with the virtual cell. *Trends in cell biology*, 13(11):570–576, 2003.
- Sliwoski, G., Kothiwale, S., Meiler, J., and Lowe Jr, E. W. Computational methods in drug discovery. *Pharmacological reviews*, 66(1):334–395, 2014.
- Sontheimer-Phelps, A., Hassell, B. A., and Ingber, D. E. Modelling cancer in microfluidic human organs-on-chips. *Nature Reviews Cancer*, 19(2):65–81, 2019.
- Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. Multimodal deep learning for biomedical data fusion: a review. *Briefings in bioinformatics*, 23(2):bbab569, 2022.
- Stevens, J. A., Grunewald, F., Tilburg, P. van, König, M., Gilbert, B. R., Brier, T. A., Thornburg, Z. R., Luthey-Schulten, Z., and Marrink, S. J. Molecular dynamics simulation of an entire cell. *Frontiers in Chemistry*, 11:1106495, 2023.
- Sun, G., Ahn-Horst, T. A., and Covert, M. W. The e. coli whole-cell modeling project. *EcoSal plus*, 9(2):eESP–0001, 2021.
- Sverchkov, Y. and Craven, M. A review of active learning approaches to experimental design for uncovering biological networks. *PLoS computational biology*, 13(6):e1005466, 2017.
- Swinney, D. C. and Anthony, J. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011.
- Sypetkowski, M., Wenkel, F., Poursafaei, F., Dickson, N., Suri, K., Fradkin, P., and Beaini, D. On the scalability of gnns for molecular graphs. *arXiv preprint arXiv:2404.11568*, 2024.
- Tomita, M. Whole-cell simulation: a grand challenge of the 21st century. *TRENDS in Biotechnology*, 19(6):205–210, 2001.
- Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C., et al. E-cell: software environment for whole-cell simulation. *Bioinformatics (Oxford, England)*, 15(1):72–84, 1999.
- Tossou, P., Wognum, C., Craig, M., Mary, H., and Noutahi, E. Real-world molecular out-of-distribution: Specification and investigation. *Journal of Chemical Information and Modeling*, 64(3):697–711, 2024.
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477, 2019.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
- Vogel, C. and Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews genetics*, 13(4):227–232, 2012.
- Wang, G., Liu, T., Zhao, J., Cheng, Y., and Zhao, H. Modeling and predicting single-cell multi-gene perturbation responses with sclambda. *bioRxiv*, pp. 2024–12, 2024.

- Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.
- Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M., Pairaudeau, G., Pennie, W. D., Pickett, S. D., Wang, J., et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews Drug discovery*, 14(7):475–486, 2015.
- Wenkel, F., Tu, W., Masschelein, C., Shirzad, H., Eastwood, C., Whitfield, S. T., Bendidi, I., Russell, C., Hodgson, L., Ding, J., Fay, M. M., Earnshaw, B., Noutahi, E., and Denton, A. K. TxPert: Leveraging Biochemical Relationships for Out-of-Distribution Transcriptomic Perturbation Prediction. arXiv preprint, 2025. URL <https://arxiv.org/abs/XXXX.XXXX>. Preprint, not yet peer-reviewed.
- Winnifirth, A., Outeiral, C., and Hie, B. Generative artificial intelligence for de novo protein design. *arXiv preprint arXiv:2310.09685*, 2023.
- Winsberg, E. *Science in the age of computer simulation*. University of Chicago Press, 2019.
- Wognum, C., Ash, J. R., Aldeghi, M., Rodríguez-Pérez, R., Fang, C., Cheng, A. C., Price, D. J., Clevert, D.-A., Engkvist, O., and Walters, W. P. A call for an industry-led initiative to critically assess machine learning for real-world drug discovery. *Nature Machine Intelligence*, 6(10):1120–1121, 2024.
- Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pp. 2024–11, 2024.
- Wong, C. H., Siah, K. W., and Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.
- Ye, C., Xu, N., Gao, C., Liu, G., Xu, J., Zhang, W., Chen, X., Nielsen, J., and Liu, L. Comprehensive understanding of *Saccharomyces cerevisiae* phenotypes with whole-cell model *wm\_s288c*. *Biotechnology and Bioengineering*, 117(5):1562–1574, 2020.
- Ye, C., Ho, D. J., Neri, M., Yang, C., Kulkarni, T., Randhawa, R., Henault, M., Mostacci, N., Farmer, P., Renner, S., et al. Drug-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nature communications*, 9(1):4307, 2018.
- Yuan, Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harbor perspectives in medicine*, 6(8):a026583, 2016.
- Zámečník, L. Causal and non-causal explanations in code biology. *Biosystems*, 209:104499, 2021.
- Zhang, J., Ubas, A. A., Borja, R. de, Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan, U., Jones, M. G., Tran, V., Pangallo, J., Papalexi, E., Sapre, A., Nguyen, H., Sanderson, O., Nigos, M., Kaplan, O., Schroeder, S., Hariadi, B., Marrujo, S., Salvino, C. C. A., Gallareta Olivares, G., Koehler, R., Geiss, G., Rosenberg, A., Roco, C., Merico, D., Alidoust, N., Goodarzi, H., and Yu, J. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025a. doi: 10.1101/2025.02.20.639398. URL <https://www.biorxiv.org/content/early/2025/02/24/2025.02.20.639398>.
- Zhang, K., Yang, X., Wang, Y., Yu, Y., Huang, N., Li, G., Li, X., Wu, J. C., and Yang, S. Artificial intelligence in drug development. *Nature Medicine*, pp. 1–15, 2025b.
- Zhou, J., Chen, Y., Hong, Z., Chen, W., Yu, Y., Zhang, T., Wang, H., Zhang, C., and Zheng, Z. Training and serving system of foundation models: A comprehensive survey. *IEEE Open Journal of the Computer Society*, 2024.
- Zuheng, Xu, Jain, M., Denton, A., Whitfield, S., Didolkar, A., Earnshaw, B., and Hartford, J. Automated discovery of pairwise interactions from unstructured data, 2024. URL <https://arxiv.org/abs/2409.07594>.

## A Considerations for benchmarking virtual cells

Here we present both scientific and practical considerations for benchmarking virtual cells. We begin by acknowledging that creating useful benchmarks in biology presents unique challenges not often found in other areas of ML/AI. In particular:

- Ground truth observations are often missing, incomplete, or biased
- Most biological observations are often static, capturing only a single point in time
- Many biological factors are highly dependent on each other, shaped by complex evolutionary pressures
- Biological data is often highly variable, due to a combination of both technical and systemic issues.

These factors make benchmarking virtual cells substantially more difficult than evaluating models in domains like vision or language, where data is more abundant, better characterized, and largely static. Moreover, we argue that they expose critical limitations in many existing benchmarks of virtual cell performance. For example, many of the datasets used in current benchmarks are biased toward transcriptomic readouts, primarily because they are the largest publicly-available resources. However, these datasets typically span only a narrow range of cellular contexts and perturbations. More fundamentally, transcriptomic changes alone capture only a partial view of cellular functional responses, as mRNA levels often do not correlate with protein activity, phenotypic outcomes, or dynamic state shifts, due to post-transcriptional regulation and other factors (Vogel & Marcotte, 2012; Liu et al., 2016; Buccitelli & Selbach, 2020). Despite these limitations, the same transcriptomic datasets are reused across benchmarks without sufficient consideration of their biases, leading to correlated, easy-to-obtain metrics that do not meaningfully drive improvement toward predictive, explanatory, and actionable models.

We also note that it can be challenging to perform better than simple baselines<sup>5</sup>, and that the most performant models are not always the most advanced from an ML perspective, e.g., simple models like scVI (Lopez et al., 2018) already offer strong baselines when making predictions about transcriptomic data (Bendidi et al., 2024). We therefore strongly recommend that all virtual cell models begin their evaluations by comparing to simple baselines in order to assess whether they can sufficiently represent the data space.

In general, we believe that virtual cell benchmarks should be:

- As generally-applicable as possible by adhering to the principles described in Appendix A.1 and Appendix A.2.
- As specifically-defined as possible by being associated with one or more of the capabilities described in Section 3.
- Reviewed and updated frequently in order to maintain relevance as biological knowledge and datasets evolve.

### A.1 Scientific considerations

The following principles address scientific considerations for virtual cell benchmarks:

**Focus on functional response.** Evaluation metrics should prioritize functional outputs that are directly relevant to drug efficacy.

**Favor explainability.** Whenever possible, benchmarks should encourage explainable models, since explainability will drive biological understanding beyond the ability to predict.

**Be biologically consistent.** A virtual cell should recapitulate widely-accepted biology. For example, models should generally respect known relationships between molecular layers (such as typical correlations between transcript level and protein abundance), but also accommodate cases where regulation

---

<sup>5</sup>E.g., the average response of all perturbed training samples.

occurs post-transcriptionally, translationally, or through protein degradation. Benchmarks should penalize grossly-imausible biological predictions without rigidly enforcing incomplete or oversimplified biological graphs.

**Be statistically significant.** Conclusions should be statistically valid and supported by data. Evaluations should be sufficiently powered to declare statistical significance, and not encourage the comparison of mean performance alone (as is typically done using the common “bold table” format).

**Be independent of modality.** To avoid modality-specific biases, benchmarks should be defined in a way that does not require a specific data modality to produce the measure, promoting models that generalize across different types of biological measurements.

**Have biologically informed splits.** When possible, virtual cell benchmarks should design biologically-informed splits, e.g., use time-based splits to simulate knowledge acquisition, or evolution-based splits to account for the relationship between genes and their products.

**Have complementary metrics.** When possible, evaluation should report multiple complementary metrics, as biological systems often behave in counterintuitive ways that are difficult to anticipate. A broad panel of metrics provides a more comprehensive view of model performance and robustness across diverse biological scenarios.

**Have local and global metrics.** A good benchmarking suite should include both specific and high-level tasks to ensure that models are improving locally (e.g., predicting response of single-cell RNAseq readouts in HUVECs) and globally (e.g., recapitulating known biology across cell types).

## A.2 Practical considerations

The following principles address practical considerations for virtual cell benchmarks:

**Update frequently.** The complexity of developing a benchmarking suite demands an iterative approach. With the help of interdisciplinary experts, the design of the entire benchmarking suite, including metrics, tasks, and data splits, needs to be periodically reviewed, refined, and updated based on factors like the desire for additional tasks and the availability of new datasets.

**Encourage prospective evaluation.** Prospective evaluation on a blind test set is the gold standard for unbiased evaluation in ML research. Regularly-organized competitions have proven extremely useful in driving progress by providing opportunities to assess progress and disseminate results that inform future research directions. For example, CASP played a critical role in driving progress on protein structure prediction.

**Simplify adoption.** Benchmarking suites should be made accessible via production-ready, open-source software, with modular, standardized components to encourage adoption, reproducibility, and transparency. Various software modules should provide a foundation for follow-up research and should make independently-generated results more comparable by standardizing the evaluation logic and operating as a source of truth.