How to Build the Virtual Cell with Artificial Intelligence: Priorities and Opportunities

Cell 187, December 12, 2024

Literature Report

12212859 Sijie Li

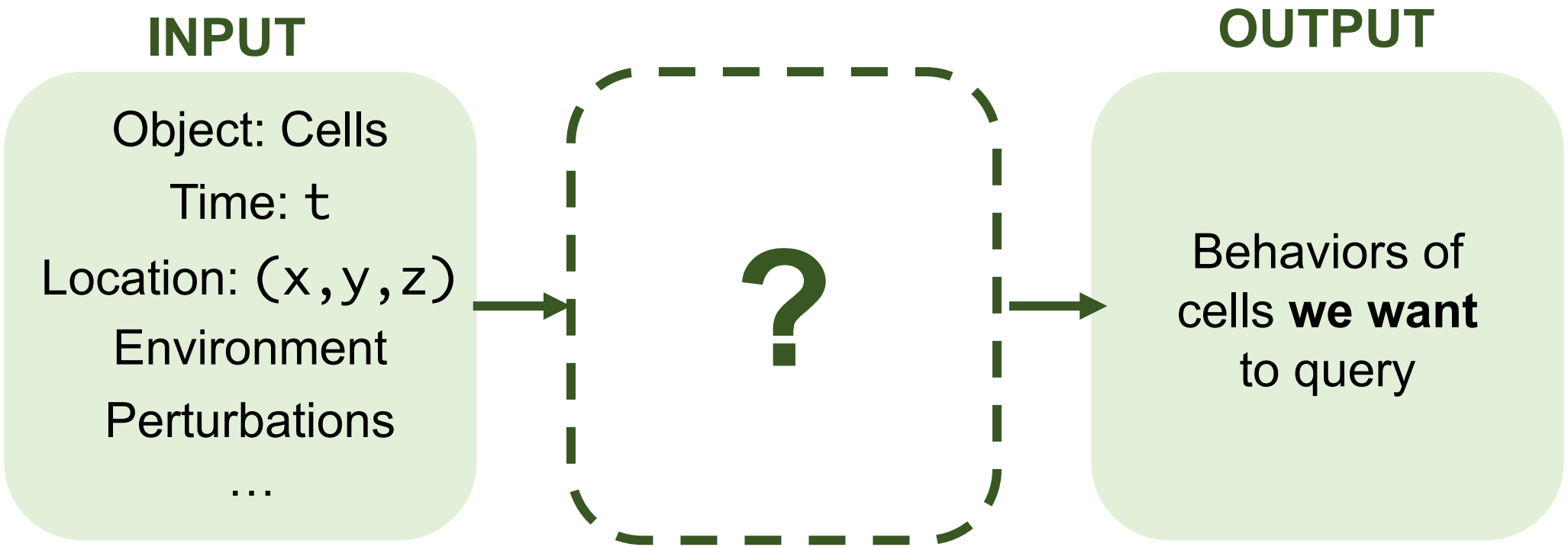
June 4th, 2025

GOAL

How to Build the **Virtual Cell** with Artificial Intelligence

➢ **Definition**

“simulator of cells and cellular systems under varying conditions and changing contexts”

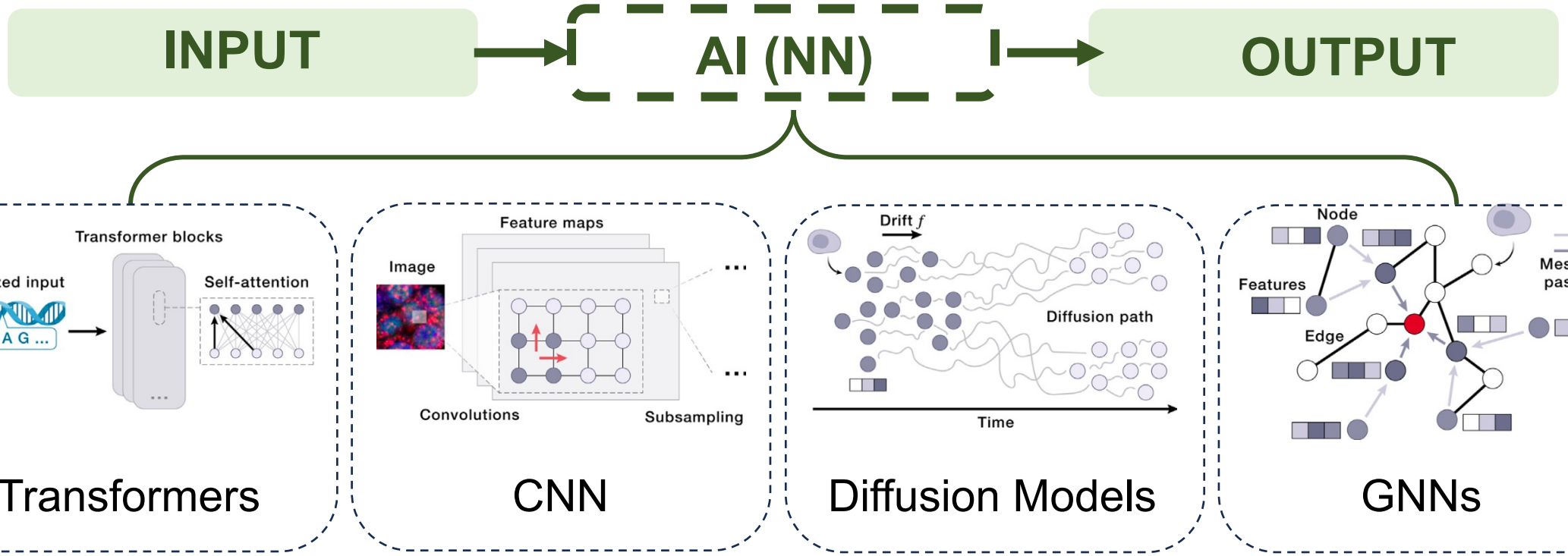


TOOL

How to Build the Virtual Cell with **Artificial Intelligence**

➢ Definition

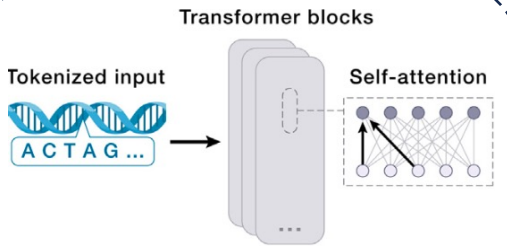
the ability of machines to **perform tasks** that **require human intelligence**



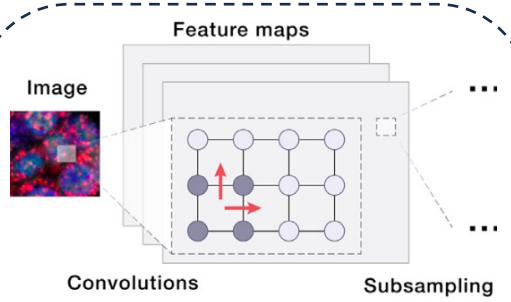
Artificial Intelligence (AI) → Machine Learning (ML) → Deep Learning (DL) → Neural Networks (NN)

TOOL

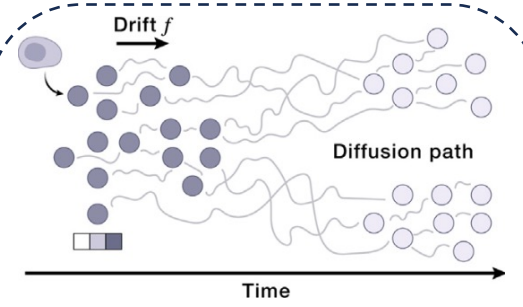
How to Build the Virtual Cell with **Artificial Intelligence**



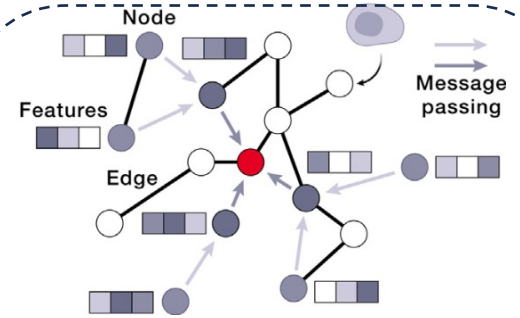
Transformers
sequences
→multimodalities



CNN
regular “network”
e.g. images



Diffusion Models
data generation
e.g. time trajectory



GNNs
node + edge objects
e.g. protein structure

TOOL

How to Build the Virtual Cell with **Artificial Intelligence**

Transformers | **CNN** | **Diffusion Models** | **GNNs**

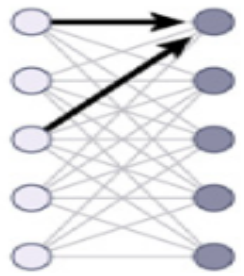
Raw Data —(**Primary coding**)→ **Primary Embedding** —(**Nueral network**)→ **Latent Space Embedding**

TANSFORMER

Gene Expression —(**Simple Mapping**: Linear projection or **table lookup**)→ **token embedding** —(**e.g. self-attention**)→ **contextual embedding**

Gene A, B, C, D, E
[5.2, 0.0, 1.8, 3.3, 7.1]

Gene A (5.2) → [0.4, 1.1, -0.2]
Gene B (0.0) → [0.0, 0.0, 0.0]
Gene C (1.8) → [0.3, 0.2, 0.1]
Gene D (3.3) → [0.6, 0.7, -0.1]
Gene E (7.1) → [1.0, 1.2, 0.4]



× n layers

[[0.48, 0.90, 0.03],
[0.10, 0.05, 0.12],
[0.30, 0.35, 0.25],
[0.52, 0.60, 0.10],
[0.95, 1.00, 0.60]]

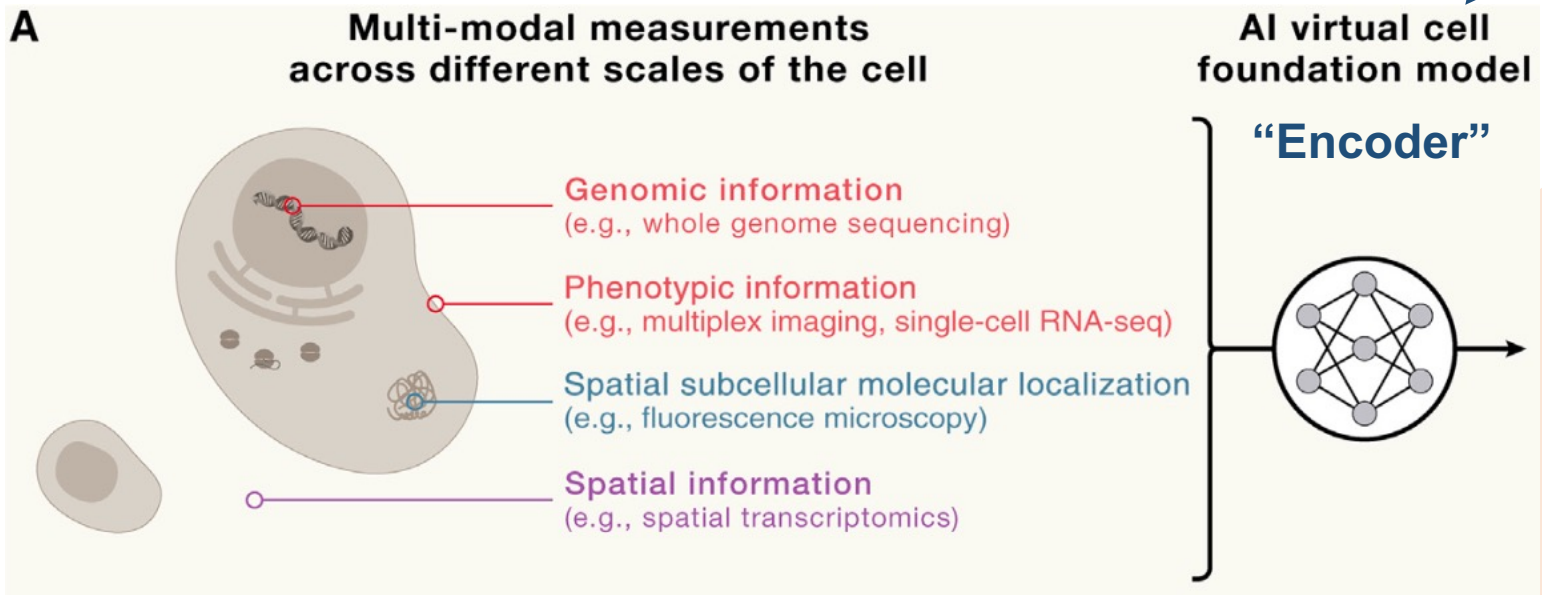
mean pooling↓
[0.47, 0.58, 0.22]

➢ **Embedding**: a representation for computation
① numerical ② preserve structure ③ enable further manipulation

Method

How to Build the Virtual Cell with Artificial Intelligence

Raw Data → Primary coding → Primary Embedding → Nueral Network → Latent Space Embedding



✓ Multi-modality
✓ Biological structures

URs
- Molecular
- Cellular
- Multicellular

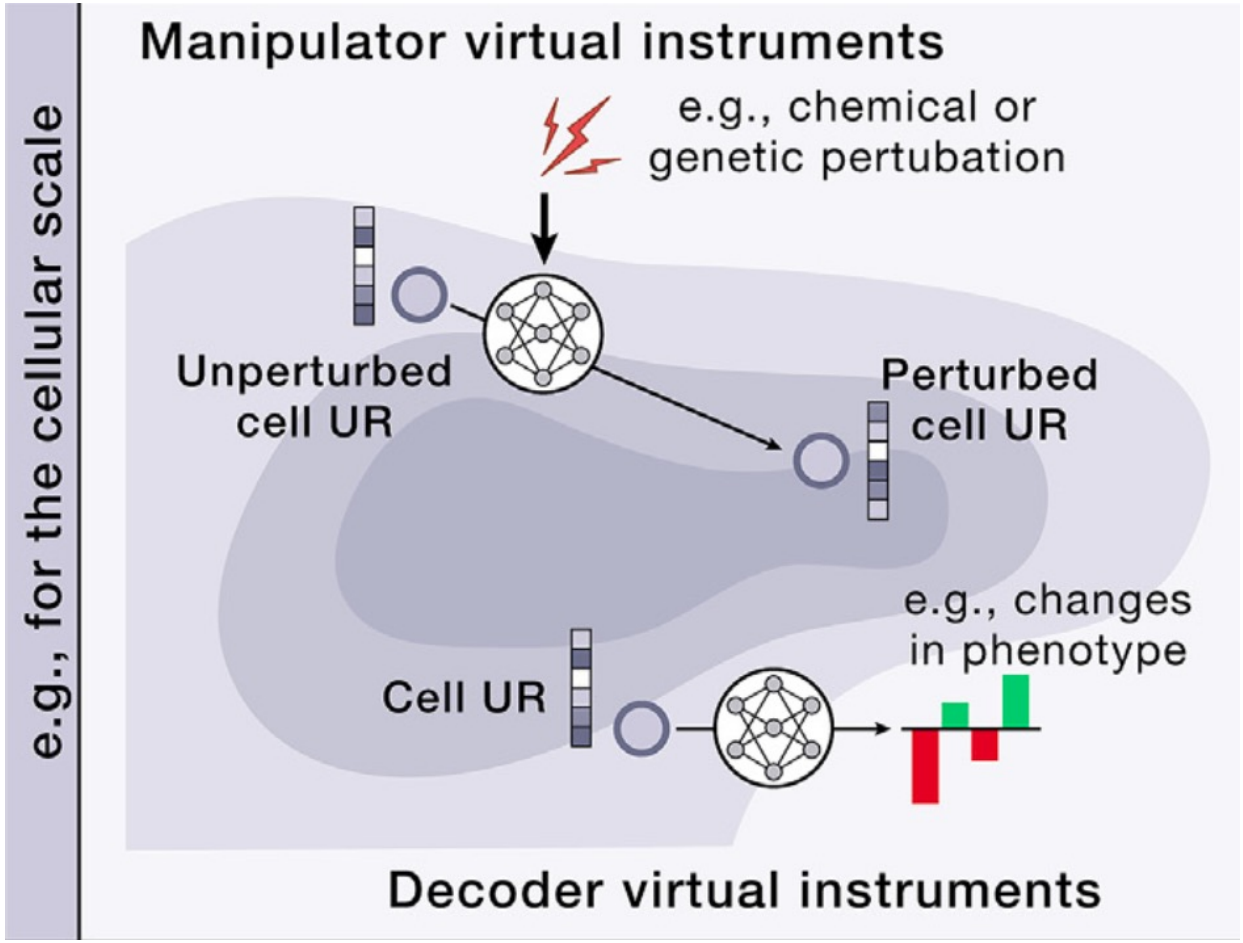
UR abstraction ↓ Physical scales

➢ **Universal Representations (URs)**

"**embeddings** produced by a **multi-modal** AIVC **foundation model**"

How to Build the Virtual Cell with Artificial Intelligence

Raw Data → **AIVC Foundation Model** → **UR** → **Neural Network** → **Cells Behavior**



➢ Virtual Instruments (VIs)

Neural networks that take URs as input and produce a desired output

pre UR [0.47, 0.58, 0.22]

↓ **Manipulator VI**
e.g. drug administration

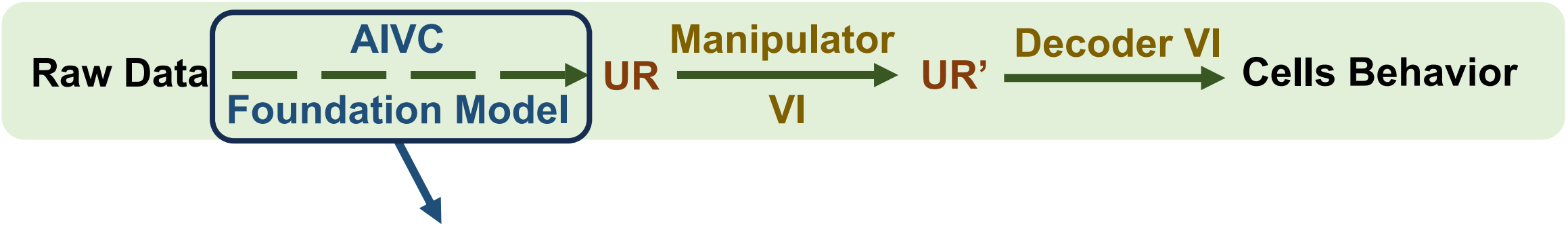
pro UR [0.97, 0.04, 0.23]

↓ **Dcoder VI**

Human understanderble

Gene A, B, C, D, E
[4.9, 30.0, 1.7, 17.8, 0.7]

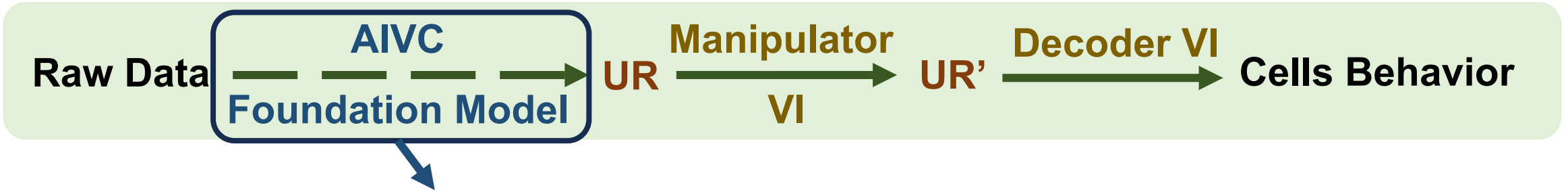
AIVC: Priorities and Opportunities



➢ **Tool Building:** Architectures for **multi-modal biological** data

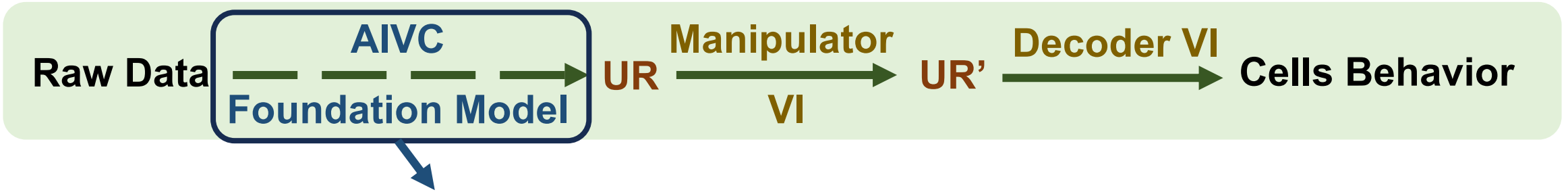
- "**Interconnection** of many large models to **capture everything** in UR" ---- Charlotte Bunne
- Self-consistent across scale, modality, and context
- **LACK!** DNA ≠ texts, medical images ≠ images

AIVC: Priorities and Opportunities



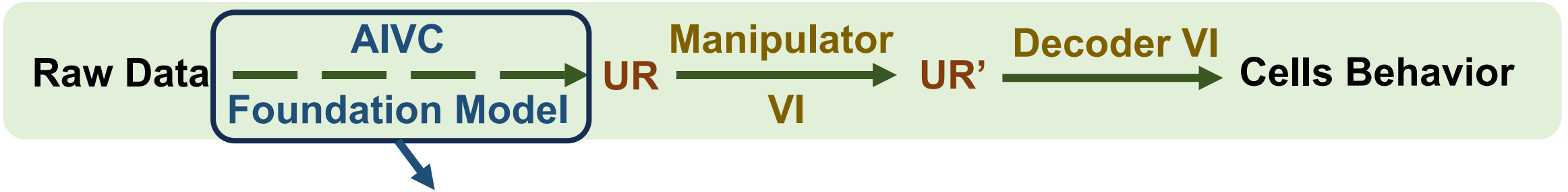
- **Tool Building:** Architectures for **multi-modal biological** data
- **Data Collection**
  - Multi-modal, multi-scale, multi-dimension (t, x), multi-species
  - **Information amount**: short read 14 pb > 1,000×dataset used to train ChatGPT **Redundant?**
  - **Perturbation**: large **combinatorial space**; in vivo? → organoid…

AIVC: Priorities and Opportunities



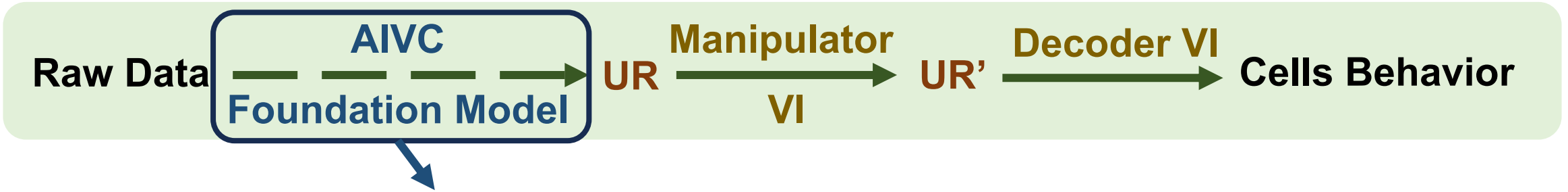
- **Tool Building:** Architectures for **multi-modal biological** data
- **Data Collection**
- **Model Evaluation**
  - Generalizability
  - Discovering new biology, ↑understanding

AIVC: Priorities and Opportunities



- **Tool Building:** Architectures for **multi-modal biological** data
- **Data Collection**
- **Model Evaluation**
- **Interpretability VS Utility**
  - Modular structure enable pinpointing
  - Project result to **interpretable space** (decoder VI)

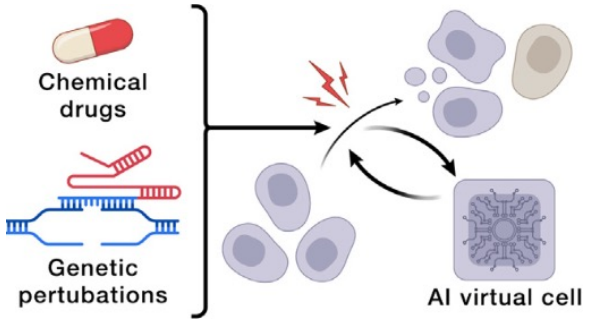
AIVC: Priorities and Opportunities



- **Tool Building:** Architectures for **multi-modal biological** data
- **Data Collection**
- **Model Evaluation**
- **Interpretability VS Utility**
- **Community**: non-expert interface, for collaboration!
- **Humanity Consideration**: diversity, privacy, ethics…

AIVC: Priorities and Opportunities

Raw Data --▸ **UR** — **VI** —▸ **Cells Behavior**



➢ Predict Response

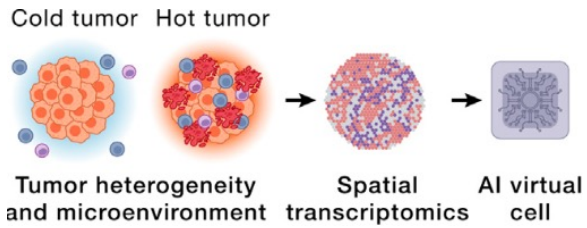
Cell state 1
e.g. expression profile

Engineering
- Cell therapy
- Drug delivery

Cell state 2 (Response)

Optimization

➢ Unlock Biological Laws



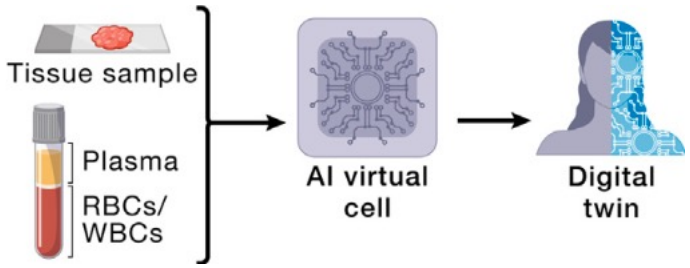
Tumor spatial **omics +**
Microenvironment **conditions**

Laws of tumor
Heterogeneity insights
Pan-cancer markers

"Machine Learning is the formalism through which we understand high dimensional data."

AIVC: Priorities and Opportunities

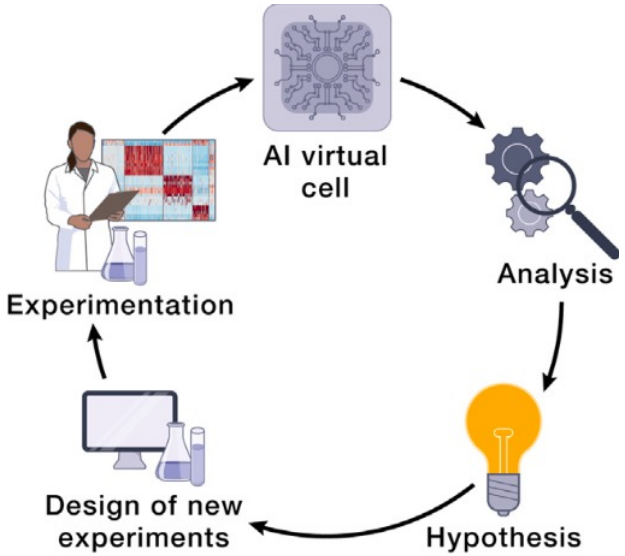
Raw Data --▸ **UR** —VI—▸ **Cells Behavior**



➢ Digital Twin

Patient samples
(easy access)
e.g. blood, images etc.

—▸ **Digital twin**
(inaccessible cell types)
→ personalized simulations



➢ Guide Experiments

Cellular system state —intervention—▸ **Result with confidence**

Gene A, B, C, D, E
[5.2, 0.0, 1.8, 3.3, 7.1]

Gene A, B, C, D, E
[4.9, 30.0, 1.7, 17.8, 0.7]
Possibility = 0.0001

Low!
Test it!

New Biology or
Model Optimization

◂—

- Generate **hypothesis**
- Identify **valuable data** to collect

"Everything we do in science is hypothesis driven"

AIVC: Priorities and Opportunities For us!

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,*] and Stephen R. Quake[3,7,49,*]

- 42-author 50-institute blueprint
- A **roadmap** for a once-vague dream
- Require interdisciplinary collaboration
- BOLD! **Doable**?
- **Together**, I believe **YES!**

[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
[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

References

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., Califano, A., Cool, J., Dernburg, A. F., Ewing, K., Fox, E. B., Haury, M., Herr, A. E., Horvitz, E., Hsu, P. D., … Quake, S. R. (2024). How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, *187*(25), 7045–7063. https://doi.org/10.1016/j.cell.2024.11.015

Topol, E. (2025, March 18). Steve Quake and Charlotte Bunne: The Holy Grail of Biology. *Ground Truths*. https://erictopol.substack.com/p/the-holy-grail-of-biology