How to build the virtual cell with artificial intelligence: Priorities and opportunities

Charlotte Bunne[1,2,3,4,50], **Yusuf Roohani**[1,3,5,50], **Yanay Rosen**[1,3,50], **Ankit Gupta**[3,6], **Xikun Zhang**[1,3,7], **Marcel Roed**[1,3], **Theo Alexandrov**[8,9], **Mohammed AlQuraishi**[9], **Patricia Brennan**[3], **Daniel B. Burkhardt**[11], **Andrea Califano**[10,12,13], **Jonah Cool**[3], **Abby F. Dernburg**[14], **Kirsty Ewing**[3], **Emily B. Fox**[1,15,16], **Matthias Haury**[17], **Amy E. Herr**[16,18], **Eric Horvitz**[19], **Patrick D. Hsu**[5,18,20], **Viren Jain**[21], **Gregory R. Johnson**[22], **Thomas Kalil**[23], **David R. Kelley**[24], **Shana O. Kelley**[25,26], **Anna Kreshuk**[27], **Tim Mitchison**[28], **Stephani Otte**[17], **Jay Shendure**[29,30,31,32], **Nicholas J. Sofroniew**[33], **Fabian Theis**[34,35,36], **Christina V. Theodoris**[37,38], **Srigokul Upadhyayula**[14,16,39], **Marc Valer**[3], **Bo Wang**[40,41], **Eric Xing**[42,43], **Serena Yeung-Levy**[1,44], **Marinka Zitnik**[45,46,47], **Theofanis Karaletsos**[3,*], **Aviv Regev**[2,*], **Emma Lundberg**[3,6,7,48,*], **Jure Leskovec**[1,3,*], **Stephen R. Quake**[3,7,49,*]

[1]Department of Computer Science, Stanford University, Stanford, CA, USA

[2]Genentech, South San Francisco, CA, USA

[3]Chan Zuckerberg Initiative, Redwood City, CA, USA

[4]School of Computer and Communication Sciences and School of Life Sciences, EPFL, Lausanne, Switzerland

[5]Arc Institute, Palo Alto, CA, USA

[6]Department of Protein Science, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden

[7]Department of Bioengineering, Stanford University, Stanford, CA, USA

[8]Department of Pharmacology, University of California, San Diego, San Diego, CA, USA

[9]Department of Bioengineering, University of California, San Diego, San Diego, CA, USA

[10]Department of Systems Biology, Columbia University, New York, NY, USA

[11]Cellarity, Somerville, MA, USA

[12]Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY, USA

*Correspondence: tkaraletsos@chanzuckerberg.com (T.K.), regev.aviv@gene.com (A.R.), emmalu@stanford.edu (E.L.), jure@cs.stanford.edu (J.L.), steve@czbiohub.org (S.R.Q.).

[13]Chan Zuckerberg Biohub, New York, NY, USA

[14]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA

[15]Department of Statistics, Stanford University, Stanford, CA, USA

[16]Chan Zuckerberg Biohub, San Francisco, CA, USA

[17]Chan Zuckerberg Institute for Advanced Biological Imaging, Redwood City, CA, USA

[18]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, USA

[19]Microsoft Research, Redmond, WA, USA

[20]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

[21]Google Research, Mountain View, CA, USA

[22]NewLimit, San Francisco, CA, USA

[23]Schmidt Futures, New York, NY, USA

[24]Calico Life Sciences LLC, San Francisco, CA, USA

[25]Chan Zuckerberg Biohub, Chicago, IL, USA

[26]Northwestern University, Evanston, IL, USA

[27]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany

[28]Department of Systems Biology, Harvard Medical School, Boston, MA, USA

[29]Department of Genome Sciences, University of Washington, Seattle, WA, USA

[30]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA

[31]Seattle Hub for Synthetic Biology, Seattle, WA, USA

[32]Howard Hughes Medical Institute, Seattle, WA, USA

[33]EvolutionaryScale, PBC, New York, NY, USA

[34]Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany

[35]School of Computing, Information and Technology, Technical University of Munich, Munich, Germany

[36]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany

[37]Gladstone Institute of Cardiovascular Disease, Gladstone Institute of Data Science and Biotechnology, San Francisco, CA, USA

[38]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA

[39]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[40]Department of Computer Science, University of Toronto, Toronto, ON, Canada

[41]Vector Institute, Toronto, ON, Canada

[42]Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, USA

[43]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

[44]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[45]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[46]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA

[47]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[48]Department of Pathology, Stanford University, Stanford, CA, USA

[49]Department of Applied Physics, Stanford University, Stanford, CA, USA

[50]These authors contributed equally

SUMMARY

Cells are essential to understanding health and disease, yet traditional models fall short of modeling and simulating their function and behavior. Advances in AI and omics offer groundbreaking opportunities to create an AI virtual cell (AIVC), a multi-scale, multi-modal large-neural-network-based model that can represent and simulate the behavior of molecules, cells, and tissues across diverse states. This Perspective provides a vision on their design and how collaborative efforts to build AIVCs will transform biological research by allowing high-fidelity simulations, accelerating discoveries, and guiding experimental studies, offering new opportunities for understanding cellular functions and fostering interdisciplinary collaborations in open science.

INTRODUCTION

The cell, the fundamental unit of life, is a wondrously intricate entity with properties and behaviors that challenge the limits of physical and computational modeling. Every cell is a dynamic and adaptive system in which complex behavior emerges from a myriad of molecular interactions. Some aspects are remarkably robust to perturbations, such as the elimination of genes or their replacement with homologs from different species. Other aspects are sensitive to even seemingly minor disruptions, such as a point mutation or an external factor that tips cells into dysfunction and disease.

To understand a cell's function, scientists have attempted to construct virtual cell models to simulate, predict, and steer cell behavior.[1–6] Building on this vision, we use the term virtual cell to define a computational model that simulates the biological functions and interactions of a cell. Existing cell models are often rule-based and combine assumptions about the underlying biological mechanisms with parameters fit from observational data. They generally rely on explicitly defined mathematical or computational approaches, such as differential equations,[7–9] stochastic simulations,[10,11] or agent-based models.[12,13] They vary in complexity and cover different defined aspects of cell biology, such as transcription and translation,[14] cytoskeletal driven cell behavior,[15,16] biochemical networks,[17] or metabolic flux.[18,19] The first whole-cell model was developed in 2012, representing all 482 genes and

molecular functions known for an organism: the bacteria *Mycobacterium genitalium*.[8] Since this pioneering work, genome-wide models have been developed to represent other bacterial organisms, including *Escherichia coli*.[8,20–22]

Despite their widespread use in modeling biological systems, approaches to date fall short of capturing many aspects of the operations of both bacterial and more complex systems, such as human cells. Challenges include: (1) Multi-scale modeling: cells operate on multiple scales across both time and space, from atomic to molecular to cellular and histological, with functional properties emerging through nonlinear transformation from one scale to another. (2) Diverse processes with massive numbers of interacting components: cellular function encompasses numerous interacting processes, such as gene regulation, metabolic pathways, and signal transduction. Each process involves a multitude of biomolecular species, in diverse and dynamic configurations and states. (3) Nonlinear dynamics: most cellular processes are highly nonlinear, such that small changes in inputs can lead to complex changes in outputs. Thus, despite progress in modeling specific cellular processes, these factors collectively pose a substantial roadblock to the construction of a virtual cell.

Two exciting revolutions in science and technology—in AI and in omics—now enable the construction of cell models learned directly from data. These parallel revolutions provide an unprecedented opportunity for an ambitious vision of an AI virtual cell (AIVC), a multi-scale, multi-modal, large-neural-network-based model that can represent and simulate the behavior of molecules, cells and tissues across diverse states (Figure 1).

Experimentally, the exponential increase in the throughput of measurement technologies has led to the collection of large and growing reference datasets within and across different cell and tissue systems,[23–25] with data doubling every 6 months for the past several years,[26] along with the ability to couple these measurements with systematic perturbations.[27–29]

Computationally, concurrent advances in AI have enhanced our ability to learn patterns and processes directly from data without needing explicit rules or human annotation.[30,31] Such modeling paradigms have been used successfully in the biomolecular realm, for example, to predict three-dimensional (3D) molecular structures from sequences[32–34] and interactions between different molecular components.[35–38] Recent modeling methodologies in AI provide representation and inference tools that satisfy the trifecta of being predictive, generative, and queryable, which are key utilities for advancing biological research and understanding. By building on these properties, we argue that we now have the methods to develop a fully data-driven neural network-based representation of an AIVC that can accelerate research in biomedicine by enabling fast-paced *in silico* studies, as well as powerful bridges between computational methods and confirmatory wet-lab experimentation (Figure 1).

The creation of an AIVC will enable a new era of high-fidelity simulation in biology, in which cancer biologists model how specific mutations transition cells from healthy to malignant; developmental biologists forecast how developmental lineages evolve in response to perturbations in specific progenitor cells; and microbiologists predict the effects of viral infection on not just the infected cell but also its host organism. AIVCs will empower

experimentalists and theorists alike, by transforming the means by which hypotheses are generated and prioritized and allowing biologists to span a dramatically expanded scope, better fitting the enormous scales of biology. Although the cellular models may not always directly identify mechanistic relationships, they can be viewed as tools for effectively narrowing the search space for mechanistic hypotheses, thereby accelerating the discovery of underlying factors behind cellular function.

This perspective article is based on extensive community discussions, including a workshop hosted by the Chan Zuckerberg Initiative, and aims to ignite the formation of a collaborative research agenda for a large-scale, long-term initiative with a roadmap for developing, implementing, and deploying AIVCs. We describe a vision catalyzed by emerging advances in AI in cell biology and their application to constructing virtual representations of cells. We lay out priorities and opportunities across data generation, AI models, benchmarking, interpretation, and ensuring biological veracity and safety (Box 1). By encouraging interdisciplinary collaborations in open science—spanning academia, philanthropy, and the biopharma and AI industries—we posit that a comprehensive understanding of cellular mechanisms is within reach. AIVCs have the potential to revolutionize the scientific process, lead to the understanding of novel biological principles, and augment human intelligence to underpin future breakthroughs in programmable biology, drug discovery, and personalized medicine (Box 2).

AIVCs

Our view of an AIVC is a learned simulator of cells and cellular systems under varying conditions and changing contexts, such as differentiation states, perturbations, disease states, stochastic fluctuations, and environmental conditions (Figure 1). In this context, a virtual cell should integrate broad knowledge across cell biology. VCs must work across biological scales, over time, and across data modalities and should help reveal the programming language of cellular systems and provide an interface to use it for engineering purposes.

In particular, an AIVC needs to have capabilities that allows researchers to (1) create a universal representation (UR) of biological states across species, modalities, datasets, and contexts, including cell types, developmental stages, and external conditions; (2) predict cellular function, behavior, and dynamics, as well as uncover the underlying mechanisms; and (3) perform *in silico* experiments to generate and test new scientific hypotheses and guide data collection to efficiently expand the virtual cell's abilities.

Next, we elaborate on these key capabilities and discuss approaches for how to achieve them.

URs

An AIVC would map biological data to UR spaces (Figure 1A), facilitating insights into shared states and serving as a comprehensive reference. These URs should integrate across three physical scales—molecular, cellular, and multicellular—and accommodate contributions from any relevant modality and context (Figure 1A). This integration will allow researchers to complement new data with existing information within the AIVC,

leveraging its extensive biological knowledge to bridge gaps between different data. Such a comparison with prior data would provide a comprehensive context for every analysis.

Importantly, the multilevel representation should generalize to new states that are not present within the data used to train the AIVC. Such an emergent capability would unlock discoveries about biological states that have not been directly observed or might not even occur in nature. For instance, the AIVC's exposure to similar instances during training, such as inflammatory states in macrophages, might enable it to predict previously unknown inflammatory states in microglia. Additionally, the AIVC should be able to predict novel states resulting from interventions (or, equivalently, interventions needed to achieve a novel specified state) offering a range of downstream applications in cell engineering and synthetic biology.

Predicting cell behavior and understanding mechanisms

A defining function of an AIVC will be its ability to model cellular responses and dynamics. By training on a wide range of snapshots, time-resolved, non-interventional, and interventional datasets collected across contexts and scales, the AIVC can develop an understanding of the molecular, cellular, and tissue dynamics that occur under natural or engineered signals. These signals include external and internal stresses or other factors such as chemical (e.g., small molecules) or genetic (engineered or natural) perturbations and their combinations. An AIVC should be able to predict responses to perturbations that have not been previously tested in the lab, while also accounting for the specific features of the cellular context within which the perturbation is being tested.

The AIVC should also have the capability to simulate the temporal evolution of alterations in cell states in response to both intrinsic and extrinsic factors, along with the resulting multicellular spatial arrangements. By modeling the transient nature of the overall cell state and the continuous flux in cellular conditions, the AIVC could uncover previously unstudied trajectories in diverse dynamic processes, such as development, maintenance of homeostasis, pathogenesis, and disease progression. Another critical challenge is understanding the molecular mechanisms underpinning observed phenotypes and trajectories. The AIVC could propose potential causal factors behind phenotypes by simulating the effects of different interventions. Through its multi-scale design, the AIVC should be able to extrapolate the basis of cellular function across scales and link intracellular processes to phenotypes at the cell and tissue level. Thus, the AIVC opens new avenues for investigating mechanisms linked to diverse phenotypes and behaviors.

Although uncovering a phenotype's causal factors may not always be feasible through computation alone, the AIVC has the potential to reduce the space of possible hypotheses. Through simulating the effects of different interventions, the AIVC could propose potential causal factors behind phenotypes with corresponding degrees of uncertainty, allowing scientists to validate claims experimentally.

In silico experimentation and guiding data generation

For real-world utility, a defining function of an AIVC will be its ability to guide data generation and experiment design. An AIVC should be queryable with computational twins

of today's laboratory experiments, here called virtual instruments (VIs). Virtual experiments could, for example, simulate experiments in a cell type that is challenging to cultivate *in vitro* or simulate expensive readouts from low-cost measurements, such as single-cell transcriptomes from label-free imaging.[39] Virtual experiments could also be used to screen a vast number of possible perturbagens at a scale that would be impossible in the lab. Such capabilities are invaluable when considering the exponentially larger search space of combinatorial perturbations involving more than one perturbagen.[40–44]

AIVCs will usher in a new pradigm of how computational systems are probed during the design of new biological experiments. In this framework, an AIVC would not only design experiments to validate specific scientific hypotheses but also to enhance its own capabilities. Equipped with the ability to assign confidence values to its predictions, an AIVC could enable interactive querying to guide experimentalists to the most efficient path for generating additional data for fine-tuned improvement in low-confidence areas. Extended to an active and iterative lab-in-the-loop process, we envision efficient and focused expansion of the AIVC's performance. Ultimately, the AIVC might even be able to identify key gaps in its own understanding of biology and propose the most efficient paths to bridge them.[45–47]

BUILDING THE AIVC

We envision an AIVC as a comprehensive AI framework composed of several interconnected foundation models that represent dynamic biological systems at increasingly complex levels of organization—from molecules to cells, tissues, and beyond. Our approach has two main components: (1) a universal multi-modal multi-scale biological state representation and (2) a set of VIs, which are neural networks that manipulate or decode these representations. Although there may be other approaches to building an AIVC, we believe this approach would provide a scaffold that can be scaled in a collaborative and open way.

We use the term UR to refer to an embedding produced by a multi-modal AIVC foundation model. An embedding is a learned numerical representation of data in a continuous vector space. The AIVC transforms high-dimensional multi-scale multi-modal biological data into embeddings that retain meaningful relationships and patterns.

The AIVC can capture cell biology at three distinct physical scales by representing (1) molecules and their structures found within individual cells, (2) individual cells, as spatial collections of those interacting molecules and structures, and (3) how individual cells interact with one another and the non-cellular environment in a tissue. Each of these scales is represented by a distinct UR, building on abstractions generated by the previous layer, thus linking the different scales.

In the context of UR, VIs are neural networks that take URs as input and produce a desired output. We describe two types of VIs: decoder VIs (or decoders) that take a UR as an input and produce human-understandable output, for example, a cell type label or a synthetic microscope image, and manipulator VIs (or manipulators), which take a UR as an input and

produce another UR as an output, for example, that of an altered cell state after perturbation. Because these instruments will operate over the same representations, they can be shared and reused across different use cases, experiments, and datasets. Thus, we envision that any scientist will be able to build a VI on top of a UR and share it with the community. The building of VIs that closely resemble real instruments, such as a microscope, has the potential to seed the development of instrument-specific lab-in-the-loop systems.

Building UR across physical scales

The AIVC would be a multi-scale foundation model that learns distinct representations of biological entities at each physical scale (Figure 2C). These representations can be aggregated together and transformed to produce representations at the next higher physical scale. This recurring architectural motif can be applied from the level of individual molecules to the scale of entire tissues and organs granting the model consistency across biological scales (Figure 2A). Each representation applies universally to a specific class of biological entities. This abstraction allows the virtual cell to seamlessly evolve and incorporate new data—whether from new modalities or from out-of-distribution sources—within this general framework.

In the following sections, we discuss design principles and data that could be used to construct each physical scale of the AIVC bottom-up. Although many existing machine learning architectures could be applied directly to the task of learning functional representations of cellular components (Box 3), we additionally suggest the incorporation of biological inductive biases into the design of these representations, and further modeling innovations should drive the refinement and success of these models.

Molecular scale: The first layer of the virtual cell represents individual molecular species (Figures 2A and 2C). Although there are many different classes of molecules present in a cell, a starting point for the AIVC will be to model the three types of molecules of the central dogma: DNA, RNA, and proteins. These can all be represented as sequences of characters—nucleotides or amino acids.[48–53] Such sequence data are particularly well suited for AI methods originally developed for natural language processing, such as large language models (LLMs) (Box 3). Given the high-throughput measurement capabilities for genomic sequences, there are substantial and growing amounts of training data available. This abundance of data, combined with simple objective functions (such as predicting masked letters in a sequence), provides the key ingredients for effectively training models to generate an initial molecular UR. Furthermore, a biological language model could be trained on all three modalities simultaneously, thus maximizing interoperability and training corpus size. Despite its inherent compatibility with transformers, specific considerations around masking and attention mechanisms must be addressed when applying these models to biological sequence data as opposed to natural language. Although language modeling approaches have been extensively studied for these core molecules and have proven successful for some of their chemical modifications[54] and various other molecules, such as glycans, lipids, and metabolites,[55,56] they may struggle with other molecular constituents of the cell. Such modeling difficulties might be exacerbated for data that are difficult to fit into a sequence or very small molecules. Given that the primary building blocks of these

entities are atoms, a neural network trained to model molecules at the atomic level[32,57] could be a more general choice for this layer. However, models with atomic resolution introduce a substantial computational burden and might be constrained by the limited availability of training data. Although atomic-based modeling is highly accurate for many static structures, it cannot yet represent the full dynamic range of chemistry that occurs at this scale. Therefore, a broader, evolutionarily informed representation such as that of sequences may be preferred.

Cellular scale: The next level of abstraction models individual cell states (Figures 2A and 2C). As cellular function is underpinned by the molecular interactions and signaling networks formed in a cell, a cellular UR can be built using representations of molecular and other (e.g., imaging) features, describing the organization and abundance of molecular components. The key step here would be to integrate learned representations of molecules with their quantities and appropriately abstracted locations and timestamps to create a unified representation of the cell.[58–60]

Data for the cellular UR consist of measurements mapped to a single-cell level, such as measurements of the transcriptome (single-cell RNA sequencing [scRNA-seq]), chromatin accessibility (scATAC-seq), chromatin modification, transcription factor binding, and proteome.[61] Imaging technologies measure cell morphology at subcellular resolution, often together with molecular information.[29,62,63] For example, fluorescence confocal microscopy can help resolve the subcellular location of the human proteome.[64] Live-cell imaging[65] enables the study of proteins in living cells using time-lapse microscopy. Cryoelectron microscopy determines biomolecular structures at near-atomic resolution.[66,67] Super-resolution microscopy offers deeper insights into molecular processes through single-molecule imaging in living systems.[68,69] Complementing imaging approaches, mass spectrometry, and proximity-dependent labeling can unveil protein-protein associations and provide deeper insights into cell structure and signaling network rewiring.[70,71]

From a model architecture perspective, vision transformers[72] or models leveraging convolutional neural networks (CNNs)[73,74] are widely applicable to biological images to model across multiple imaging channels capturing different biological features,[75,76] while being robust to distribution shift and batch variability.[77] Autoencoders and transformers have been successfully applied for learning representations for sequence-based data.[59,78,79] Using AI algorithms to integrate different data modalities collected with sequencing and imaging technologies creates a multi-view model of the cell that can be both dynamic and predictive.[80,81]

As the AIVC model grows in complexity, it is crucial to also model cellular organelles and membraneless compartments[82] as units that play specific roles within the cell. Robustly capturing the functions of these units is vital to ensure accurate predictions, mechanistic interpretability, and model generalizability.

Given their prevalence, the cellular UR will initially rely on transcriptomics measurements, whereas imaging modalities will be key for continued modeling of cellular spatial organization and dynamics.

Multicellular scale: At the third layer of abstraction, the AIVC models the organization of cells into a multicellular UR (Figures 2A and 2C). This layer allows for the exploration of how cell-cell interactions, largely governed by spatial proximity, combine into tissues, organs, and, ultimately, whole organisms. Multicellular interactions can be analyzed after tissue dissociation (such as in scRNA-seq)[83] or *in situ* in a 2D section or 3D volume, where the tissue structure is preserved. Building the AIVC will require integration across available modalities that provide spatial insights, i.e., both spatial molecular profiling, as well as non-molecular tissue imaging data.

There are multiple methods to profile the spatial location of RNA[84] and proteins[85] in cells, along with various imaging methods for select molecular species (e.g., immunohistochemistry) or with stains for tissue structure alone (e.g., hematoxylin and eosin [H&E]). Spatial molecular biology is currently a very active area of research and method development. Although publicly available data are still limited, we foresee a rapid development in this domain providing multi-omic 2D and 3D datasets. A more generalized data generation effort together with open frameworks for spatial data[86] could greatly accelerate modeling at the multicellular scale.

The relative organization of cells within a 2D tissue section and 3D tissue volume can be represented using a graph or point cloud. The multicellular UR can be derived from such data using graph-learning techniques, such as graph neural networks (GNNs) and equivariant neural networks (ENNs). For image-based data, convolutional neural networks or vision transformers can be applied (Box 3).

Predicting cell behavior and understanding mechanisms

VIs are the "tools" that operate on UR embeddings and perform various functions and tasks. By altering URs of molecules, cells, and tissues, manipulators can abstract complex dynamic processes (Figure 2B) more simply as transitions between (distributions of) their representations (Figure 2D). Similarly, decoders can take an embedding of biological entities and predict one or more concrete properties, for example, physical structure, cell type/state, fitness, expression, or drug response.

The design of a wide array of manipulators provides us with an unprecedented set of tools for modeling cell behavior and dynamics: generative AI approaches, such as diffusion models[87] or autoregressive transformers,[88] i.e., model architectures that capture heterogeneity and parameterize continuous time dynamics, can predict a future state or evolution of a cell or molecular state (Box 3).[57,89] Using integrated data from time-lapse imaging,[65] gene expression profiles,[83] and other modalities, manipulators can allow inferring the phenotypic progression from stem cell to differentiated cell, while capturing the influence of both genetic factors and environmental conditions—through learned interpolations and extrapolations between multi-scale URs of different cell states. Similarly, they allow predicting the effect of treatments on patients, given a virtual representation of a patient's molecular profile.

Furthermore, variations in cellular URs can be linked to corresponding changes in molecular states or their spatial localization, influenced by downstream factors, such as genetic variants

or functional changes in proteins, which are represented in a lower scale of the AIVC. Leveraging the ability of manipulators to model temporally resolved molecular and cellular events, decoders of the AIVC could potentially identify cellular components, molecular pathways, and their interactions that contribute to each prediction and process. As such, the multi-scale design of the AIVC may unveil mechanistic hypotheses of such processes.

Despite the remarkable advancements in protein modeling, the field continues to struggle in modeling dynamic molecular processes using foundation models. There will likely be areas of cell modeling, including dynamics, which pose similar challenges. For instance, the modeling of intricate networks of transient and weak molecular interactions, which play a crucial role in rapid fine-tuning of cellular signaling and formation of cell biological features such as condensates, may pose similar challenges. Consequently, we foresee a need for advanced data collection and modeling methodologies capable of capturing the dynamics of cellular processes, akin to those encountered in protein modeling. At the same time, although some functionalities of the AIVC heavily depend on such solutions, others (e.g., certain predictive functionalities) may be successful even without them. That is one of the appealing properties of multi-modal AI models with emergent properties and why developing the AIVC now is so compelling.

In silico experimentation and guiding data generation

Manipulator VIs operating in the UR space could further enable the exploration of a broad range of hypotheses through *in silico* experiments that virtually perturb a cell model. This might be achieved by predicting changes in the URs following a perturbation prompt (Figure 2D).[40,42–44]

The design of manipulators that predict transitions in the UR upon an *in silico* input can build on conditional generative models: deep learning architectures such as conditional deep generative models[31] allow generating the desired UR based on the property or context of interest (Box 3). Here, high-throughput perturbation screens—based on RNA-seq,[28,83,90] optical pooled screens (OPS),[29,39,91] or other technologies—offer a rich resource through which the AIVC can be trained to predict these effects. By conditioning on specific perturbations—such as environmental changes, genetic mutations, or chemical treatments— the generative model might produce a new UR reflecting the predicted cellular response. This conditioning could be achieved through learned or pre-computed embeddings of the affected molecular targets. Chemical compounds, small molecules, and metabolites could be embedded based on their chemical properties. Additionally, LLMs trained on comprehensive scientific literature and biological databases, such as Gene Ontology or drug banks, could further provide a rich contextual background used for conditioning the generative model, e.g., through considering wide range of interactions and side effects.

VIs can be designed so that predictions are accompanied by estimates of model uncertainty.[92] Under a Bayesian formulation of its predictive function, the predictions made for cell perturbation outcomes could include an uncertainty score, either implicitly via inference, deep kernels,[93,94] or through explicit estimation of the full posterior over model parameters.[95,96] Some practical approaches utilize model ensembles[97] or conformal predictions.[98,99] By assigning specific confidence levels to its predictions, the AIVC can call

methods for computing the expected value of additional data or approximations referred to in machine learning as active learning to guide experimental data collection[45] for expanding its UR. Alternatively, methods for computing the expected value of information could be used to guide data generation with the goal of optimizing a desired biological property.[92] Lastly, through its ability to conduct *in silico* experiments and suggest additional informative experiments, the AIVC could become an integrative part of lab-in-the-loop schemes. This allows not only for a seamless experimental validation of its predictions but also a sequence of experiments, predictions, and generations of hypotheses that gradually improve our systematic understanding of molecular circuits that drive biological functions.

DATA NEEDS AND REQUIREMENTS

A key consideration for the AIVC is which datasets and modalities must be collected to enable its effective construction. Unlike traditional experimental design, where data are generated to test specific scientific hypotheses, data collection for training the AIVC should be focused on ensuring the broad applicability and generalizability expected of the AIVC. To meet these ambitions, data would ideally span different domains and modalities, capture the heterogeneity and diversity of biological variability, and enable models to distinguish between technical (measurement) noise, stochastic biological variation, and physiological differences.

Data generation will require simultaneous exploration of temporal and physical scales, while allowing for system perturbations. Here, classical imaging technologies,[65,100,101] including live-cell, and newer structural imaging technologies, such as cryoelectron tomography and soft X-ray tomography,[66,102,103] as well as novel spatial omics technologies,[104,105] offer opportunities to model biomolecules and functions across scales. Furthermore, biological processes span a vast range of timescales, from the fastest reactions happening in picoseconds to a cell division progressing over hours to a day, tumor development occurring over years, and neurodegeneration over decades. The recent construction of universal cell atlases[101,106] may serve as a powerful resource for modeling cellular behavior over longer timescales, such as tissue formation. New approaches will be needed to build comparable datasets that capture the behavior of cells on shorter timescales, e.g., through methods such as live-cell imaging. Besides molecular measurements, an important aspect of data collection will lie in the measurement of biophysical and biochemical cellular properties to provide boundaries of physical and chemical realism to the AIVC.

Another important driver for the development of AIVCs will be multi-modal datasets. For example, datasets that bridge molecular and spatial scales, such as single-cell transcriptomics data combined with histology to understand how cells interact and what molecular signatures underpin the formation of specialized spatial niches.[107] Further technological development is needed to collect multi-modal data that better capture the relationship between molecular signatures, cell behavior, cellular regulation, and organization.

Although a core interest of virtual cell modeling will focus on human datasets for the purpose of understanding disease and aiding the development of novel therapeutics, human

datasets are limited in our ability to perform controlled experimentation and perturbations *in vivo*.

Here, the field of 3D tissue biology, including culture systems, such as organoids, is emerging as a tool to study the complexities of tissue architecture and function[108] in a 3D environment, while allowing perturbations of the system. Another critical avenue to surpass this limitation will be to perform diverse, organism-wide profiles of species spanning evolutionary history, across perturbations and under various conditions.[109–111] Ideally, large datasets could be collected across all three physical scales, allowing the AIVC to extend beyond disease research into other areas such as industrial biotechnology, agricultural biotechnology, infectious diseases, and climate change. However, based on data collection trends for the cellular and multicellular scales, modeling animal cells remains the most realistic.

Finally, a key aspect of biological data generation will be the exploration of combinatorial spaces: biological spaces are commonly high dimensional, and enumerating their variants is intractable in general, e.g., when considering all possible variants of a genome. Even for combinations of a small number of entities, exemplified in the case of enumerating pairs or sets of perturbations,[47,90] experimental design becomes exceedingly challenging. Because combinatorial possibilities quickly expand well beyond what is practical experimentally, or even computationally, new methods for their exploration must be developed.

How much data are needed to build the AIVC?

The scale of raw biological data is undeniable, but so is the sheer nominal size of even one human cell system, making first principle estimates challenging. For instance, the Short Read Archive of biological sequence data holds over 14 petabytes of information,[112] which is more than 1,000 times larger than the dataset used to train ChatGPT.[113] Large parts of these data may be redundant or have diminishing returns if used for training, and the scaling laws for models' performances must be investigated thoroughly.

In addition to data size, data diversity and quality are critical to ensure model performance.[114] Data from humans and model organisms, such as mice and *Escherichia coli*, are unequally represented in sequence and literature databases, which when used for training, encode strong species biases.[114] Other biases, for example, in terms of sex, specific diseases, or human ancestral populations could also reduce the impact of AIVC models.[115]

Although efforts on the data side are required, the AI models driving the AIVC must be designed to withstand and adapt to these challenges, i.e., exhibit robustness in their ability to integrate datasets of various origins and quality. This is crucial given both the rapid pace of advances in lab technologies (which preclude standardization on a single platform) and the broad diversity of modalities and cell systems that must be encompassed by the AIVC. As virtual cell efforts mature, the dialog between the scientists who develop models, those who generate experimental data, and funding organizations must be further intensified.

MODEL EVALUATION

A more important question for the development of AIVCs may not be "how do we build them?" but rather "how do we build trust in their competence and fidelity?" To this end, a comprehensive and adaptable benchmarking framework will be needed. Although various frameworks already exist for tackling specific biological questions (for example, protein structure prediction models[89] were developed in the context of the CASP evaluation framework), the AIVC will need to demonstrate generalizability across numerous biological contexts and downstream tasks. It must account for dynamic distributions that evolve due to environmental changes, infections, genetic variants, and other such factors causing distribution shifts.[116]

Even beyond generalizability, emergent capabilities, such as those associated with LLMs, could enable AIVC models to extrapolate to truly out-of-distribution data. In a biological context it may be difficult to decide how this boundary is defined during evaluation. New molecules, new cell states, and even new species could be considered within the training distribution. A new molecule could have homologs, including remote homologs, within the dataset. A new cell type or state could execute gene programs and regulatory networks found in existing cell types. A new strain could be closely related to existing species in the training data or live in similar environmental niches. Extrapolation to new data could then be limited to consider only the design of biological entities that do not naturally occur. This type of evaluation is already considered within the molecular design space because language-model-created proteins, such as esmGFP[52] or OpenCrispr1,[53] highlight how different they are from any of their naturally occurring counterparts. If extrapolation is a goal when designing these models, it is possible that additional inductive biases, fine tuning, or preference optimization using biomechanical, physics-based, or mechanistic modeling[117] would prove necessary.

The evaluation of AIVCs should prioritize both generalizability, as well as discovering new biology. Generalizability measures how well the model performs in unseen contexts, such as novel cell types and genetic backgrounds. It can be evaluated through a cross-modal reconstruction task, such as predicting gene expression given the morphology of a previously unseen cell or the next image in a sequence of microscopy images of cell state. Assessing generalizability builds confidence in the AIVC's ability to capture core biological processes and understand how they vary across different contexts. Establishing such cross-modal benchmarks to link scales and modalities in cell biology is of imminent priority to the research community because these tasks are both biologically useful and well defined.

Ultimately, AIVC models should be judged on their ability to unlock new ways of understanding biology. Such an evaluation will ensure that model development is aligned with biological relevance. The most useful initial accomplishments will likely be to generate valuable testable hypotheses. For this purpose, validation datasets that are related to phenotypes that are experimentally verifiable may be suitable, such as growth rate of cells, molecular profiles, disrupted protein-protein interactions, or transcription factor binding.

As the capabilities of AIVCs improve, we must consider whether statistical measures of performance are adequate, or if interpretability and biological causality would be core requirements.

INTERPRETABILITY AND INTERACTION

One of the hallmarks of scientific discovery in biology has been the creation of mechanistic models of a phenomenon under observation. When creating virtual cells, we may have to forgo our ability to build fully mechanistic models in favor of learning interactions that will generalize from data and predict beyond the observations. However, it is still desirable to strive toward increased interpretability.

Every AIVC prediction could be substantiated with the corresponding multi-scale interactions that determine resulting states, e.g., understanding how a cellular subsystem or protein complex is disrupted in a diseased tissue can aid development of therapeutic interventions.[118,119] The modular structure of the AIVC will enable researchers to pinpoint specific genes, proteins, or molecular processes involved in each predicted behavior. Patterns in the wiring of large models can also be leveraged to uncover combinatorial biological interactions, such as those between proteins, which can be projected to interpretable spaces without restricting the generality of the original model. Although many capabilities of the AIVC rely on predictive tasks, generating mechanistic hypotheses could provide experimental routes to understand and explore the AIVC's predictions further and will be vital for the adoption and use of AIVCs.

Ultimately, it will be of key interest to build an interactive layer for the AIVC that enables researchers of varying expertise to grasp and utilize its predictions effectively. AI agents, built using LLMs, could serve as virtual research assistants, providing an intuitive interface for non-experts.[46,120] Leveraging their extensive knowledge of scientific literature, these language models can offer deeper insights into the predictions made by the AIVC.

AN OPEN COLLABORATIVE APPROACH

Creating an AIVC requires tremendous investment, diverse backgrounds, and many iterations and can only be advanced by a concerted open science effort. As a scientific community, we must strive to ensure that both the development and usage of virtual cells are accessible and responsive to the entire scientific community. These efforts would greatly benefit from open data resources and data standards, a collaborative platform for cell modeling, and, especially, open benchmark datasets and common validation strategies to ensure their biological fidelity and real-world utility. Such a collaborative program could greatly accelerate progress across individual efforts and unify scientific research at a global scale, connecting myriad smaller-scale efforts.

To achieve this, multiple key parameters need to be considered. First, we must ensure that AIVCs represent and benefit all of humanity, with open data that captures human ancestral, sex, and geographic diversity.[121] Ensuring that such datasets reflect human diversity, while safeguarding individuals' privacy is a principal challenge. Second, as the size of AIVC models increases, the cost of training, fine tuning, or using them as is will also grow.

Investments in diverse data collection, infrastructure, and a platform for hosting virtual cell models will be critical to ensure representation, accessibility, and benefit to the broader scientific community. The platform should foster open and collaborative development of AIVCs, enabling active collaboration between biologists, clinicians, statisticians, and computer scientists. This platform should facilitate swift iterations between the lab and the modeling environment and offer opportunities to quickly test and benchmark new models. Third, synergistic collaboration among stakeholders is needed across the biomedical ecosystem, including philanthropy, academia, biopharma, and the AI industry. Pre-competitive collaborations can greatly accelerate our collective progress toward creating AIVCs. Besides the synchronization with data generators and other modeling efforts, collaboration with regulatory authorities and bioethics experts are crucial for benchmarking and establishing new norms that will expedite the deployment of AIVCs, while complying with legal requirements and setting standards for ethical issues for responsible use of virtual cells.

This article is intended to serve as a primer for the formation of a collaborative research agenda and roadmap for a large-scale, long-term initiative for developing and implementing AI-powered VCs. If successful, such interactive AIVC models, capable of simulating cellular biology, have the potential to fundamentally change how cell biology research is done. We foresee a future where AIVC platforms function as open, interconnected hubs for collaborative development and broad deployment of cell models to researchers but also as education hubs delivering training to researchers, as well as providing engagement activities for educators, patients, and the public.

OUTLOOK AND REASONS FOR OPTIMISM

The genetics and genomics communities have created large reference datasets, such as the human genome project,[23] HapMap,[122] the Cancer Genome Atlas (TCGA),[123] ENCODE,[124] the Genotype-Tissue Expression (GTEx) project,[125] the Human Protein Atlas (HPA),[64,126] the Human Cell Atlas (HCA),[24] and a growing number of deeply phenotyped, population-scale biobank efforts.[127] Thanks to these projects, massive reference data are now available to train machine learning models. Although these efforts will continue to grow, they also catalyze a new, parallel effort: creating a virtual simulation of cell biology, a new process for scientific inquiry.

The result, the AIVC has the potential to revolutionize the scientific process, leading to future breakthroughs in biomedical research, personalized medicine, drug discovery, cell engineering, and programmable biology. Acting as a virtual laboratory, the AIVC could facilitate a seamless interface between data derived from *in silico* experimentation and results from physical laboratories. As such, we expect the AIVC to contribute to a more unified view of biological processes, fostering alignment among scientists on how emergent properties in biology arise.

By bridging the worlds of computer systems, modern generative AI and AI agents, and biology, the AIVC could ultimately enable scientists to understand cells as information processing systems and build virtual depictions of life. As the AIVC expands the

understanding of cellular and molecular systems, it will also increasingly allow us to program them and design novel synthetic ones. AI models have already been used to design new CRISPR enzymes,[53] functional proteins,[128] and even entire prokaryotic genomes.[51] The rapid progress in the precision of cell and genome engineering tools will accelerate this shift and different instantiations of the AIVC will compete in their ability to engineer new, functional biology capabilities as much as in their ability to represent and simulate biology.

Finally, we staunchly advocate the role for open science approaches, where the scientific community readily shares data, models, and benchmarks, where findings and insights are contextualized, and where a climate of perpetual improvement is fostered. We welcome and encourage all stakeholders across sectors and domains to engage in this endeavor. With a massive scientific undertaking and shared goals, open sharing of insights, and the power of safe, ethical, and reliable AI, we believe that we are stepping into a new era of scientific exploration and understanding. The confluence of AI and biology, as encapsulated by AIVCs, signals a paradigm shift in biology and shines as a beacon of optimism for unraveling multiple mysteries of the cell.

ACKNOWLEDGMENTS

REFERENCES

1. Slepchenko BM, Schaff JC, Macara I, and Loew LM (2003). Quantitative cell biology with the Virtual Cell. Trends Cell Biol. 13, 570–576. [PubMed: 14573350]

2. Johnson GT, Agmon E, Akamatsu M, Lundberg E, Lyons B, Ouyang W, Quintero-Carmona OA, Riel-Mehan M, Rafelski S, and Horwitz R (2023). Building the next generation of virtual cells to understand cellular biology. Biophys. J 122, 3560–3569. [PubMed: 37050874]

3. Marx V (2023). How to build a virtual embryo. Nat. Methods 20, 1838–1843. [PubMed: 38057522]

4. Goldberg AP, Szigeti B, Chew YH, Sekar JA, Roth YD, and Karr JR (2018). Emerging whole-cell modeling principles and methods. Curr. Opin. Biotechnol 51, 97–102. [PubMed: 29275251]

5. Georgouli K, Yeom J-S, Blake RC, and Navid A (2023). Multi-scale models of whole cells: progress and challenges. Front. Cell Dev. Biol 11, 1260507. [PubMed: 38020904]

6. Marucci L, Barberis M, Karr J, Ray O, Race PR, de Souza Andrade M, Grierson C, Hoffmann SA, Landon S, Rech E, et al. (2020). Computer-aided whole-cell design: Taking a holistic approach by integrating synthetic with systems biology. Front. Bioeng. Biotechnol 8, 942. [PubMed: 32850764]

7. Lauffenburger DA, and Linderman JJ (1996). Receptors: models for binding, trafficking, and signaling (Oxford University Press).

8. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, and Covert MW (2012). A whole-cell computational model predicts phenotype from genotype. Cell 150, 389–401. [PubMed: 22817898]

9. Mangan S, and Alon U (2003). Structure and function of the feed-forward loop network motif. Proc. Natl. Acad. Sci. USA 100, 11980–11985. [PubMed: 14530388]

10. Zopf CJ, Quinn K, Zeidman J, and Maheshri N (2013). Cell-cycle dependence of transcription dominates noise in gene expression. PLoS Comput. Biol 9, e1003161. [PubMed: 23935476]

11. Eling N, Morgan MD, and Marioni JC (2019). Challenges in measuring and understanding biological noise. Nat. Rev. Genet 20, 536–548. [PubMed: 31114032]

12. Hellweger FL, Clegg RJ, Clark JR, Plugge CM, and Kreft J-U (2016). Advancing microbial sciences by individual-based modelling. Nat. Rev. Microbiol 14, 461–471. [PubMed: 27265769]

13. Gorochowski TE (2016). Agent-based modelling in synthetic biology. Essays Biochem. 60, 325–336. [PubMed: 27903820]

14. Thiele I, Jamshidi N, Fleming RM, and Palsson B (2009). Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. PLoS Comput. Biol 5, e1000312. [PubMed: 19282977]

15. Odell GM, and Foe VE (2008). An agent-based model contrasts opposite effects of dynamic and stable microtubules on cleavage furrow positioning. J. Cell Biol 183, 471–483. [PubMed: 18955556]

16. Popov K, Komianos J, and Papoian GA (2016). MEDYAN: mechanochemical simulations of contraction and polarity alignment in actomyosin networks. PLoS Comput. Biol 12, e1004877. [PubMed: 27120189]

17. Burke PEP, Campos CBL, Costa LDF, and Quiles MG (2020). M. G. A biochemical network modeling of a whole-cell. Sci. Rep 10, 13303. [PubMed: 32764598]

18. Li G, Liu L, Du W, and Cao H (2023). Local flux coordination and global gene expression regulation in metabolic modeling. Nat. Commun 14, 5700. [PubMed: 37709734]

19. Fang X, Lloyd CJ, and Palsson BO (2020). Reconstructing organisms in silico: genome-scale models and their emerging applications. Nat. Rev. Microbiol 18, 731–743. [PubMed: 32958892]

20. Stevens JA, Grünewald F, van Tilburg PAM, König M, Gilbert BR, Brier TA, Thornburg ZR, Luthey-Schulten Z, and Marrink SJ (2023). Molecular dynamics simulation of an entire cell. Front. Chem 11, 1106495. [PubMed: 36742032]

21. Maritan M, Autin L, Karr J, Covert MW, Olson AJ, and Goodsell DS (2022). Building structural models of a whole mycoplasma cell. J. Mol. Biol 434, 167351. [PubMed: 34774566]

22. Ahn-Horst TA, Mille LS, Sun G, Morrison JH, and Covert MW (2022). An expanded whole-cell model of E. coli links cellular physiology with mechanisms of growth rate control. npj Syst. Biol. Appl 8, 30.

23. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. (2001). The sequence of the human genome. Science 291, 1304–1351. [PubMed: 11181995]

24. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M, et al. (2017). The human cell atlas. eLife 6, e27041. [PubMed: 29206104]

25. CZI Single-Cell Biology Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, Bezzi E, Batuhan C, Jim C, Chambers S, et al. (2023). CZ CELL×GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Preprint at bioRxiv.

26. Heimberg G, Kuo T, DePianto D, Heigl T, Nathaniel D, Salem O, Scalia G, Biancalani T, Turley S, Rock J, et al. (2023). Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. Preprint at bioRxiv.

27. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. (2016). Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 167, 1853–1866.e17. [PubMed: 27984732]

28. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, Pliner HA, Jackson DL, Daza RM, Christiansen L, et al. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. Science 367, 45–51. [PubMed: 31806696]

29. Feldman D, Funk L, Le A, Carlson RJ, Leiken MD, Tsai F, Soong B, Singh A, and Blainey PC (2022). Pooled genetic perturbation screens with image-based phenotypes. Nat. Protoc 17, 476–512. [PubMed: 35022620]

30. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, and Polosukhin I (2017). Attention is all you need. Adv. Neural Inf. Process. Syst 30.

31. Rombach R, Blattmann A, Lorenz D, Esser P, and Ommer B High-resolution image synthesis with latent diffusion models. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 10684–10695.

32. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. [PubMed: 34265844]

33. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876. [PubMed: 34282049]

34. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130. [PubMed: 36927031]

35. Gomes J, Ramsundar B, Feinberg EN, and Pande VS (2017). Atomic convolutional networks for predicting protein-ligand binding affinity. Preprint at arXiv.

36. Cunningham JM, Koytiger G, Sorger PK, and AlQuraishi M (2020). Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. Nat. Methods 17, 175–183. [PubMed: 31907444]

37. Torng W, and Altman RB (2019). High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics 35, 1503–1512. [PubMed: 31051039]

38. Corso G, Stärk H, Jing B, Barzilay R, and Jaakkola T (2023). DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking The Eleventh International Conference on Learning Representations.

39. Kudo T, Meireles AM, Moncada R, Chen Y, Wu P, Gould J, Hu X, Kornfeld O, Jesudason R, Foo C, et al. (2024). Multiplexed, image-based pooled screens in primary cells and tissues with perturbview. Nat. Biotechnol, 1–10. [PubMed: 38191665]

40. Roohani Y, Huang K, and Leskovec J (2023). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat. Biotechnol 42, 927–935. [PubMed: 37592036]

41. Bunne C, Stark SG, Gut G, Del Castillo JS, Levesque M, Lehmann K-V, Pelkmans L, Krause A, and Ratsch G (2023). Learning single-cell perturbation responses using neural optimal transport. Nat. Methods 20, 1759–1768. [PubMed: 37770709]

42. Lotfollahi M, Klimovskaia Susmelj A, De Donno C, Hetzel L, Ji Y, Ibarra IL, Srivatsan SR, Naghipourfar M, Daza RM, Martin B, et al. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. Mol. Syst. Biol 19, e11517. [PubMed: 37154091]

43. Bunne C, Krause A, and Cuturi M (2022). Supervised Training of Conditional Monge Maps. Adv. Neural Inf. Process. Syst 35, 6859–6872.

44. Bereket M, and Karaletsos T (2024). Modelling Cellular Perturbations with the Sparse Additive Mechanism Shift Variational Autoencoder. Adv. Neural Inf. Process. Syst 36.

45. Huang K, Huang K, Lopez R, Hutter J-C, Kudo T, Rios A, and Regev A (2023). Sequential Optimal Experimental Design of Perturbation Screens Guided by Multi-modal Priors. In International Conference on Research in Computational Molecular Biology (Springer Nature), pp. 17–37.

46. Roohani YH, Vora J, Huang Q, Liang P, and Leskovec J (2024). BioDiscoveryAgent: An AI Agent for Designing Genetic Perturbation Experiments Preprint at arXiv.

47. Cleary B, and Regev A (2024). The necessity and power of random, undersampled experiments in biology. Preprint at arXiv.

48. Ji Y, Zhou Z, Liu H, and Davuluri RV (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 37, 2112–2120. [PubMed: 33538820]

49. Brandes N, Ofer D, Peleg Y, Rappoport N, and Linial M (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38, 2102–2110. [PubMed: 35020807]

50. Celaj A, Gao AJ, Lau TTY, Holgersen EM, Lo A, Lodaya V, Cole CB, Denroche RE, Spickett C, Wagih O, et al. (2023). An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. Preprint at bioRxiv.

51. Nguyen E, Poli M, Durrant MG, Kang B, Katrekar D, Li DB, Bartie LJ, Thomas AW, King SH, Brixi G, et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. Science 386, eado9336. [PubMed: 39541441]

52. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, et al. (2024). Simulating 500 million years of evolution with a language model. Preprint at bioRxiv.

53. Ruffolo JA, Nayfach S, Gallagher J, Bhatnagar A, Beazer J, Hussain R, Russ J, Yip J, Hill E, Pacesa M, et al. (2024). Design of highly functional genome editors by modeling the universe of CRISPR-cas sequences. Preprint at bioRxiv.

54. Peng Z, Schussheim B, and Chatterjee P (2024). PTM-mamba: a PTM-aware protein language model with bidirectional gated mamba blocks. Preprint at bioRxiv.

55. Dai B, Mattox DE, and Bailey-Kellogg C (2021). Attention please: modeling global and local context in glycan structure-function relationships. Preprint at bioRxiv.

56. Yu T, Yao T, Sun Z, Shi F, Zhang L, Lyu K, Xuan B, Liu A, Zhang X, Zou J, et al. (2024). LipidBERT: A Lipid Language Model Pretrained on METiS de novo Lipid Library. Preprint at arXiv.

57. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR, et al. (2024). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science 384, eadl2528. [PubMed: 38452047]

58. Rosen Y, Brbi M, Roohani Y, Swanson K, Li Z, and Leskovec J (2024). Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with Saturn. Nat. Methods 21, 1492–1500. [PubMed: 38366243]

59. Rosen Y, Roohani Y, Agrawal A, Samotor an L, Tabula Sapiens Consortium, Quake SR, and Leskovec J (2023). Universal cell embeddings: A foundation model for cell biology. Preprint at bioRxiv.

60. Chen Y, and Zou J (2024). GenePT: A Simple but Effective Foundation Model for Genes and Cells Built from ChatGPT. Preprint at bioRxiv.

61. Mahdessian D, Cesnik AJ, Gnann C, Danielsson F, Stenström L, Arif M, Zhang C, Le T, Johansson F, Schutten R, et al. (2021). Spatiotemporal dissection of the cell cycle with single-cell proteogenomics. Nature 590, 649–654. [PubMed: 33627808]

62. Chandrasekaran SN, Cimini BA, Goodale A, Miller L, Kost-Alimova M, Jamali N, Doench JG, Fritchman B, Skepner A, Melanson M, et al. (2024). Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nat. Methods 21, 1114–1121. [PubMed: 38594452]

63. Carlson RJ, Leiken MD, Guna A, Hacohen N, and Blainey PC (2023). A genome-wide optical pooled screen reveals regulators of cellular antiviral responses. Proc. Natl. Acad. Sci. USA 120, e2210623120. [PubMed: 37043539]

64. Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, et al. (2017). A subcellular map of the human proteome. Science 356, eaal3321. [PubMed: 28495876]

65. McDole K, Guignard L, Amat F, Berger A, Malandain G, Royer LA, Turaga SC, Branson K, and Keller PJ (2018). In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. Cell 175, 859–876.e33. [PubMed: 30318151]

66. Nogales E, and Mahamid J (2024). Bridging structural and cell biology with cryo-electron microscopy. Nature 628, 47–56. [PubMed: 38570716]

67. Bauda E, Gallet B, Moravcova J, Effantin G, Chan H, Novacek J, Jouneau PH, Rodrigues CDA, Schoehn G, Moriscot C, et al. (2024). Ultrastructure of macromolecular assemblies contributing to

bacterial spore resistance revealed by in situ cryo-electron tomography. Nat. Commun 15, 1376. [PubMed: 38355696]

68. Lelek M, Gyparaki MT, Beliu G, Schueder F, Griffié J, Manley S, Jungmann R, Sauer M, Lakadamyali M, and Zimmer C (2021). Single-molecule localization microscopy. Nat. Rev. Methods Primers 1, 39. [PubMed: 35663461]

69. Möckl L, and Moerner WE (2020). Super-resolution microscopy with single molecules in biology and beyond–essentials, current trends, and future challenges. J. Am. Chem. Soc 142, 17828–17844. [PubMed: 33034452]

70. Cesnik A, Schaffer LV, Gaur I, Jain M, Ideker T, and Lundberg E (2024). Mapping the multiscale proteomic Organization of cellular and Disease Phenotypes. Annu. Rev. Biomed. Data Sci 7, 369–389. [PubMed: 38748859]

71. Qin Y, Huttlin EL, Winsnes CF, Gosztyla ML, Wacheul L, Kelly MR, Blue SM, Zheng F, Chen M, Schaffer LV, et al. (2021). A multi-scale map of cell structure fusing protein images and interactions. Nature 600, 536–542. [PubMed: 34819669]

72. Dosovitskiy A (2020). An image is worth 16×16 words: transformers for image recognition at scale. Preprint at arXiv.

73. Fukushima K (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern 36, 193–202. [PubMed: 7370364]

74. LeCun Y, and Yoshua B (1995). Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks 3361, 255–258.

75. Bao Y, Sivanandan S, and Karaletsos T (2023). Channel Vision Transformers: An Image Is Worth c × 16 × 16 WordsThe Twelfth. International Conference on Learning Representations 4.

76. Kraus O, Kenyon-Dean K, Saberian S, Fallah M, McLean P, Leung J, Sharma V, Khan A, Balakrishnan J, Celik S, et al. (2024). Masked autoencoders for microscopy are scalable learners of cellular biology. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11757–11768.

77. Bao Y, and Karaletsos T (2023). Contextual vision transformers for robust representation learning. Preprint at arXiv.

78. Lopez R, Regier J, Cole MB, Jordan MI, and Yosef N (2018). Deep generative modeling for single-cell transcriptomics. Nature Methods 15, 1053–1058. [PubMed: 30504886]

79. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, et al. (2023). Transfer learning enables predictions in network biology. Nature 618, 616–624. [PubMed: 37258680]

80. Kobayashi-Kirschvink KJ, Comiter CS, Gaddam S, Joren T, Grody EI, Ounadjela JR, Zhang K, Ge B, Kang JW, Xavier RJ, et al. (2024). Prediction of single-cell RNA expression profiles in live cells by Raman microscopy with Raman2RNA. Nat. Biotechnol 42, 1726–1734. [PubMed: 38200118]

81. Ryu J, Lopez R, Bunne C, and Regev A (2024). Cross-modality matching and prediction of perturbation responses with labeled Gromov-Wasserstein optimal transport. Preprint at arXiv.

82. Saar KL, Scrutton RM, Bloznelyte K, Morgunov AS, Good LL, Lee AA, Teichmann SA, and Knowles TPJ (2024). Protein Condensate Atlas from predictive models of heteromolecular condensate composition. Nat. Commun 15, 5418. [PubMed: 38987300]

83. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas A, Kamitaki N, Martersteck E, et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214. [PubMed: 26000488]

84. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, Giacomello S, Asp M, Westholm JO, Huss M, et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353, 78–82. [PubMed: 27365449]

85. Lundberg E, and Borner GHH (2019). Spatial proteomics: a powerful discovery tool for cell biology. Nat. Rev. Mol. Cell Biol 20, 285–302. [PubMed: 30659282]

86. Marconato L, Palla G, Yamauchi KA, Virshup I, Heidari E, Treis T, Vierdag WM, Toth M, Stockhaus S, Shrestha RB, et al. (2024). SpatialData: an open and universal data framework for spatial omics. Nat. Methods 10.1038/s41592-024-02212-x.

87. Somnath VR, Pariset M, Hsieh Y-P, Martinez MR, Krause A, and Bunne C (2023). Aligned Diffusion Schrödinger Bridges. In Uncertainty in Artificial Intelligence, pp. 1985–1995.

88. Katharopoulos A, Vyas A, Pappas N, and Fleuret F (2020). Fast autoregressive transformers with linear attention. In International Conference on Machine Learning.

89. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500. [PubMed: 38718835]

90. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, Gilbert LA, and Weissman JS (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science 365, 786–793. [PubMed: 31395745]

91. Lawson MJ, Camsund D, Larsson J, Baltekin Ö, Fange D, and Elf J (2017). In situ genotyping of a pooled strain library after characterizing complex phenotypes. Mol. Syst. Biol 13, 947. [PubMed: 29042431]

92. Papamarkou T, Skoularidou M, Palla K, Aitchison L, Arbel J, Dunson D, Filliponne M, Fortuin V, Hennig P, Hernandez-Lobato JM, et al. (2024). Position: bayesian deep learning is needed in the age of large-scale AI. In Forty-First International Conference on Machine Learning.

93. D'Angelo F, and Fortuin V (2021). Wenzel F.On Stein Variational Neural Network. Ensembles Preprint at arXiv.

94. Ober SW, Rasmussen CE, and van der Wilk M (2021). The promises and pitfalls of deep kernel learning. In Conference on Uncertainty in Artificial Intelligence, pp. 1206–1216.

95. Karaletsos T (2020). Bui T.D.Hierarchical Gaussian Process Priors for Bayesian Neural Network Weights. Adv. Neural Inf. Process. Syst 33, 17141–17152.

96. Kapoor S, Maddox WJ, Izmailov P, and Wilson AG (2022). On uncertainty, tempering, and data augmentation in bayesian classification. Adv. Neural Inf. Process. Syst 35, 18211–18225.

97. Lakshminarayanan B, Pritzel A, and Blundell C (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Adv. Neural Inf. Process. Syst 30.

98. Angelopoulos AN, and Bates S (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Preprint at arXiv.

99. Cherian JJ, Gibbs I, and Candès EJ (2024). Large language model validity via enhanced conformal prediction methods. Preprint at arXiv.

100. Cho NH, Cheveralls KC, Brunner AD, Kim K, Michaelis AC, Raghavan P, Kobayashi H, Savy L, Li JY, Canaj H, et al. (2022). Open-Cell: endogenous tagging for the cartography of human cellular organization. Science 375, eabi6983. [PubMed: 35271311]

101. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. (2015). Proteomics. Tissue-based map of the human proteome. Science 347, 1260419. [PubMed: 25613900]

102. Berger C, Premaraj N, Ravelli RBG, Knoops K, López-Iglesias C, and Peters PJ (2023). Cryo-electron tomography on focused ion beam lamellae transforms structural cell biology. Nat. Methods 20, 499–511. [PubMed: 36914814]

103. Loconte V, Chen JH, Vanslembrouck B, Ekman AA, McDermott G, Le Gros MA, and Larabell CA (2023). Soft X-ray tomograms provide a structural basis for whole-cell modeling. FASEB J. 37, e22681. [PubMed: 36519968]

104. Moffitt JR, Lundberg E, and Heyn H (2022). The emerging landscape of spatial profiling technologies. Nat. Rev. Genet 23, 741–759. [PubMed: 35859028]

105. Vandereyken K, Sifrim A, Thienpont B, and Voet T (2023). Methods and applications for single-cell and spatial multi-omics. Nat. Rev. Genet 24, 494–515. [PubMed: 36864178]

106. Tabula Sapiens Consortium*, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, Yosef N, Bulthaup B, Brown P, et al. (2022). The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science 376, eabl4896. [PubMed: 35549404]

107. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, Maaskola J, Lundeberg J, and Zou J (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. Nat. Biomed. Eng 4, 827–834. [PubMed: 32572199]

108. Bock C, Boutros M, Camp JG, Clarke L, Clevers H, Knoblich JA, Liberali P, Regev A, Rios AC, Stegle O, et al. (2021). The organoid cell atlas. Nat. Biotechnol 39, 13–17. [PubMed: 33384458]

109. Tabula; Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators (2018). Single-cell transcriptomics of 20 mouse organs creates a tabula muris. Nature 562, 367–372. [PubMed: 30283141]

110. Li H, Janssens J, De Waegeneer M, Kolluru SS, Davie K, Gardeux V, Saelens W, David FPA, Brbi M, Spanier K, et al. (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. Science 375, eabk2432. [PubMed: 35239393]

111. Lange M, Granados A, Vijaykumar S, Bragantini J, Ancheta S, Santhosh S, Borja M, Kobayashi H, McGeever E, Solak AC, et al. (2023). Zebrahub – Multimodal zebrafish Developmental Atlas Reveals the State Transition Dynamics of Late Vertebrate Pluripotent Axial Progenitors. Preprint at bioRxiv.

112. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, and O'Sullivan C (2022). The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 50, D387–D390. [PubMed: 34850094]

113. Achiam J, et al. (2023). GPT-4 technical report. Preprint at arXiv.

114. Ding F, and Steinhardt JN (2024). Protein language models are biased by unequal sequence sampling across the tree of life. Preprint at bioRxiv.

115. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. (2023). A draft human pangenome reference. Nature 617, 312–324. [PubMed: 37165242]

116. Liu J, Shen Z, He Y, Zhang X, Xu R, Yu H, and Cui P (2021). Towards out-of-distribution generalization: A survey. Preprint at arXiv.

117. Nisonoff H, Wang Y, and Listgarten J (2023). Coherent blending of biophysics-based knowledge with bayesian neural networks for robust protein property prediction. ACS Synth. Biol 12, 3242–3251. 10.1021/acssynbio.3c00217. [PubMed: 37888887]

118. Zheng F, Kelly MR, Ramms DJ, Heintschel ML, Tao K, Tutuncuoglu B, Lee JJ, Ono K, Foussard H, Chen M, et al. (2021). Interpretation of cancer mutations using a multiscale map of protein systems. Science 374, eabf3067. [PubMed: 34591613]

119. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, and Ideker T (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods 15, 290–298. [PubMed: 29505029]

120. Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR, Ektefaie Y, Kondic J, and Zitnik M (2024). Empowering biomedical discovery with AI agents. Cell 187, 6125–6151. [PubMed: 39486399]

121. Hurrell T, Naidoo J, Ntlhafu T, and Scholefield J (2024). An African perspective on genetically diverse human induced pluripotent stem cell lines. Nat. Commun 15, 8581. [PubMed: 39362853]

122. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, Ch'ang LY, Huang W, Shen B, Tam Y, et al. (2003). The international HapMap project. Nature 5, 467–475.

123. Cancer; Genome; Atlas; Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, and Stuart JM (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet 45, 1113–1120. [PubMed: 24071849]

124. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. [PubMed: 22955616]

125. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, and Foster B (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet 45, 580–585. [PubMed: 23715323]

126. Pontén F, Jirström K, and Uhlen M (2008). The Human Protein Atlas–a tool for pathology. J. Pathol 216, 387–393. [PubMed: 18853439]

127. Downey P, and Peakman TC (2008). Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles. Int. J. Epidemiol 37 (Suppl 1), i46–i50. [PubMed: 18381393]

128. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL, Xiong C, Sun ZZ, Socher R, et al. (2023). Large language models generate functional protein sequences across diverse families. Nat. Biotechnol 41, 1099–1106. [PubMed: 36702895]

129. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet 47, 856–860. [PubMed: 26121088]

130. Mason C, Brindley DA, Culme-Seymour EJ, and Davie NL (2011). Cell therapy industry: billion dollar global business with unlimited potential. Regen. Med 6, 265–272. [PubMed: 21548728]

131. Bashor CJ, Hilton IB, Bandukwala H, Smith DM, and Veiseh O (2022). Engineering the next generation of cell-based therapeutics. Nat. Rev. Drug Discov 21, 655–675. [PubMed: 35637318]

132. Jia Q, Wang A, Yuan Y, Zhu B, and Long H (2022). Heterogeneity of the tumor immune microenvironment and its clinical relevance. Exp. Hematol. Oncol 11, 24. [PubMed: 35461288]

133. Melssen MM, Sheybani ND, Leick KM, and Slingluff CL (2023). Barriers to immune cell infiltration in tumors. J. Immunother. Cancer 11.

134. Chow A, Perica K, Klebanoff CA, and Wolchok JD (2022). Clinical implications of T cell exhaustion for cancer immunotherapy. Nature Reviews Clinical Oncology 19, 775–790.

135. de Visser KE, and Joyce JA (2023). The evolving tumor microenvironment: from cancer initiation to metastatic outgrowth. Cancer Cell 41, 374–403. [PubMed: 36917948]

136. Barkley D, Moncada R, Pour M, Liberman DA, Dryg I, Werba G, Wang W, Baron M, Rao A, Xia B, et al. (2022). Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. Nat. Genet 54, 1192–1201. [PubMed: 35931863]

137. Schwartzberg L, Kim ES, Liu D, and Schrag D (2017). Precision oncology: who, how, what, when, and when not? American Society of Clinical Oncology Educational Book 37, 160–169. [PubMed: 28561651]

138. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, Costello CE, Cravatt BF, Fenselau C, Garcia BA, et al. (2018). How many human proteoforms are there? Nat. Chem. Biol 14, 206–214. [PubMed: 29443976]

139. Katsoulakis E, Wang Q, Wu H, Shahriyari L, Fletcher R, Liu J, Achenie L, Liu H, Jackson P, Xiao Y, et al. (2024). Digital twins for health: a scoping review. npj Digit. Med 7, 77. [PubMed: 38519626]

140. Rajewsky N, Almouzni G, Gorski SA, Aerts S, Amit I, Bertero MG, Bock C, Bredenoord AL, Cavalli G, Chiocca S, et al. (2020). Lifetime and improving European healthcare through cell-based interceptive medicine. Nature 587, 377–386. [PubMed: 32894860]

141. Alix-Panabières C, and Pantel K (2021). Liquid biopsy: from discovery to clinical application. Cancer Discov. 11, 858–873. [PubMed: 33811121]

142. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, and Regev A (2022). The evolution, evolvability and engineering of gene regulatory DNA. Nature 603, 455–463. [PubMed: 35264797]

143. Gómez-de-Mariscal E, García-López-de-Haro C, Ouyang W, Donati L, Lundberg E, Unser M, Muñoz-Barrutia A, and Sage D (2021). DeepImageJ: A user-friendly environment to run deep learning models in ImageJ. Nat. Methods 18, 1192–1195. [PubMed: 34594030]

144. Le T, Winsnes CF, Axelsson U, Xu H, Mohanakrishnan Kaimal J, Mahdessian D, Dai S, Makarov IS, Ostankovich V, Xu Y, et al. (2022). Analysis of the human protein atlas weakly supervised single-cell classification competition. Nat. Methods 19, 1221–1229. [PubMed: 36175767]

145. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, Chen B, Zhang A, Shao D, Shaban M, et al. (2024). Towards a general-purpose foundation model for computational pathology. Nat. Med 30, 850–862. [PubMed: 38504018]

146. Moen E, Bannon D, Kudo T, Graf W, Covert M, and Van Valen D (2019). Deep learning for cellular image analysis. Nat. Methods 16, 1233–1246. [PubMed: 31133758]

147. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet 53, 354–366. [PubMed: 33603233]

148. Ho J, Jain A, and Abbeel P (2020). Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst 33, 6840–6851.

149. Lipman Y, Chen RT, Ben-Hamu H, Nickel M, and Le M (2023). Flow Matching for Generative Modeling. International Conference on Learning Representations.

150. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, and Monfardini G (2009). The graph neural network model. IEEE Trans. Neural Netw 20, 61–80. [PubMed: 19068426]

151. Cao Y, and Shen Y (2020). Energy-based graph convolutional networks for scoring protein docking models. Proteins 88, 1091–1099. [PubMed: 32144844]

152. Brbi M, Cao K, Hickey JW, Tan Y, Snyder MP, Nolan GP, and Leskovec J (2022). Annotation of spatially resolved single-cell data with STELLAR. Nat. Methods 19, 1411–1418. [PubMed: 36280720]

153. Wu Z, Trevino AE, Wu E, Swanson K, Kim HJ, D'Angio HB, Preska R, Charville GW, Dalerba PD, Egloff AM, et al. (2022). Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. Nat. Biomed. Eng 6, 1435–1448. [PubMed: 36357512]

154. Hamilton W, Ying Z, and Leskovec J (2017). Inductive Representation Learning on Large Graphs. Adv. Neural Inf. Process. Syst 30.

Box 1.

Grand challenges for building the AIVC

OUTLINING CAPABILITIES AND DESIGNING EVALUATION FRAMEWORKS

The burgeoning number of foundation models in biology perform a subset of the capabilities of virtual cells outlined in this perspective. Given the diversity of these approaches, it is important to define what the core capabilities of AIVCs should be and how those capabilities can be evaluated. For every capability, proper metrics must be designed, and comprehensive evaluation data be collected. Models' capabilities should be assessed on general performance as well as on their ability to answer specific biological questions. It is imperative to continuously improve benchmarking strategies along with AIVC models and ensure that they align with biologically meaningful objectives. As the field develops better alignment on these questions, collaborative opportunities will arise, and the speed at which virtual cells can be generated will accelerate.

ESTABLISHING SELF-CONSISTENCY ACROSS VARYING CONTEXTS WITH DIFFERENT ARCHITECTURES

Biology is tremendously complex: it operates across different scales, in different contexts, and is measured with different modalities. AIVC models must be self-consistent across all of these axes. Models should propagate function across physical scales— interactions between molecules should have consistent effects when measuring binding affinity, gene expression, cell-cell communication, or tissue organization. As physical and dynamic scales increase in scope and size, additional context, for example, species, cell type, tissue, disease status, etc., should fine-tune predictions made at smaller resolutions, while also accounting for stochasticity. Model predictions should also be agnostic to their input and output modalities. The same entity, profiled with different technologies, should have the same internal representation in an AIVC. To properly model such complex behaviors, many machine learning approaches should be explored and their merits carefully judged.

BALANCING INTERPRETABILITY AND BIOLOGICAL UTILITY

A consistent trend in the application of deep learning methods to biology, accelerated by the rise of large foundation models, has been the implicit trade-off between models' performance gains and their increasingly uninterpretable "black box" natures. AIVC models will ultimately be judged on their ability to expand our understanding of biology, either by providing novel insights to biological processes or by accelerating the scientific process. To achieve this goal, AIVC models must make highly accurate and well-calibrated predictions that simulate biology, and the trade-off between actionability and interpretability will have to be balanced. Actionable model outputs are those of high utility to design affordable and efficient validation experiments and are key for initial real-world use. Various approaches exist for explaining model predictions, including causal modeling, sparse featurization, and counterfactual reasoning, and this is a highly active research area. Building intuitive interfaces that facilitate the study

and interpretation of AIVCs via other models, such as AI research agents, will further increase downstream utility.

## CONSTRUCTING A FRAMEWORK FOR COLLABORATIVE CELL MODELING

The successful development of AIVCs will require collaboration across disciplines. We foresee a future where AIVC platforms function as open, interconnected hubs for collaborative development and broad deployment of cell models to researchers and as education hubs delivering training to researchers, as well as providing engagement activities for educators, patients, and the public. Thus, investments in infrastructure fostering open and collaborative development of AIVCs should be of high priority.

ENSURING AIVCs BENEFIT ALL AND PROMOTE ETHICAL AND RESPONSIBLE USE

Generating large open datasets that reflect human diversity—datasets integral for training AIVC models—poses a substantial challenge. Developers will have to use the utmost care to ensure these datasets are used ethically and transparently while building AIVCs and develop strategies to mitigate risks of model contamination with falsified data. Early adopters of AIVCs will have a key role in promoting and demonstrating responsible use of these models. Furthermore, the development of chat-based interfaces could be crucial in democratizing access to AIVCs. Close collaboration with ethics and regulatory experts from the outset is paramount for establishing new regulatory norms that will facilitate the responsible use of AIVCs.

UNDERSTANDING THE VALUE OF DIFFERENT DATA TYPES TO PRIORITIZE LARGE-SCALE DATA GENERATION
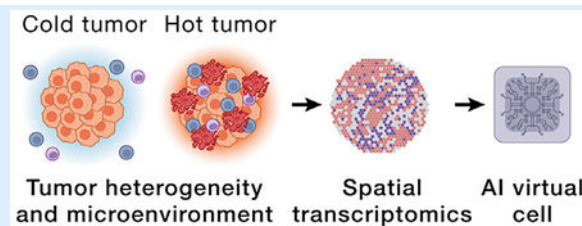
A fundamental question for the collaborative development of AIVCs is which data and modalities should be collected to enable generalization across biological contexts and scales. These data will need to encompass the breadth of biology in different species, domains, and modalities, representing the heterogeneity of life, while maintaining depth sufficient to distinguish true signals from noise. A key aspect of data generation will be the simultaneous measurement of temporal and physical scales, while also allowing perturbations of the system.

Box 2.

Vignettes

CELL ENGINEERING TO ENABLE PHENOTYPIC DRUG DISCOVERY AND CELL-BASED THERAPEUTICS
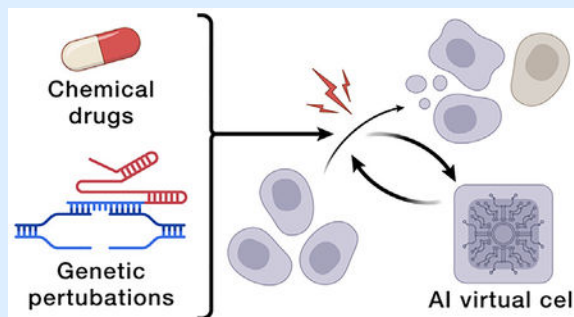
One challenge in developing successful therapies is the difficulty in incorporating the full underlying genetic, molecular and cellular basis of disease during drug discovery and development.[129] These context-specific underpinnings are not fully specified and often vary between human patients and model systems used in pre-clinical studies. By integrating biological data from various sources relevant to specific disease contexts, the AIVC could generate an environment for testing different therapeutic interventions *in silico* and identify approaches for engineering cells to reverse disease phenotypes, while accounting for the effects of varying both treatments and patient profiles. By representing the overall disease phenotype specific to patient populations (rather than one specific biochemical target at a time), the AIVC can enable virtual phenotypic screens. Although *in silico* experiments may not always be fully accurate, by prioritizing virtual hits with higher chances of success, the AIVC can lower experimentation costs and accelerate the process.

The AIVC has potential to push the cell therapy frontier. With growing evidence affirming the efficacy and safety of cell-based therapies for rare diseases and cancer,[130,131] the AIVC can improve systematization and precision to cell engineering. For example, virtual cell-based engineering could enable targeted modifications to pancreatic beta cells to create individualized beta cell replacement therapies for type 1 diabetes. By simulating the biological phenotype of individual patients, *in silico* experiments within the AIVC could identify interventions that help drive the differentiation of beta cells from progenitors, cloak them from the immune system, and maintain their function, with the ultimate goal of either transplanting these engineered cells into patients or engineering them *in situ*.

UNLOCKING THE POWER OF SPATIAL BIOLOGY TO FIGHT CANCER

Cold tumor Hot tumor

Tumor heterogeneity and microenvironment → Spatial transcriptomics → AI virtual cell

Spatial structures in cancer, specifically within the tumor microenvironment (TME), are critical drivers of cancer progression and can drive resistance to the immune system and limit drug efficacy.[132] Malignant cells within a tumor can engage in active immune evasion by either blocking immune infiltration,[133] evading immune recognition, or dampening immune cell function.[134] Thus, immune resistance must be understood in the spatial context of the cellular neighborhood to identify the specific cell states and gene signatures involved. Although next-generation spatial profiling methods enable researchers to experimentally investigate the heterogeneity of the TME,[135] an AIVC could extend these analyses to a universal, pan-cancer framework, which can be personalized to individual patients. Using an AIVC model, cancer researchers should be able to identify TME niches shared across multiple cancer types from many patients. Identifying pan-cancer markers can drive cancer treatment both by highlighting new targets and also by identifying existing treatments that can be applied to new cancer types.[136] In this setting, the AIVC would help identify the interactions associated with TME cell states and would search for similar states from any disease where existing treatments exist.

Finally, the AIVC could greatly advance precision oncology.[137] Given that the AIVC will capture intrinsic variation, the genetic diversity of individual patient's cancers will be represented in any analyses. Although the AIVC would already accurately qualify the change in the expression of genes, tumor sequencing data would allow it to model the change of function of those genes, for example, through loss of function, change in post-translational modifications, or rewiring of protein-protein interactions and signaling networks.[138]

DIAGNOSTIC VIRTUAL CELL MODELS FOR INDIVIDUAL PATIENTS

Tissue sample

Plasma
RBCs/WBCs → AI virtual cell → Digital twin
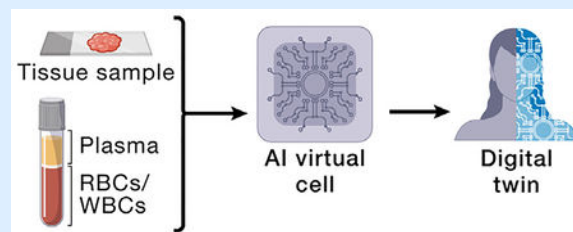
The AIVC could introduce a new approach to diagnostics that incorporates a personalized AIVC (or a digital twin[139]) to track a patient's health and suggest suitable interventions. The AIVC would create a detailed representation of each patient's cells by incorporating specific patient data, such as genetic sequences, single-cell profiles from blood, and tissue pathology images, along with additional clinical information from their health records. Periodic updates to each patient's AIVC instance enable monitoring

of evolving health conditions, prediction of upcoming adverse events, and potential therapeutic outcomes.

Through additional updates from less costly assays, this virtual patient model could be progressively refined and made more robust.[140] For example, transcriptomic or genetic liquid biopsies can reveal significant and diverse characteristics of a patient from a single test and could greatly aid in the diagnosis of a broad spectrum of conditions.[141] Through the virtual cell's implicit and structured representation of universal cell types and states, one can envision the creation of patient models of inaccessible cell types, such as beta cells in the pancreas or neurons in the brain, generated after sampling accessible cell types such as blood or skin.

A HYPOTHESIS-GENERATING FRAMEWORK FOR SCIENTIFIC RESEARCH
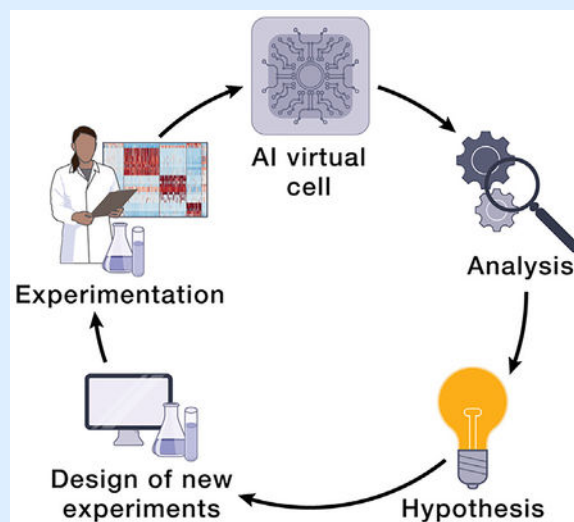

Traditionally, the biological research community has relied on computational models for analyzing data from past experiments based on an existing hypothesis. The virtual cell could switch the paradigm by computationally exploring a vast array of possible hypotheses through *in silico* experimentation. It could identify the most informative experiments for addressing specific biological questions, shifting the role of computational models from merely validating hypotheses or processing observations without a particular goal to generating specific sets of hypotheses to pursue.

This shift could greatly enhance the scientific discovery process: instead of conducting a single experiment followed by an in-depth analysis, scientists can engage in a dynamic iterative interaction with the virtual cell. With each new piece of data, they can refine their understanding of the biological system and consult the virtual cell to identify what additional experimental data could be valuable. Ultimately, we may be able to perform active learning with biologists in the loop and construct self-driving labs for efficient and unbiased generation of virtual cells.

Box 3.

AI techniques for building the AIVC

The AIVC will connect a number of diverse neural network architectures. Although these architectures may not have been purpose-built for biological applications, they have each demonstrated success when matched with specific biological modalities and inductive biases. In many cases, these architectures may be exchangeable, and one must weigh their individual trade-offs in accuracy, speed, and generalizability. Beyond this, the community is actively developing AI architectures tailored to the characteristics of (large) biological datasets.

TRANSFORMERS
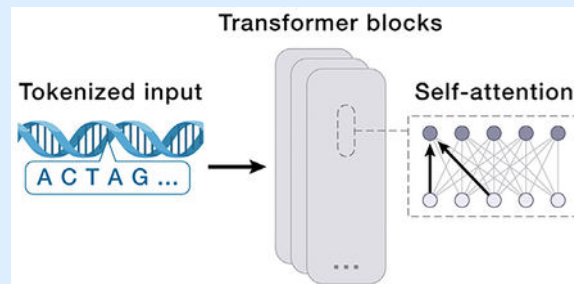
A transformer neural network[30] comprises multiple transformer layers, each taking a series of tokens (discrete pieces of information such as words, RNA molecules, or gene representations) as input—initial tokens for the first layer and outputs from the preceding layer for subsequent ones. Within each layer, tokens use self-attention to integrate context from other tokens, enhancing their own representations, which are then processed through a feed forward network. This architecture, which fundamentally requires only a collection of tokens, adapts well across various applications and use cases.

The collection of tokens passed to a transformer does not have any ordering by default. Additionally, the self-attention mechanism, the core of the success of the transformer, can be taken as a strong biological inductive bias. For instance, in representing cells through their RNA molecules detected via scRNA-seq, each RNA molecule, represented as a token, interacts with others, modeling gene interactions through self-attention.[30] Customizing input tokens with numerical representations of genes further allows the integration of diverse biological data scales, from individual genes to whole cells.[59,60]

Additionally, introducing positional encodings to tokens enables transformers to process sequences, such as natural language,[30] or biological sequences, such as DNA,[48,142] by incorporating sequence-specific dependencies. This approach is crucial in applications such as masked language modeling, where the model predicts missing tokens in sequences, enhancing its understanding of contextual relationships within data. Innovations continue to refine transformers, increasing their capacity to handle longer sequences and improving efficiency, with advancements such as state-space models enabling the generation of extensive DNA sequences.[51]

CNN
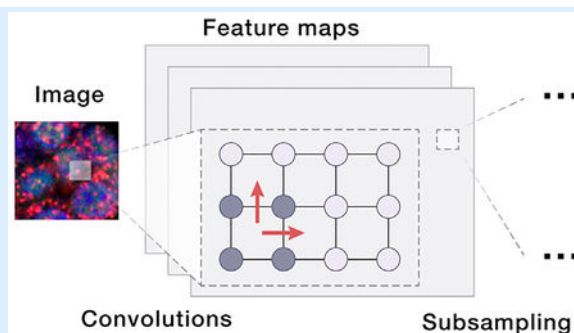
A CNN is a deep learning model primarily used for analyzing images.[73,74] It consists of multiple layers that automatically and adaptively learn spatial hierarchies of features through backpropagation. This learning is facilitated by convolutional layers that apply filters to local patches of input data, pooling layers that reduce dimensionality, and fully connected layers that interpret the features extracted to make decisions.

In the field of biology, CNNs have proven invaluable for tasks involving image data due to their ability to detect complex patterns and structures, such as microscope images of cells and tissues. Here, CNNs play a critical role in multiplex imaging,[143] where multiple targets within a single sample are labeled and visualized simultaneously. This technique is particularly useful in studying the complex interactions of different molecules or cell types within a heterogeneous tissue environment.[144] Another notable application is in the analysis of H&E-stained tissue sections, commonly used in clinical pathology.[145] Lastly, in live-cell imaging, CNNs are employed to track dynamic changes within cells or even single molecules over time, providing insights into cell migration, responses to treatment, or the movement and interaction of individual molecules within cells, revealing crucial biological processes at a molecular level.[146]

Beyond their traditional use in image processing, CNNs can also be applied to model sequence data, such as DNA sequences, where they identify patterns and features that are predictive of biological functions.[147] Despite their extensive utility, CNNs are increasingly being supplemented or replaced by vision transformer models,[72] which leverage self-attention mechanisms to process entire images in parallel. These models can often achieve higher accuracy on tasks where understanding the global context within the image is crucial.

DIFFUSION MODELS
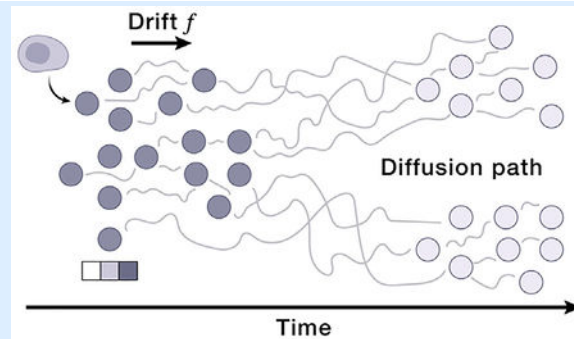
Diffusion models are a class of generative deep learning models that have recently gained attention for their ability to generate high-quality, diverse samples across various domains.[148] They operate by gradually transforming a distribution of random noise into a structured output (images, text, cellular states, etc.) through a process that mimics a physical diffusion process. Building up on diffusion model architectures, approaches such as flow matching methods can also model the distributional evolution over time,[149] making them especially powerful in biological applications where dynamic changes and temporal progression are critical. Flow matching methods thus capture and generate sequences of data that reflect continuous transformations, such as the developmental stages of cells over time and space or the response of biological systems to treatments.[87] The ability of diffusion and flow matching models to learn and replicate complex distributions, combined with the temporal and spatial modeling capabilities of flow matching methods, makes them particularly suited for tasks that involve high-dimensional, intricate data structures typical of biological systems.

GNNs
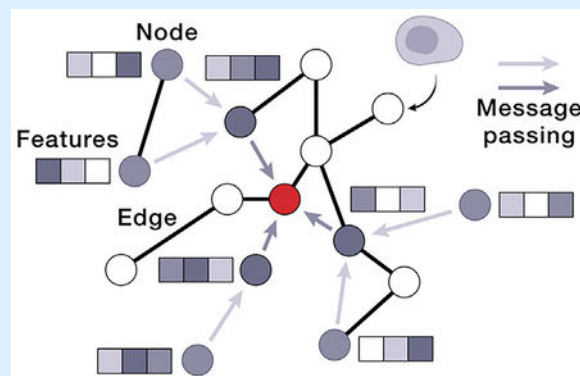
GNNs are a set of architectures that can model graphical data.[150] Graphs, sets of nodes connected by edges, are useful representations for many kinds of biological data. When modeling a biological system, a GNN could be a good choice if a graph structure represents some core inductive bias. For example, a protein structure[151] can be thought of as a graph where residues are nodes, and their bonds are edges. Cells in a tissue form a graph: each cell is a node, and the cells it is physically proximal to are connected by edges.[152,153] In both cases, the graph represents how nodes are physically proximal

to each other. For spatially organized cells, the graph represents how they may pass chemical signals between one another.

GNNs can be used to make predictions about individual nodes, edges, or the graph as a whole.[154] For simplicity, in the following section, we describe a node-based GNN. At each layer, a node updates its representation using a neural network, which can take in that node's current representation, in addition to the representations of the node's neighbors, which are connected by an edge. By stacking GNN layers, a node can receive "messages" from neighboring nodes at increasing distances, "hops," from it. Nodes and edges can both be initialized with different features, which control their final representation and what messages they pass to their neighbors. For example, a GNN trained on spatial transcriptomic data could take node features to be the virtual cell representation of each cell's gene expression. The GNN would then update those representations to include context about each cell's neighbors, helping to identify spatial interactions and niches.[153]
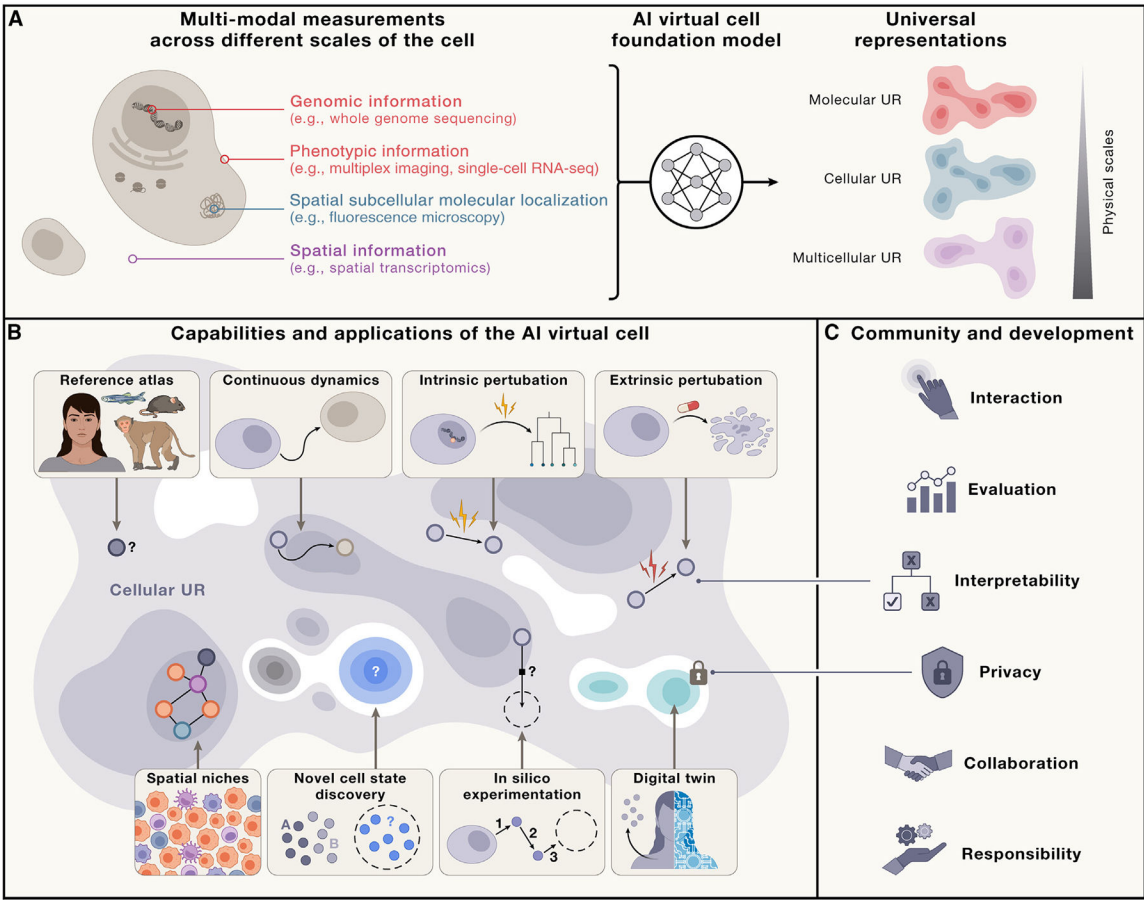
Figure 1. Capabilities of the AIVC

(A) The AIVC provides a universal representation (UR) of a cell state that can be obtained across species and conditions and generated from different data modalities across scales (molecular, cellular, and multicellular).

(B) The AIVC possesses capabilities to represent and predict cell biology. This universality allows the representation to act as a reference that can generalize to previously unobserved cell states, providing guidance for future data generation. Because the representation is shared across modalities, it also remains invariant to the specific data type used to generate it, serving as a virtual representation for unified analysis across modalities. The AIVC also allows modeling the dynamics of cells as they transition between different states, whether naturally due to processes such as differentiation or due to genetic variation or artificially through engineered perturbations. Thus, the AIVC enables *in silico* experimentation that would otherwise be cost-prohibitive or impossible in a lab.

(C) The utility of the AIVC depends on its interactions with humans at different levels. At the individual scientist level, it must be accessible through open licenses and the democratization of computing resources. Interpretability can be established through intermediary layers, such as language models that allow the virtual cell to communicate its results effectively. At the scientific community level, evaluating the AIVC should focus on core capabilities that move beyond narrow benchmarks. Community development will be

crucial for ongoing improvements to the virtual cell that remain accessible. At the societal level, the AIVC must ensure the privacy of its contents to protect sensitive data.
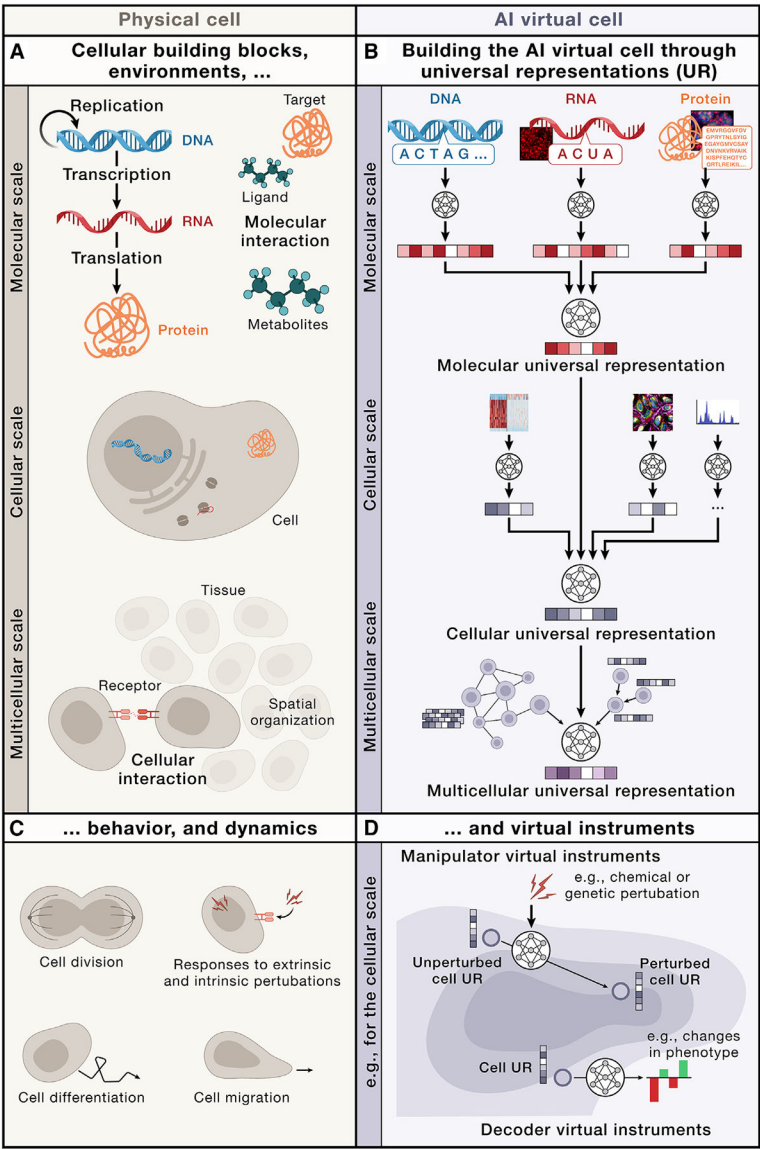
Figure 2. Overview of the AIVC

(A and B) (A) Similar to biological cells, (B) the AIVC models cell biology across different physical scales, including molecular, cellular, and multicellular. Along the physical dimension, the first scale models the state and interactions of individual molecules, such as those of the central dogma, as well as additional molecules, such as metabolites. Molecules can be represented as sequences or atomic structures. The next scale represents cells as collections of these molecules. For example, such cells contain a genetic sequence, RNA transcripts, and some quantities of proteins. Molecules within cells have specific locations that may be related to their function. The final scale models the interactions between cells and how they communicate and form complex tissues. Each scale relies on universal representations that are learned from multi-modal data and are integrating URs from the previous scale.

(C and D) (C) To capture the behavior and dynamics of physical cells, its components, or collections, (D) the AIVC comprises virtual instruments. On the cellular scale, for example, manipulator VIs simulate how cell states change as cells divide, migrate, develop from progenitor states, or respond to perturbations through learned transitions in the URs. Decoder VIs allow for the decoding of the cell UR, e.g., to understand phenotypic properties.