Grow AI Virtual Cells: Three Data Pillars and Closed-Loop Learning

Liujia Qian[1,2,3], Zhen Dong[1,2,3], Tiannan Guo[1,2,3*]

1 Affiliated Hangzhou First People's Hospital, State Key Laboratory of Medical Proteomics, School of Medicine, Westlake University, Hangzhou, Zhejiang Province, China.

2 Westlake Center for Intelligent Proteomics, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang Province, China.

3 Research Center for Industries of the Future, School of Life Sciences, Westlake University, Hangzhou, Zhejiang Province, China.

* Correspondence: guotiannan@westlake.edu.cn

Cells, the fundamental units of life, are crucial for understanding health, aging, and disease, and are essential tools in drug development and synthetic biology. However, cell-based experiments are resource-intensive and prone to variability, contributing to the reproducibility concerns in biomedical research.

While the first carbon-based cell emerged through billions of years of evolution, the development of the first silicon-based cell now presents a transformative opportunity for the science community. The concept of virtual cell or digital cell, introduced circa 2000, initially relies on traditional low-throughput biochemical experiments to quantify spatiotemporal changes in substances involved in specific biological processes [1]. These early models employed differential equations and stochastic simulations to model specific cellular processes. Pioneering whole-cell virtual models, such as those for *Mycoplasma* [2], *Escherichia coli* [3] and *Saccharomyces cerevisiae* [4,5], were primarily based on *a priori* knowledge. However, they lack rigorously designed matched perturbation omics data and spatiotemporal imaging data. While groundbreaking, these early models are limited in their ability to fully capture the dynamic nature and complexity of living cells, underscoring the need for more comprehensive data integration and advanced modeling approaches.

Recent advancements in high-throughput technologies and artificial intelligence (AI) have paved the way for more sophisticated virtual cell simulations. Bunne *et al.* recently proposed the concept of Artificial Intelligence Virtual Cells (AIVCs), which integrate AI and multi-modal data to create comprehensive computational models of cellular functions [6]. These AIVCs promise to enable precise and scalable *in silico* experimentation, potentially revolutionizing biomedical research by complementing, or even replacing conventional experiments in certain scenarios, with high-throughput simulations.

Despite the promising outlook of AIVCs, several critical questions remain unanswered. Just as a cell culture medium nourishes biological cells, what constitutes the ideal 'culture medium' for growing these digital entities? Which cell types should we prioritize for virtual cultivation? Addressing these challenges will be crucial for realizing the full potential of

AIVCs and their impact on drug development, disease modeling, and fundamental biological research. As we stand at the threshold of this new era in cellular modeling, the scientific community should collaborate to establish standards and best practices for AIVC development and validation.

Three data pillars for growing AIVCs

We propose here that the evolution or growth of AIVC relies on three essential building blocks and nutrients: *a priori* knowledge, static architecture, and dynamic states. These data pillars, when combined with deep learning algorithms, form the foundation for AIVC development (**Figure 1**).

The biomedical community has generated vast amounts of cell-related data, including text-based literature, molecular expression data, and multi-scale imaging from the organismal to the nanoscale level. With the rapid advancement of AI, we speculate that a comprehensive foundation model integrating all these data sources could be developed, serving as a fundamental basis for constructing AIVC. We designate *a priori* knowledge as the first pillar of AIVC construction. Despite its vast size and diversity, this knowledge base primarily consists of fragmented information across different cell types and populations. While it is unrealistic to build a fully functional AIVC for a specific cell type solely from these data, they encapsulate fundamental cell biological mechanisms essential for model construction. Furthermore, as these data already exist and require no additional generation cost for AIVC developers, they provide an ideal starting point for building AIVC.

However, while the *a priori* knowledge pillar is rich in diverse cell biology information, it cannot be directly used to construct a specific AIVC model. To achieve this, a comprehensive characterization of a specific cell is required, capturing its complete cellular structures at both the morphological and molecular expression levels, along with their interactions. We define this second essential pillar of AIVC construction as static architecture, which integrates nanoscale molecular structures and spatially resolved data from molecular modeling, cryo-electron microscopy, cryo-electron tomography, correlative light and electron microscopy, super-resolution fluorescence imaging, spatial omics, and other multi-scale data. Additionally, tissue expansion techniques[7,8] can further enhance spatial resolution, complementing the high-resolution imaging methods and omics technologies mentioned above. This integrated approach provides a detailed three-dimensional context essential for accurate AIVC modeling.

While static data provide a *bona fide* snapshot of the cell, they fail to capture the dynamic nature of living systems. To construct a live AIVC, we introduce dynamic states as the third pillar for AIVC development. These data encompass natural processes such as aging, development, and carcinogenesis, as well as induced perturbations, including physical, chemical, and genetic interventions. Cellular dynamics are historically studied by measuring the expression or activity of one or a few molecules at a time. With advancements in high-throughput omics technologies—such as transcriptomics, proteomics, and metabolomics—it is now possible to profile thousands of molecules across diverse cellular states. To build an effective AIVC, it is essential to comprehensively capture a wide range of cellular states and maximize their diversity to ensure high accuracy in differentiating them, requiring large

volumes of dynamic, cell-specific data. Given the limited number of naturally occurring states, artificial perturbations serve as effective tools to generate different cellular states. Among these, perturbation proteomics is particularly valuable, as proteins are the primary structural components and catalysts of cellular biochemical processes[9]. A recent AIVC pilot study integrating perturbation proteomics and AI algorithms demonstrates accurate prediction of drug efficacy and synergistic combinations, highlighting the critical role of dynamic perturbation proteomics data in constructing robust virtual cell models for drug discovery and cellular simulation [10].

Although single-cell omics technologies provide large datasets of millions of cells, the value for constructing AIVC is limited due to the similarity of the cells' states. Antibody-based methods, such as those used in the Human Protein Atlas [11], are valuable, but mass spectrometry-based proteomics offers distinct advantages in measuring 1000s of proteins, protein post-translational modifications, and complex dynamics without the need for affinity-based reagents[12]. To better understand the impact of perturbations on cellular behavior, emerging spatial omics technologies enable large-scale mapping of molecular distributions, providing insights into how perturbations alter cellular processes in their native spatial context. In particular, spatial proteomics represents the forefront of this advancement [8,13]. In addition, innovative sample preparation methods now allow simultaneous multi-omics analysis of the same sample [14]. We argue that the AI-driven integration of static and dynamic data is essential for constructing a functionally robust and predictive AIVC.

The integration of multimodal data from *a priori* knowledge, static architecture, and dynamic states demands sophisticated AI frameworks capable of hierarchical reasoning, cross-modal alignment, and predictive simulation. Foundational architectures such as transformers, convolutional neural networks (CNNs), and diffusion models provide critical building blocks for data processing and feature extraction. Future advancements in AI algorithms will further enhance the fidelity, generalizability, and predictive power of AIVCs.

The ultimate purpose of these models is multifaceted, addressing key challenges in systems biology and personalized medicine. AIVCs aim to infer molecular states across omics layers, forecast molecular states based on physiological inputs, and predict cellular outcomes following perturbations or in specific conditions based on baseline molecular states. By integrating diverse data types and extracting complex, non-linear relationships across biological scales, AIVCs leverage their capacity to provide unprecedented insights into cellular behavior.

Evolution of AIVC: Closed-Loop Active Learning Systems

In our vision for the future of AIVC, we foresee a transition from static, data-driven models to adaptive systems capable of evolutionary intelligence. While traditional approaches relied on passive data assimilation, modern closed-loop architectures enable AIVCs to actively interrogate biological reality through AI model, autonomous robots, and dynamic data. The operational framework for closed-loop AIVC development draws inspiration from recent breakthroughs in autonomous chemistry laboratories [15]. Central to this vision is the establishment of closed-loop frameworks that integrate computational prediction with robotic experimentation, specifically targeting gaps in dynamic state data (**Figure 1**). Unlike static

data repositories or fixed simulation parameters, closed-loop systems establish a self-optimizing workflow where AI algorithms continuously identify knowledge gaps in dynamic response patterns and automatically design and execute multiplexed perturbation experiments to resolve uncertainties in molecular interaction networks. The cycle is completed as the system validates predictions through real-time comparison of *in silico* and *in vitro* outcomes. This autonomous experimentation cycle fundamentally transforms the temporal resolution of model refinement. While classical approaches required years of manual hypothesis testing in exploratory synthetic chemistry research, closed-loop systems can achieve equivalent knowledge gains through mere weeks of targeted robotic experimentation, dramatically accelerating the pace of scientific discovery and understanding [15].

A critical challenge in modeling dynamic states is the combinatorial complexity of cellular responses to perturbations. While existing datasets capture snapshots of induced or natural states, they often lack systematic coverage of the parameter space. Our proposed closed-loop active learning systems could prioritize high-impact perturbations—such as CRISPR-based gene knockouts, small-molecule treatments, or optogenetic triggers—based on their potential to reduce model uncertainty or reveal novel regulatory mechanisms. For example, an AIVC trained on baseline proteomic profiles might identify understudied phosphorylation events in stress response pathways, prompting robotic platforms to perform time-resolved phosphoproteomics under targeted metabolic perturbations. This feedback loop would not only refine the model's understanding of signaling dynamics but also generate purpose-built datasets that maximize biological insight per experiment. As robotic throughput increases and multimodal data integration matures, AIVCs may soon autonomously guide the resolution of longstanding questions in cell biology—from decoding context-specific protein functions to engineering synthetic cellular behaviors.

Low-hanging fruits

Selecting a proper cellular model for the inaugural AIVC is a crucial decision that will shape the development and validation of this groundbreaking technology. Several candidates merit consideration, each with its own advantages and limitations. As one of the simplest self-replicating organisms, mycoplasma offers a minimalist system for modeling. However, its unique biology may limit broader applicability. *E. coli*, the well-studied prokaryote, provides a wealth of existing data and a simpler cellular structure. Its rapid growth and ease of genetic manipulation are advantageous, but it lacks the complexity of eukaryotic cells. As a eukaryote with subcellular organelles similar to human cells, yeast offers a balance between simplicity and relevance to higher organisms. Its genetic tractability and established role in biotechnology make it an attractive candidate. The immortalized human cancer cell lines (*e.g.*, HeLa, HEK293) are widely used in research and associated with vast amounts of phenotypic and omics data.

While extensive experimental data exist for these cell types, significant gaps remain across the three pillars, particularly in their dynamic states. Notably, perturbation proteomics data, which are essential for building a comprehensive and dynamic AIVC, are scarce across all the candidate cell types.

We propose starting with a relatively simple yet informative model, such as *S. cerevisiae*. This organism has relatively small genomes and boast a wealth of perturbation omics and imaging data, coupled with established protocols for genetic manipulation and high-throughput experiments. As a eukaryote, yeast presents a multi-compartmented cellular structure, providing an excellent platform to model complex intracellular organization and dynamics. This feature allows for a more comprehensive representation of eukaryotic cellular processes, bridging the gap between prokaryotic and higher eukaryotic systems. Moreover, *S. cerevisiae*'s relevance extends beyond basic research to applied fields such as synthetic biology and drug screening, enhancing the potential impact of the AIVC.

Although *S. cerevisiae* presents compelling advantages as an initial model, human cancer cell lines remain pivotal candidates for subsequent AIVC development. Their pervasive use in biomedical research, immediate relevance to human pathophysiology, and potential to revolutionize drug discovery and personalized medicine render them invaluable targets for AIVC modeling.

Developing an AIVC for these simpler organisms can serve as a proof of concept, allowing us to address fundamental questions posed by Bunne *et al*.[6]: What are the specific data needs and requirements for building an AIVC? How much data is necessary to construct a robust and predictive AIVC? How can we develop a comprehensive and adaptable benchmarking framework to evaluate AIVC performance? By tackling these questions with a simpler model organism, we can refine our methodologies and establish best practices before advancing to more complex cellular systems. This stepwise approach will provide valuable insights into the scalability of the AIVC concept and inform future efforts to model more complex eukaryotic cells and cell populations.

Conclusion and outlook

As we create and grow AIVCs in the digital petri dish of modern biomedical research, we must carefully consider the 'nutrients' that will nourish their growth. Our proposed three data pillars of *a priori* knowledge, static architecture, and dynamic states forms the essential medium for these *in silico* entities. Among these, perturbation-based omics data - transcriptomics, proteomics, and metabolomics - emerge as the critical growth factors.

To efficiently generate this wealth of perturbation data, we envision Closed-Loop Active Learning Systems as the next evolutionary step. These systems, inspired by autonomous chemistry laboratories, will seamlessly integrate AI-driven predictions with robotic experimentation. Like a skilled gardener, they will identify knowledge gaps, design targeted experiments, and continuously refine our understanding of cellular complexities. The journey from static models to adaptive, self-optimizing AIVCs promises to revolutionize drug discovery, disease modeling, and fundamental biological research. We also propose the low-hanging fruits along the journey. Creating and growing a virtual yeast cell might be a valid option. As we stand on the brink of this exciting frontier, the collaborative efforts of the scientific community will be crucial in realizing the full potential of AIVC and driving the future of *in silico* life sciences.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (Major Research Plan, Grant No. 92259201 to T.G.), the National Natural Science Foundation of China (Key Joint Research Program, Grant No. U24A20476 to T.G.), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024SSYS0035 to T.G.), and the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation (BX20240330 to L.Q.). We thank heated discussion in the 2024 Westlake Symposium for the Future of Proteomics and the 2nd π-hub Global Summit. Special thanks to Dr. Ruedi Aebersold and Dr. Dangsheng Li for discussion.

Declaration of interests

T.G. is the shareholder of Westlake Omics (Hangzhou) Biotechnology Co., Ltd. The remaining authors declare no competing interests.

References

1 Slepchenko, B. M., Schaff, J. C., Macara, I. & Loew, L. M. Quantitative cell biology with the Virtual Cell. *Trends Cell Biol* **13**, 570-576, doi:10.1016/j.tcb.2003.09.002 (2003).

2 Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389-401, doi:10.1016/j.cell.2012.05.044 (2012).

3 Macklin, D. N. *et al.* Simultaneous cross-evaluation of heterogeneous E. coli datasets via mechanistic simulation. *Science* **369**, doi:10.1126/science.aav3751 (2020).

4 Ye, C. *et al.* Comprehensive understanding of Saccharomyces cerevisiae phenotypes with whole-cell model WM_S288C. *Biotechnol Bioeng* **117**, 1562-1574, doi:10.1002/bit.27298 (2020).

5 Osterlund, T., Nookaew, I., Bordel, S. & Nielsen, J. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Syst Biol* **7**, 36, doi:10.1186/1752-0509-7-36 (2013).

6 Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045-7063, doi:https://doi.org/10.1016/j.cell.2024.11.015 (2024).

7 Wang, S. *et al.* Single-shot 20-fold expansion microscopy. *Nat Methods* **21**, 2128-2134, doi:10.1038/s41592-024-02454-9 (2024).

8 Dong, Z. *et al.* Spatial proteomics of single cells and organelles on tissue slides using filter-aided expansion proteomics. *Nat Commun* **15**, 9378, doi:10.1038/s41467-024-53683-7 (2024).

9 Qian, L. *et al.* AI-empowered perturbation proteomics for complex biological systems. *Cell Genomics* **4**, doi:10.1016/j.xgen.2024.100691 (2024).

10 Sun, R. *et al.* A perturbation proteomics-based foundation model for virtual cell construction. *bioRxiv*, 2025.2002.2007.637070, doi:10.1101/2025.02.07.637070 (2025).

11 Sjostedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, doi:10.1126/science.aay5947 (2020).

12 Guo, T., Steen, J. A. & Mann, M. Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature*, In press (2025).

13 Nordmann, T. M., Mund, A. & Mann, M. A new understanding of tissue biology from MS-based proteomics at single-cell resolution. *Nat Methods* **21**, 2220-2222, doi:10.1038/s41592-024-02541-x (2024).

14 Li, W. *et al.* Integral-Omics: Serial Extraction and Profiling of Metabolome, Lipidome, Genome,

Transcriptome, Whole Proteome and Phosphoproteome Using Biopsy Tissue. *Anal Chem* **97**, 1190-1198, doi:10.1021/acs.analchem.4c04421 (2025).

15 Dai, T. *et al.* Autonomous mobile robots for exploratory synthetic chemistry. *Nature* **635**, 890-897, doi:10.1038/s41586-024-08173-7 (2024).

Figure 1

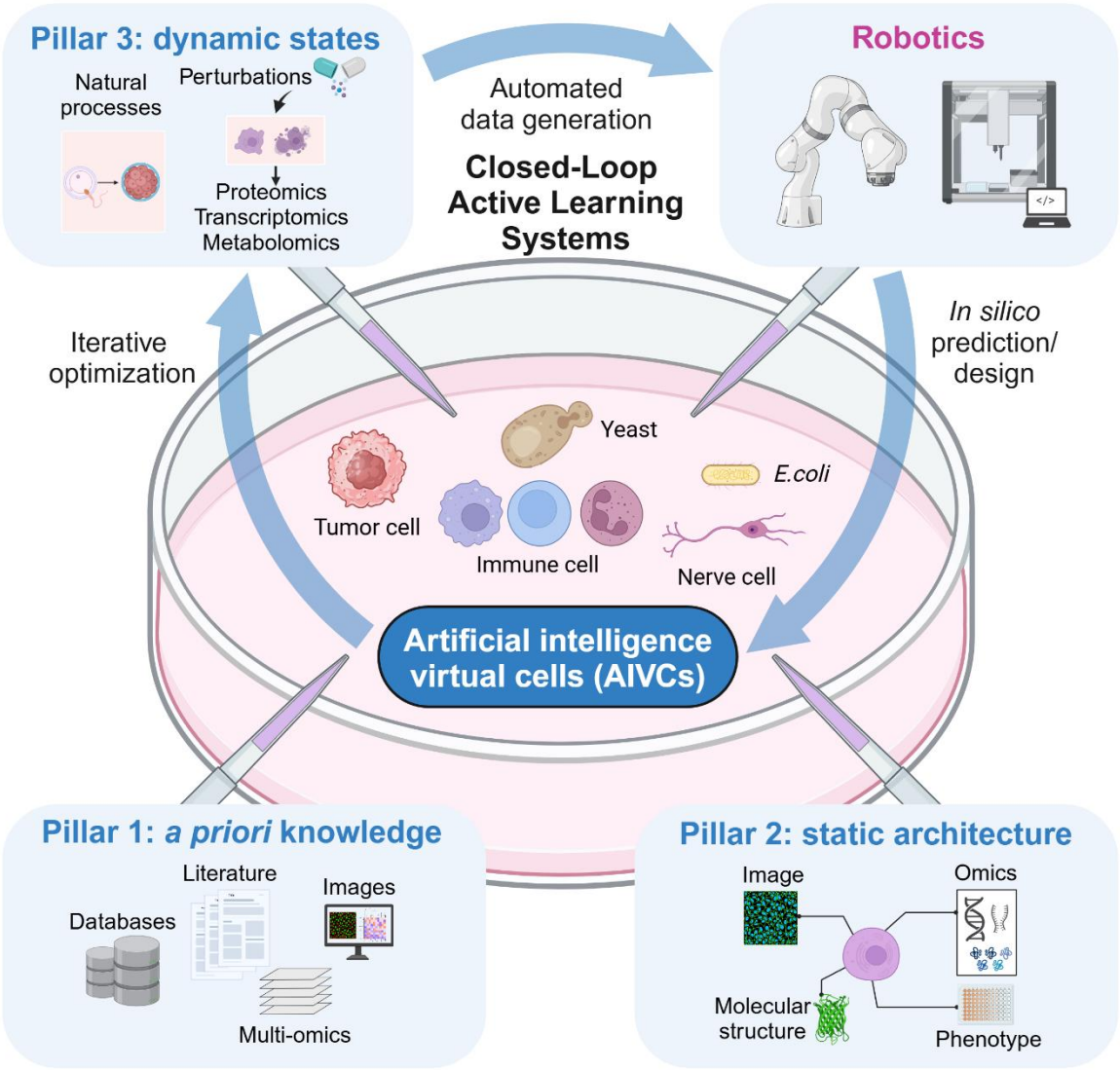

Figure 1. Data pillars for AIVC growing and evolution through closed-loop learning. This schematic illustrates the three key pillars for growing AIVCs: *a priori* knowledge, static architecture, and dynamic states. These are integrated using AI algorithms to model cellular behavior, with examples of model organisms like *E. coli*, and yeast. It also showcases the evolution of AIVCs towards closed-loop active learning systems. In this advanced framework, computational predictions guide automated experimentation, with a particular focus on perturbation omics. AIVC – Artificial intelligence virtual cell.