Briefings in Bioinformatics, 2026, 27, bbag104
https://doi.org/10.1093/bib/bbag104
Published: 8 March 2026
Review

OXFORD

Artificial intelligence-enabled multi-scale virtual cell: perspective, challenges, and opportunities

Huasen Jiang[1], Xiaoyu Huang[1], Xiangpeng Bi[1], Wenjian Ma[1], Haibo Ni[2], Zhiqiang Wei[1,3], Pin Sun[4,*], Henggui Zhang[5], Shugang Zhang[1,*]

[1]College of Computer Science and Technology, Ocean University of China, 238 Songling Road, Laoshan District, Qingdao 266404, China
[2]Medical School of Nanjing University, 22 Hankou Road, Gulou District, Nanjing 210008, China
[3]College of Computer Science and Technology, Qingdao University, 308 Ningxia Road, Shinan District, Qingdao 266071, China
[4]Department of Echocardiography, The Affiliated Hospital of Qingdao University, 16 Jiangsu Road, Shinan District, Qingdao 266000, China
[5]School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

*Corresponding authors. Pin Sun, Department of Echocardiography, The Affiliated Hospital of Qingdao University, 16 Jiangsu Road, Shinan District, Qingdao 266000, China. E-mail: sunpin@qduhospital.cn; Shugang Zhang, College of Computer Science and Technology, Ocean University of China, 238 Songling Road, Laoshan District, Qingdao 266404, China. E-mail: zsg@ouc.edu.cn.

Abstract

As the fundamental unit of life, cells coordinate biological activities through the interaction between microscopic molecular mechanisms and macroscopic tissue organization. Traditional research studies, experiments, and biochemical analyses, give rise to important insights, although they are restricted in spatiotemporal resolution and processing power, thereby precluding the understanding of dynamic cross-scale biological events . Breakthroughs in artificial intelligence (AI) have given birth to the AI virtual cell (AIVC) as a new way to do research. By integrating multi-omics data and mixing methods from multidisciplinary models, AIVC establishes a digital twin system to simulate cell functions and behaviors. AIVC still faces a number of pressing challenges that need to be addressed in its current development stage. In this review, we are proposing a unified definition and technical framework for AIVC and analyze in detail the cross-scale coupling mechanisms of the "gene–protein–pathway–cell" hierarchy. Furthermore, we decompose the technical construction framework of AIVC from cross-scale representation engineering, functional submodule design, and multi-component dynamic regulation mechanisms. Additionally, we summarize the existing models and datasets in the field to provide reference resources for researchers. Finally, we deeply discuss the challenges faced by AIVC, such as data heterogeneity and model interpretability, and aim to accelerate the research progress in the AIVC field while driving the life sciences to shift from observational analysis to a paradigm that integrates predictability and innovation. Despite being in the early stage, AIVC is a trending topic that has garnered widespread interest. This review aims to integrate existing models, datasets, and technical ideas to provide a unified framework for field development.

Keywords virtual cell, AI for science, smart healthcare, bioinformatics

Introduction

Cells, as the fundamental structural and functional units of life, have long held a central position in life sciences research [1, 2]. At the microscopic level, various organelles and biomolecules within cells cooperate with one another to carry out basic life activities such as material metabolism, energy conversion, and information transmission [3]. On the other hand, at the macroscopic level, cells constitute the basis of tissues, organs, and even entire organisms. Different cell types combine and differentiate in specific ways to form tissues and organs with specialized functions [4]. Comprehending the mechanisms of cell proliferation, differentiation, and apoptosis processes is essential for uncovering the mechanisms of individual development and tissue regeneration and repair [5].

In the field of traditional cell research, researchers have long relied on experimental observation and manual data analysis as core technical supports. These research methods mostly depend on conventional biochemical technology systems, and explore the regulatory mechanisms of cellular behaviors by analyzing the molecular composition and metabolic pathway characteristics within cells [6, 7]. For instance, in genetic research, researchers often employ traditional techniques such as gene knockout and gene editing to investigate the functions of specific genes in cells [8]. In the field of cellular research, methods like radioactive labeling and Western

Received: December 12, 2025. **Revised:** February 5, 2026. **Accepted:** February 11, 2026
© The Author(s) 2026. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

blotting are used to track the transmission process of signaling molecules [9]. However, traditional cell research is constrained by equipment resolution and field of view, making it difficult to capture subtle and rapid dynamic changes within cells, while also entailing substantial investment of time, labor, and material resources [10]. In addition, with the exhaustive breakthroughs in biological data analysis technologies, biological data have grown exponentially, and this renders traditional biological research methods increasingly inadequate to meet current demands [5, 11, 12]. Therefore, there is an immediate requirement for us to develop more efficient, exhaustive, and intelligent research approaches to tackle these challenges.

In recent years, the rapid advancement of artificial intelligence (AI) technology has brought new opportunities in the field of cell research. Scientists have gradually begun to construct virtual cell models based on AI technology, aiming to replace traditional cell research methods [13–16]. To achieve this goal, researchers have carried out extensive explorations and attempts in virtual cell modeling [17, 18]. The concept of virtual cell modeling traces back to mechanistic frameworks like the "Virtual Cell (VCell)" software proposed by Loew *et al.*, which simulates cellular processes based on established biochemical rules [19], alongside representative counterparts, including E-Cell [20] and CellML [21]. Specifically, VCell and E-Cell specialize in the systematic simulation of cellular biochemical networks, with VCell emphasizing broad adaptability across cell types and E-Cell focusing on organelle-scale metabolic modeling. In contrast, CellML serves as a standardized markup language to address cross-platform incompatibility of biological models, facilitating collaborative development, and reuse of simulation tools. With the surge of multi-omics data and AI advancements, the field has gradually shifted from pure mechanistic simulation to exploring integration with data-driven prediction—an evolution that informs the artificial intelligence virtual cell (AIVC) framework proposed herein. As an emerging research approach, the AIVC integrates multidisciplinary knowledge including cell biology, biophysics, mathematical modeling, and AI algorithms, and is capable of building highly realistic virtual cell models in computers [22, 23]. AIVC is not a single implemented system but an interdisciplinary research paradigm integrating cell biology, AI algorithms and multi-scale modeling, whose core value lies in integrating scattered technologies to achieve a leap from single-scale simulation to full-chain dynamic prediction. Representative methods related to AIVC are summarized in Table 1. For example, in 2022, the University of Illinois at Urbana-Champaign developed a computational model of the minimized synthetic cell JCVI-syn3A that exhaustively simulates cellular metabolism, gene expression, and growth processes [24]. Hao *et al.* developed a large-scale pretrained model called scFoundation, with 100 million parameters, pretrained on over 50 million human single-cell transcriptomic profiles. This model effectively captures the complex contextual relationships among genes across various cell types and states [25]. Fu *et al.* [26] developed an interpretable foundational model, GET, which achieves high-precision prediction of gene expression in 213 human cell types based on chromatin accessibility and sequence information. This model has precisely pinpointed the distal regulatory regions and functional transcription factor interaction mechanisms in fetal red blood cells. In addition, many experts in the field have conducted exhaustive and systematic reviews of the model development and research progress of AIVC [13, 27–29].

Although the existing AIVC models have been widely applied, they still fail to achieve refined simulation of cellular behaviors and multi-scale functional prediction. We have summarized the following three main issues:

- Mathematical modeling challenges in cross-scale coupling. Cell research spans multiple functional scales, and microscopic responses often undergo complex nonlinear transitions when moving from one scale to another.
- Heterogeneous fusion bottlenecks for multi-modal biological data. Functional dimensions like genes, proteins, and signaling pathways correspond to diverse data modalities. Semantic mismatches between these modalities impede the synergistic complementary effects across different modal data.
- Computational dilemma in multi-scale interaction. The cross-coupling of multi-scale parameters causes dimensionality explosion, exceeding current computational limits. Meanwhile, the inherent fault tolerance of biological systems fundamentally conflicts with deterministic modeling, making it challenging to balance efficiency, and reliability in large-scale simulations.

Recently, the Chan Zuckerberg Initiative unveiled its "Billion Cell Project"—an effort to assemble an unparalleled dataset of one billion cells that delivers vital resources to advance biomedical AI models [30]. To drive standardization in this field, the Arc Institute has teamed up with Stanford University, the University of California, San Francisco, and other partners to launch the world's first Virtual Cell Challenge [31]. These developments make it clear that the field of AIVC is entering a key period of opportunity, where its innovative, revolutionary technologies are set to open new frontiers in cell biology research.

AIVC integrates multi-omics data and cross-scale dynamic modeling to build a digital platform that simulates cellular processes (Fig. 1). This data-driven approach overcomes traditional experimental limits in spatiotemporal resolution, enabling real-time simulation of cellular activities, prediction of regulatory mechanisms, and identification of key disease nodes. Fueled by these capabilities, AIVC is set to spark a biological revolution that moves humanity from merely observing life to entering a new era of programming life.

Basic concepts and related theories of cross-scale virtual cell construction

Definition of AIVC

Initially, the concept of a virtual cell aimed to reconstruct cellular structure and function through computational integration of multi-scale data [43, 44]. To advance this concept with a focus on contemporary AI-driven approaches, this review defines AIVC as a multi-scale AI-driven computational framework that integrates the gene–protein–pathway–cell biological hierarchies via multi-omics and multimodal data fusion, with the core goal of building a digital twin system for simulating cellular functions and behaviors. With the multi-scale modeling as the core characteristic, multi-omics and multimodal data serve as essential sources upon its construction. Following this particular scope, AI methodology regarding the four involved scales (i.e. gene, protein, pathway, and cell) is included, and the methods covering multiple scales are particularly emphasized. In relation with this, available metrics for evaluating AIVC are sorted out in a scale-dependent way. Meanwhile, since each scale corresponds to biological data of multiple modalities, we consider their fusion as a necessary step that should be introduced, but we do not elaborate on this in detail. AIVC is still in its early stage and evolves rapidly, the definition of AIVC here aims to

Table 1 Summary of models related to AIVC

Model	Training data (Input)	Key contributions and features	Output	Coverage scope	Data regime	Strengths	Limitations	Benchmark tasks	Model type	Ref
JCVI-syn3A	Synthetic minimal cell genome, biochemical reaction rules	Comprehensive simulation of JCVI-syn3A cell metabolism, gene expression, and growth processes	Simulation results of cell metabolism, gene expression, and growth processes	Whole-cell	Multi-modal (Sequence + Rules)	First comprehensive simulation of a minimal synthetic cell; covers full metabolic processes	Limited generalizability; only applicable to synthetic cells	Cell growth simulation, metabolic pathway reconstruction	Mechanistic	[24]
BoolNet	Small gene networks (10–100 genes)	Pioneered Boolean network modeling for cell state transitions, simulating gene switch behaviors	Gene switch state transitions, cell fate predictions	Gene	Single-modal (Gene network data)	Pioneered Boolean network modeling; high computational efficiency	Restricted to small-scale gene networks; lacks dynamic details	Gene regulatory network simulation, cell state transition prediction	Mechanistic	[32]
CellOrganizer	2D/3D microscopy images (tens of thousands)	Generated probabilistic models of subcellular structures from images, reconstructing morphological diversity	Probabilistic models of subcellular structures, morphological diversity reconstruction results	Cell (Subcellular structure)	Single-modal (Image data)	Accurately reproduces subcellular morphology; supports spatial modeling	Focuses solely on cell morphology; no molecular mechanism integration	Subcellular structure reconstruction, cell morphology classification	Hybrid	[33]
BioNetGen	Signaling pathway rules (100+ reaction rules)	Rule-based biochemical network modeling, simulating protein interaction cascades	Protein interaction cascade simulations, pathway activation results	Pathway	Single-modal (Rule data)	Enables complex signaling pathway simulation; high rule interpretability	Relies on manual rule definition; poor scalability	Signaling pathway activation simulation, protein interaction prediction	Mechanistic	[34]
iSEEEK	Over 10 million cells	Gene ranking-based model trained by predicting gene rankings	Gene expression rankings, cell type representations	Gene-Cell	Single-modal (Transcriptomic data)	Excels at large-scale data integration; accurate gene ranking prediction	Relies solely on transcriptomic data; lacks structural information	Cell type annotation, gene expression association analysis	Data-driven	[35]
tGPT	22.3 million cells (human and mouse)	Autoregressive modeling of gene rankings, processing gene sequences sorted by expression levels	Autoregressive prediction results of gene rankings	Gene-Cell	Single-modal (Transcriptomic data)	Supports cross-species prediction; strong large-scale data processing capability	Focuses only on gene rankings; ignores molecular interactions	Cross-species gene expression prediction, cell state inference	Data-driven	[36]
Geneformer	30 million cells	Predicted gene positions in cellular contexts, generating ranking-based gene embeddings	Gene position representations in cellular context, gene embedding vectors	Gene-Cell	Single-modal (Transcriptomic data)	Captures gene contextual associations; high-quality gene embeddings	Lacks cross-scale information transmission mechanisms	Gene regulatory network reconstruction, cell state transition prediction	Data-driven	[37]

(continued)

Table 1 Continued.

Model	Training data (Input)	Key contributions and features	Output	Coverage scope	Data regime	Strengths	Limitations	Benchmark tasks	Model type	Ref
scBert	Millions of cells (human)	Discretized gene expression values into "bins," transforming continuous prediction into classification	Cell type classification results, discretized gene expression representations	Gene-Cell	Single-modal (Transcriptomic data)	Strong noise resistance; high cell type annotation accuracy	Discretizes continuous expression values; loses partial information	Cell type annotation, gene expression pattern recognition	Data-driven	[38]
scGPT	Over 33 million cells (human)	Segmented gene expression values and used self-attention masking for autoregressive prediction	Masked gene expression recovery results, cell state representations	Gene-Cell	Single-modal (Transcriptomic data)	Enables fine-grained gene expression prediction; efficient self-attention mechanism	Covers only transcriptomic scale; no protein/pathway integration	Gene expression recovery, cell subtype identification	Data-driven	[39]
UCE	Over 36 million cells	Integrated genetic data using protein language models, predicting gene expression via self-supervised learning	Universal cell embeddings, gene expression prediction results	Gene-Protein-Cell	Multi-modal (Transcriptomic + Sequence)	Integrates genetic and protein data; strong generalizability	Lacks pathway-level regulatory mechanism modeling	Cross-cell type prediction, gene function annotation	Hybrid	[40]
scFoundation	~50 million cells (human)	Used masked autoencoder (MAE) to directly predict raw gene expression values	Cell state embeddings, gene regulatory network mapping results	Gene-Cell	Single-modal (Transcriptomic data)	High scalability for large-scale data; accurate gene regulatory network mapping	Lacks multi-modal data integration	Cell type annotation, drug response prediction	Data-driven	[25]
GET	213 human, fetal, and adult cell types	Enabled regulatory inference across diverse cell types, revealing universal and cell type-specific TF interaction networks	Gene expression levels across 213 cell types, regulatory element localization results	Gene-Protein	Multi-modal (Chromatin + Sequence)	High cell type-specific prediction accuracy; precise regulatory element identification	Limited to gene expression prediction tasks	Regulatory element localization, gene expression regulation prediction	Data-driven	[26]
GeneCompass	~50 million (human) + 50 million (mouse)	Incorporated four biological priors to enhance understanding of gene regulatory mechanisms	Gene regulatory mechanism representations, cross-species gene association results	Gene-Protein-Cell	Multi-modal (Transcriptomic + Species information)	Incorporates 4 biological priors; strong cross-species prediction capability	Weak pathway-scale simulation capability	Cross-species gene regulatory prediction, cell fate inference	Hybrid	[41]
CellFM	~100 million cells (human)	Adopted ERetNet architecture, excelling in multi-task prediction with the largest single-species dataset to date	Cell phenotypes, gene network reconstruction results	Gene-Cell	Single-modal (Transcriptomic data)	Largest single-species dataset; strong resistance to noise and batch effects	Weak pathway-scale simulation and cross-scale coupling capabilities	Cell state transition prediction, gene network reconstruction	Data-driven	[42]

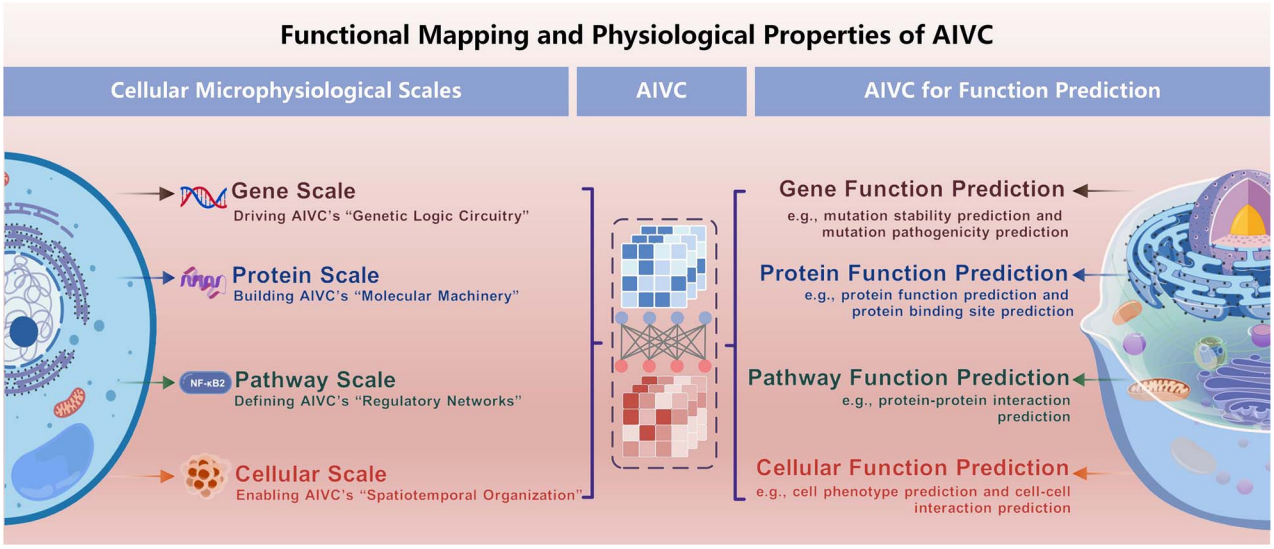


Fig. 1. Functional characteristics of AIVC.

offer a unified and practical framework with a particular focus on the multi-scale modeling, rather than a rigid or immutable concept.

The inherent complexity of biological systems, exemplified by the nonlinear dynamics where subtle perturbations at critical nodes can cascade through feedback loops, poses a significant challenge to the construction of AIVC. This is exemplified by KRAS hotspot mutations (e.g. G12, G13, and Q61), which disrupt GTP/GDP cycling and lock the protein in a constitutively active GTP-bound state. This, in turn, abnormally activates downstream oncogenic pathways such as MAPK and PI3K. Its smooth protein surface and strong nucleotide affinity make targeted therapy difficult [45]. In addition, covalent inhibitors developed for the KRAS G12C mutant block the signal by binding to the allosteric pocket of the GDP-bound protein. Although breakthroughs have been achieved in lung cancer, differences in tissue-specific efficacy and secondary mutations still limit their clinical application [46].

Existing biological research methods struggle to achieve the systematic integration of cross-scale cell behaviors. Traditional molecular dynamics simulations can resolve protein conformational changes with atomic-level precision. However, they incur extremely high computational costs and are generally only applicable to the nanosecond scale and local structures [47–49]. On the other hand, omics technologies can provide global data at the cellular or tissue level but have difficulty capturing the real-time dynamic characteristics of molecular interactions. For example, it is hard to determine how protein conformational changes affect the spatiotemporal distribution of signal transduction networks [50–52]. This gap between the microscopic and macroscopic levels means that many biological processes, such as the cascading effects of oncogenic mutations and the allosteric regulation of drug targets, cannot be fully analyzed.

AIVC technology connects the divide between different scales by constructing a multi-scale computational framework. Its core architecture relies on three key modules: multi-scale representation descriptors, adaptive representation encoders, and dynamic function predictor. Among them, the multi-scale representation descriptors are responsible for digitally representing and modeling entities at different biological levels to ensure a continuous mapping from genes to cell behaviors, such as gene sequence representation, protein structure representation, and interaction networks. These descriptors not only need to independently characterize the features of each scale but also establish cross-scale association rules. The adaptive feature encoder is the learning core of the architecture. This module achieves the fusion of multi-source data and feature capture through a deep neural network. The dynamic function debugger is responsible for the selection of specific task classifiers and the iterative optimization of the model, enabling function prediction at different microscopic scales and prediction of the spatiotemporal evolution of cell behaviors.

Concept of cross-scale

As a complex biological system, the functions and behaviors of cells are regulated by multi-scale and multi-dimensional molecular mechanisms. To systematically analyze the complexity of cells, the construction of AIVC is mainly divided into four functional scales: the gene scale, the protein scale, the signaling pathway scale, and the cell scale. These four scales do not exist in isolation but jointly drive the complexity and diversity of cell behaviors through multi-scale interactions. They are detailed as follows. First, the gene scale defines the foundational genetic framework, encompassing DNA sequence information, transcriptional regulation, and epigenetic modifications. Next, the protein scale describes the molecular machinery, including protein synthesis, modification, and interaction networks. The signaling pathway scale forms the regulatory bridge, translating signals into functional responses via molecular cascades and determining cellular adaptability. Finally, the cell scale integrates all underlying information, manifesting the integrated spatial, interactive, and behavioral outcomes of the cell.

Interactions across scales form the core framework of virtual cell research, shaping a full-chain regulatory network from molecular mechanisms to cellular behaviors. Specifically, AIVC incorporates the following adjacent-scale couplings:

(1) **Gene–protein coupling**. Gene expression levels directly dictate protein synthesis quantity. Epigenetic modifications influence transcriptional efficiency by altering chromatin accessibility. For example, mutations in the tumor–suppressor gene BRCA1 lead to the loss of its protein function, thereby affecting the DNA repair mechanism [53].

(2) **Protein–pathway coupling**. The modification state of proteins directly regulates the activity of signaling pathways. For instance, the phosphorylation of AKT activates the PI3K/AKT pathway, promoting cell survival [54]. The activation of the Caspase family of proteins initiates the apoptosis signaling pathway [55].

(3) **Pathway–cell coupling**. The dynamic changes in signaling pathways ultimately determine cell behaviors and fates. For example, the activation of the TGF-β signaling pathway induces epithelial–mesenchymal transition, enhancing the migration ability of tumor cells [56]. The activation of the p53 pathway triggers cell-cycle arrest or apoptosis [57].

The construction of AIVC should conform to the above scale divisions. A multi-level dynamic coupling architecture capable of integrating multi-functional predictions should be constructed for each corresponding scale. The required data and construction process will be described in the subsequent sections.

Artificial intelligence in cross-scale virtual cells

Datasets

For multiscale AIVC modeling, the foundation lies in exhaustive, high-quality datasets that encompass the four core biological scales: genetic, protein, pathway, and cellular. A summary of relevant datasets is shown in Table 2.

Gene-scale datasets focus on DNA sequences and mutation–disease associations and provide foundational information on genetics for AIVCs. As the two most widely used core databases in global pathogenic mutation research, ClinVar contains over 6 million records of gene variant–disease associations, with >90% of entries accompanied by clear clinical assertion criteria. It also classifies variant types such as germline and somatic, covering multiple disease categories including tumors and rare diseases [58]. HGMD, on the other hand, has accumulated and updated ∼300 000 manually rigorously curated pathogenic variant records [59]. Together, these two databases serve as the gold-standard data sources for the mutation pathogenicity prediction task of AIVCs. ProThermDB [60] and FireProtDB [61] specialize in protein mutation stability data, containing ∼31 500 and 13 274 experimental data points on the thermal stability of protein mutants, respectively. In addition, UK Biobank [63] and GWAS Catalog [64] are core public resources for macro genotype–phenotype association research, with applications spanning complex disease mechanisms, drug response prediction, and other fields.

Protein-scale datasets center on protein 3D structures, sequences, and functions—providing critical support for AIVC molecular modeling. PDB has curated over 200 000 experimentally resolved 3D structures, with determinations spanning cryo-electron microscopy and other techniques [67]. These structures cover >90% of human proteins. It stands as a fundamental data resource for structural biology and drug design. On the other hand, UniProt integrates over 500 000 entries of protein sequences, functional annotations, and literature-associated data, enabling full-chain information retrieval spanning sequence, structure, and function [68]. The synergy between PDB and UniProt is adopted by nearly all protein research teams worldwide. SKEMPI 2.0 focuses on the impact of mutations on protein–protein interactions (PPIs), containing >6000 experimental data points on the effects of mutations on PPI binding affinity [62]. PDBbind [73] and MISATO [74] cater to research on drug–target binding: the former includes 30 000 entries of protein–ligand complex structures and their binding affinity data, while the latter supplements PDBbind with dynamic trajectory data of protein–ligand interactions.

Pathway-scale datasets focus on intermolecular interactions and signal transduction, serving as a bridge for the regulatory networks modeling of AIVC. STRING [69] and BioGRID [70] collaborate to construct high-confidence pathway networks. STRING contains 27.541 billion protein–protein interaction records, which are categorized into four levels based on confidence scores (≥0.900, ≥0.700, ≥0.400, and full confidence), corresponding to 332 million, 977 million, 3.885 billion, and 27.541 billion records, respectively. BioGRID covers 2 898 895 protein and genetic interactions, 31 540 chemical interactions, and 1 128 339 post-translational modification records. Additionally, KIBA, Davis, and Human form a protein–ligand affinity dataset system: KIBA includes over 110 000 affinity scores integrated from multiple indicators, Davis contains >30 000 dissociation constant data points for kinase inhibitors, and Human provides over 3000 binary labels for compound–protein interactions [75–77].

Cell-scale datasets focus on single-cell transcriptomes, imaging, and cell–cell interactions. Their large-scale and multi-dimensional nature provides a macroscopic perspective for the spatiotemporal organization modeling of AIVCs. The HCA [78] and Tabula Sapiens [79] have applications covering cell type annotation, tissue development mechanism analysis, and other fields. The HCA contains a single-cell transcriptomic atlas of over 60 million cells, spanning multiple human organs. Tabula Sapiens focuses on multi-organ single-cell data, including gene expression profiles of >5 million cells, which enable the classification of cell subtypes and functional states. It also reveals transcription factor expression, aging effects, and sex-specific characteristics across cell types from 28 human organs and tissues. The IDR [82] and CellPhoneDB [83] supplement spatial and interaction information. The IDR archives over 8 million high-resolution cellular microscopy images and ∼6 million tissue images. CellPhoneDB includes 2911 manually curated cell–cell interaction records, among which 1882 are mediated exclusively by proteins and 1029 by non-protein ligands. Approximately 50% of the curated interactions involve at least one heteromeric complex. Additionally, in June 2025, Hani Goodarzi, together with >20 researchers in the field, jointly launched the Virtual Cell Challenge [31]. To support the training and evaluation of models for the challenge as well as meet the needs of virtual cell modeling, Tahoe-100M [80] and scBaseCount [81] have been designated as core benchmark resources and incorporated into the "Arc Virtual Cell Atlas" developed by the Arc Institute. Among them, Tahoe-100M is the world's largest-scale single-cell perturbation atlas, generated by Tahoe via its independently developed high-throughput Mosaic platform. It contains over 100 million single-cell transcriptomic profiles derived from ∼60 000 drug perturbation experiments, systematically mapping the responses of 50 cancer cell lines to >1100 small-molecule drugs. scBaseCount is the first continuously updated single-cell RNA-seq database constructed under the drive of AI. It leverages a hierarchical agent to automatically mine and conduct standardized preprocessing on Sequence Read Archive data, covering 21 organisms and 72 tissues, containing over 230 million cells, with its scale still expanding.

AIVC modeling methods

AIVC is envisioned as a data-driven AI architecture capable of cross-scale cell behavior simulation. By mapping the biological

Table 2 Summary of datasets related to AIVC construction.

Datasets	Data content	Ref	Application scenarios
Gene scale			
ClinVar	Data on the association between gene mutations and diseases	[58]	Gene pathogenicity prediction
HGMD	Database of pathogenic mutations	[59]	
ProThermDB	Data on the thermal stability of protein mutations	[60]	Mutation stability prediction
FireProtDB	Experimental data on the stability of protein mutations	[61]	
SKEMPI 2.0	Data on the stability of mutations at the protein–protein interface	[62]	
UK Biobank	Genomic and phenotypic data	[63]	Gene–phenotype association analysis
GWAS Catalog	Published results of GWAS studies	[64]	
DisGeNET	Data on the association between genes and diseases	[65]	
GEO	Gene expression and phenotypic data	[66]	
Protein scale			
PDB	3D structure data of proteins resolved by experiments	[67]	Protein function prediction, Protein structure design
UniProt	Comprehensive database of protein sequences, functions, and literature annotations	[68]	
Pathway scale			
STRING	Protein–protein interactions	[69]	Protein–protein interaction prediction
BioGRID	Experimentally verified data on protein interactions and genetic interactions	[70]	
IntAct	Experimental data on molecular interactions (including complex structures)	[71]	
Yeast	Protein interaction data related to yeast cells	[72]	
PDBbind	Structures of protein–ligand complexes and their binding affinity data	[73]	Protein–ligand interaction prediction
MISATO	Static structures and dynamic trajectories of protein–ligands	[74]	
KIBA	Protein–ligand affinity data	[75]	
Davis	Protein–ligand affinity data	[76]	
Human	Protein–ligand affinity data	[77]	
Cell-scale			
HCA	Single-cell transcriptome atlas of human whole body cell types	[78]	Cell type annotation, Cellular state characterization, Spatiotemporal cellular network modeling
Tabula Sapiens	Single-cell transcriptome data of multiple organs	[79]	
Tahoe-100M	Small-molecule perturbation single-cell transcriptomic data	[80]	
scBaseCount	Single-cell RNA-seq database	[81]	
IDR	High-resolution cell microscopy imaging dataset	[82]	Cell imaging and morphological analysis
CellPhoneDB	Cell communication database	[83]	Cell–cell interaction

scales to the hierarchical representation space of deep learning and introducing biologically inspired regulatory elements, the following aims can be achieved. The first is the horizontal modality fusion that integrates sequence modality, structure modality, and dynamic modality within a same scale to achieve modality alignment and information completion. The next is the vertical scale penetration that establishes a cross-layer feature pyramid and construct specific flow-based representations according to scale characteristics. The final aim is to accomplish desired prediction tasks by using an extendable and dynamic architecture that incorporates submodules of different levels. The conceptual diagram of the overall architecture of AIVC modeling is shown in Fig. 2.

Taxonomy of AIVC-related models

This section systematically summarizes existing modeling methods related to AIVC, see Table 1 for details. For each model, we have systematically sorted and summarized it from key attributes, including input data, key contributions, output results, and model type, etc., clearly presenting the technical characteristics and application scenarios of each method. Overall, based on the model's modeling design, data modality, and covered biological hierarchies, existing AIVC-related methods can be systematically classified into three aspects: In the aspect of model type, based on the way modeling logic is combined with biological prior knowledge, methods are divided into three categories: mechanistic, data-driven, and hybrid. Mechanistic methods, such as JCVI-syn3A and BoolNet, are constructed based on explicit biochemical rules or physical laws, featuring strong interpretability. Data-driven methods, such as Geneformer and scGPT, rely on large-scale multimodal data to autonomously learn implicit patterns, making them suitable for complex system modeling. Hybrid methods, such as UCE and GeneCompass, integrate the advantages of biological priors and data learning, balancing interpretability, and

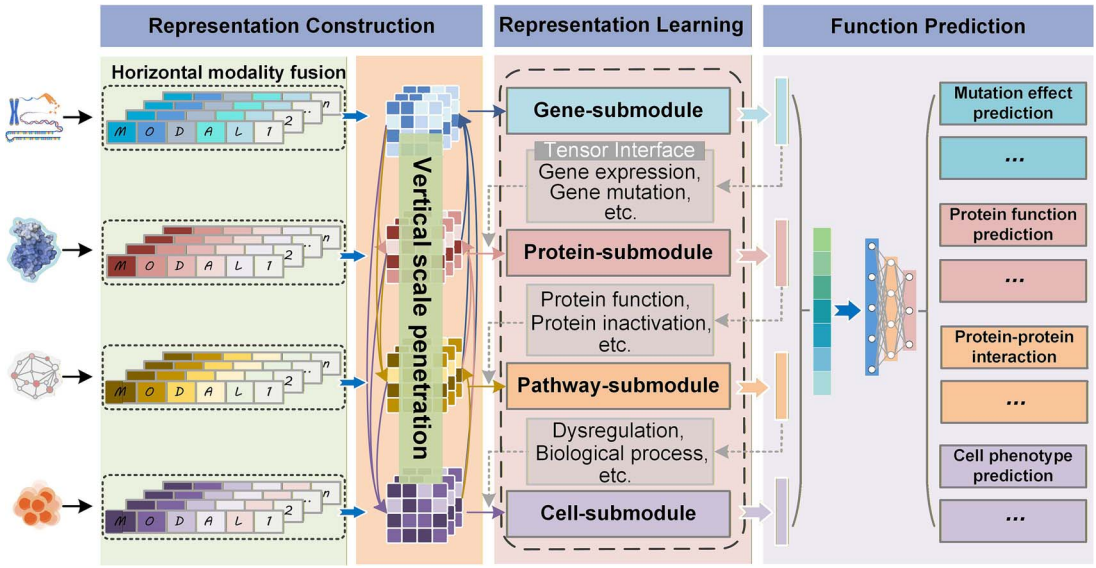


Fig. 2. Conceptual diagram of the AIVC technical architecture.

generalization ability. In the aspect of data modality, according to the source and type of input data, methods are divided into two categories: single-modal and multi-modal. single-modal methods, such as iSEEEK and CellFM, rely solely on a single type of data (e.g. transcriptomic data and image data) with a simple modeling process. Multi-modal methods, such as JCVI-syn3A and GET, integrate multiple complementary data including transcriptomic data, genomic data, and spatial coordinates, enabling more comprehensive characterization of the multi-dimensional characteristics of biological systems. In the aspect of spatial scale, according to the biological tissue hierarchies covered by the models, it covers different scales from genes to cells, including gene scale (e.g. BoolNet), gene–protein scale (e.g. GET), pathway scale (e.g. BioNetGen), cell/subcellular scale (e.g. CellOrganizer), gene–cell scale (e.g. scFoundation), gene–protein–cell scale (e.g. UCE), and whole-cell scale (e.g. JCVI-syn3A). Methods at each scale focus on the dynamic modeling needs of the corresponding hierarchy, jointly forming the multi-scale modeling system of AIVC.

Construction of cross-scale representations

This section provides an introduction to the multi-scale representation descriptors. In the AIVC framework, cross-scale representation construction centers on creating patterned inputs that link biological levels. These representations are deep abstractions and integrations of essential biological features. Their construction follows principles of hierarchical encoding and dynamic interaction, preserving feature independence at each scale while keeping cross-scale associations computationally tractable.

Gene-scale encoding converts DNA/RNA sequences into digital representations, primarily through label encoding, k-mer encoding, and one-hot encoding. These methods differ in how they preserve information, capture biological meaning, and adapt to downstream models. Among them, label encoding, as the simplest method, achieves encoding by converting the four bases (A, G, C, and T) into integers or ordinal values [84]. This scheme features fast operation speed and avoids the problem of dimensionality curse. K-mer encoding centers on substrings of length k (i.e. k-mers) extracted from genome sequences. For a sequence of length N, when the k value is specified by the user, the total number of k-mers that can be generated in the sequence is given by "N - k + 1" [85]. One-hot encoding represents individual bases or k-mers using binary vectors. For example, base A corresponds to [0, 0, 0, 1], and base G corresponds to [0, 0, 1, 0]. A DNA sequence of length N can be converted into a 4×N matrix through this encoding. When k-mers are used as tokens (e.g. k = 3), the sequence is first split into overlapping 3-base substrings, and then each k-mer is assigned a one-hot vector of length 4^3, forming a 2D matrix where rows correspond to k-mers and each column corresponds to a "word" in the dictionary [86]. The key difference between one-hot encoding and label encoding lies in that the former preserves the base order, while there is no specific order among categories in the latter. Meanwhile, ne-hot encoding can focus on retaining the precise characteristics of nucleotide motifs; this encoding method can also eliminate the implicit ordinal relationship between features in the input data, preserve positional information and the diversity of amino acids in the DNA sequence, thereby providing more biologically realistic input features for subsequent modeling. Almuhaideb *et al.* conducted a validation study on the performance of various gene sequence representation methods, including label encoding and k-mer one-hot encoding. The results showed that the combination of k-mer one-hot encoding and CNN-LSTM achieved the best performance [87].

Protein representations are commonly categorized by their information dimensions into 1D (sequence), 2D (topology), and 3D (structure) methods [88]. The 1D approach operates on the linear order of amino acids, using techniques like label or k-mer encoding to convert residue types into digital signals for rapid sequence analysis, though it captures only sequential information [89]. 2D representation builds on 1D methods to capture pairwise residue interactions. Amino acids are graph nodes, interactions are edges, and valid pairs are screened using Euclidean distance thresholds [90]. However, it only captures planar relationships, missing 3D spatial details. 3D representation, based on protein spatial conformation, breaks 1D/2D constraints [91]. It prioritizes 3D features over linear and planar data. Geometric equivariant graph neural networks are widely used to process sequences and 3D structures jointly—they capture microdetails and ensure pose-invariant representations, avoiding

spatial distortion. This provides high-fidelity data for protein function analysis and cross-scale studies [92–95].

The construction of signaling pathway-scale representations focuses on the interactions between entities. The realization of cellular functions stems from the signal interactions between functional molecules that form signaling pathways [96]. Modeling of signaling pathways can be regarded as a huge functional interaction network, where the nodes in the network represent functional molecular entities, and the edges between entities represent whether there is an interaction or the type of interaction between the two [97, 98]. The node representations in the network are inherited from the upstream protein/gene scale [99–101]. This representation method can predict the interaction potential between functional molecules and their response patterns under various perturbations, providing an important bridge for cross-scale dynamic interactions. For example, in the HiSIF-DTA framework proposed in our previous research, when constructing the signaling pathway network, the structural and functional features at the upstream protein scale are used as the core representation of network nodes. The weights of edges are employed to quantify interactions between proteins, and with the GCN model, this framework has successfully achieved high-precision prediction of drug–target interactions, providing a typical case for bridging the cross-scale analysis between protein molecular interactions and cellular function regulation [97].

Cell-scale representation stands as the most complex, requiring holistic integration of multi-modal data and multi-scale features. Representations at this scale must fuse features from underlying microscopic scales while accounting for macroscopic characteristics, including organelle spatial organization, mechanical force conduction, and metabolic flux distribution [102, 103]. Building on topological modeling approaches, the organelle interaction network can be modeled as a dynamic graph structure. Here, nodes incorporate functional state features of organelles, including mitochondria and endoplasmic reticulum, while edges encode their physical connections and material exchange relationships [104, 105]. Meanwhile, it is also necessary to quantify the spatial distribution patterns of subcellular structures and the mechanical properties of cells, such as the conduction of mechanical signals like membrane tension and adhesion [106]. This multi-modal fusion representation method can not only reflect the steady-state characteristics of cells but also predict the dynamic responses of cells to external stimuli through temporal representation, ultimately achieving a complete closed-loop modeling from molecular events to cell behaviors.

Effective cross-scale representation interaction is crucial for ensuring coordinated and consistent information flow across different biological levels. To achieve this, systematic interaction mechanisms need to be established both vertically and horizontally. For vertical interaction, a feature pyramid structure can be adopted. Low-scale features are transmitted upwards through attention guidance, while high-scale features are fed back downwards through inverse operations, thus realizing hierarchical information transfer from genes to proteins, from proteins to pathways, and from pathways to cells. Horizontal interaction relies on a dynamic gating fusion mechanism, which automatically adjusts the contribution weights of each modality according to the specific task requirements to ensure the balance and optimization of cross-scale information. Through this systematic cross-scale representation construction and interaction mechanism, the AIVC framework can start from gene mutations, predict signal pathway abnormalities through protein conformation changes, and ultimately deduce the full-chain mechanism of changes in cell migration behavior.

Design of functional submodules

This section provides an introduction to the adaptive representation encoders. The functional submodules of AIVC are the core modules of the overall AIVC framework. Their design should follow the construction concept of scale decoupling, function specialization, and system collaboration. Through the method of hierarchical modeling, joint learning across modules is achieved relying on standardized classifiers. In the architecture design, each submodule deeply integrates biological mechanisms and deep learning paradigms, forming the full-chain simulation ability from molecular regulation to cell behaviors.

(1) **Gene function submodule**. As the underlying driving unit for cross-scale modeling, this module focuses on the conversion of DNA sequence information into functional semantics. Regarding the long-range dependency characteristics of genomic sequences (such as long-distance interactions between enhancers and promoters), the Transformer-XL architecture can be adopted to break through the context length limitation of traditional Transformers. Through the segmented recurrence mechanism, continuous modeling of millions of bases can be achieved [107, 108]. Another example is the HiCDiffusion model developed by Chiliński *et al.* [108], which is a diffusion model extended based on Transformers. By integrating chromatin Hi-C data, it successfully achieves the prediction of 3D chromatin interaction patterns from DNA sequences, further verifying the advantages of this architecture in modeling long-range dependencies [108]. In addition, CRISPR screening data can be introduced as a supervisory signal, and gene editing perturbations and expression level changes can be mapped to a differentiable attention weight matrix, so as to analyze the spatial organization rules of cis-regulatory elements [109, 110]. For example, the CRISMER model proposed by Emtiaj *et al.* adopts a hybrid deep learning structure combining a Transformer architecture and multi-branch convolutional neural networks. The former is responsible for capturing long-range dependencies in genomic sequences, while the latter extracts k-mer features. Through integrated gradient analysis, it focuses on PAM-proximal regions and mismatch patterns, efficiently deciphering the synergistic interaction rules between promoters and enhancers, and providing reliable support for the prediction of regulatory element spatial organization at the genomic level [111]. This module should realize the prediction of relevant tasks such as the pathogenicity and stability of gene mutations and the identification of related oncogenes [112–114].

(2) **Protein function submodule**. To break through the excessive dependence of traditional methods on the static structure of proteins, the construction of this module can adopt a “structure-dynamics” dual-stream modeling system. In the structure encoding branch, a residue level graph structure can be constructed based on the 3D conformation predicted by AlphaFold, and a 3D graph convolutional network can be used to extract geometrically sensitive features [115–117]. For instance, Yu *et al.* [118] proposed Protein-SE(3), a multi-view generative model based on SE(3). The core SE(3) mathematical abstraction of this framework supports algorithm devel-

opment independent of explicit protein structures, which in turn strengthens the capacity of geometric equivariant models to capture proteins' 3D structural features. In addition, DeepFRI is a protein function annotation model based on graph convolutional networks (GCNs), adopting a two-stage architecture. It takes protein 3D structures as input and identifies key residues via grad-CAM, endowing structural encoding with function-oriented feature extraction capability [119]. Furthermore, MIF2GO achieves protein function annotation through multimodal fusion, integrating six biological modalities including structural information to generate unified and robust protein representations. It not only possesses cross-species generalization ability but also provides an effective paradigm for the deep integration of structural and multi-dimensional biological information [95]. In the dynamics branch, the model is trained on trajectory data from molecular dynamics simulations to capture the characteristics of conformational changes [120]. Additionally, implicit dynamic representations of proteins can also be constructed via models, with a typical example being the DynPred model proposed by Shah *et al.* [121]. This model takes sequence and structural data of over 200 000 proteins as input, processes features through techniques such as position-specific scoring matrix (PSSM) and Unirep embedding, integrates a GCN, and finally outputs the root-mean-square fluctuation (RMSF) values of alpha carbons (Cα), thereby achieving accurate prediction of protein implicit dynamics. Regarding the issue of representation fusion between the two branches, methods such as contrastive learning can be used to align the representations in the latent space and establish association constraints between static and dynamic representations. This module should implement relevant tasks such as protein function prediction, protein design, and binding site prediction [122–124].

(3) **Pathway function submodule**. Facing the highly non-linear features of biological pathways, this module should integrate graph neural networks and neural ordinary differential equations to deeply capture the characteristics of signal pathways. Firstly, a molecular interaction heterogeneous graph can be constructed based on pathway databases, and multiple types of relationships between nodes (such as activation, inhibition, and metabolic transformation) can be encoded [125]. Then, relevant deep learning networks for graph learning can be used to capture the implicit pathway information among molecules [126, 127]. For instance, Li *et al.* proposed the heterogeneous graph attention network for drug–target interaction prediction [128]. This model enhances heterogeneous graph structure learning from intra-layer and inter-layer. On one hand, it constructs an enhanced graph attention diffusion layer (EGADL) to efficiently connect non-directly connected nodes, enable information transmission from important nodes beyond multiple hops, and address the limitation that traditional graph neural networks can only aggregate information from directly connected nodes. On the other hand, it expands the receptive field from the inter-layer dimension by stacking multiple EGADLs to capture long-range dependencies in biological heterogeneous graphs. In addition, to enhance the interpretability of the model, an attention-based pathway activity scoring mechanism can be introduced to quantify the regulatory contributions of key nodes to the pathway output. This module should implement core functions, such as protein–protein interaction prediction, pathway activity prediction, and drug–target combination screening [94, 97].

(4) **Cellular function submodule**. To achieve cross-scale emergent simulation from microscopic molecular events to macroscopic phenotypes, this module can draw on the technical paradigms of existing advanced cell large models. Models like scFoundation and CellFM have established a solid basis for analyzing complex cellular functions at scale. scFoundation, developed by Tsinghua University researchers, is a 100-million-parameter foundational model trained on 50 million single-cell gene expression profiles [25]. It uses a Transformer-based asymmetric xTrimoGene backbone to process roughly 20 000 genes in parallel. A major innovation is its read-depth-aware pretraining task, which mitigates biases from varying sequencing depths, enabling the model to map gene regulatory networks and support downstream tasks such as drug response prediction. CellFM, the first global transcriptomics foundation model, was trained on 100 million human cells using an enhanced RetNet architecture with 800 million parameters [42]. It effectively handles single-cell data challenges like noise and batch effects, achieving strong performance in cell type classification and gene-network reconstruction. Building on these advances, the module abstracts organelles—such as mitochondria and endoplasmic reticulum—as autonomous agents. Each agent learns specific functions (e.g. energy metabolism, protein secretion) via deep networks, while a Transformer-based coordinator models inter-agent interactions and dependencies through multi-head attention, thereby predicting cellular adaptive behaviors under microenvironmental stimuli [129]. This extends gene-level modeling to the organelle-interaction level, enabling cellular modeling, state-transition simulation, and perturbation-response prediction [130–132].

In addition, each submodule should achieve cross-scale information transfer through standardized tensor interfaces. The output representation of the gene module serves as the initial condition for the protein module. The output representation of the protein module is mapped to the node representation of the pathway module, and the dynamic state of the pathway constitutes the observation space of the agents in the cell module.

Grouping of the dynamic prediction module

This section provides an introduction to the dynamic function predictor, a core component of AIVC that undertakes the key tasks of cross-scale temporal synchronization and multi-level functional regulation.

(1) **Temporal scale synchronization.** AIVC modeling needs to integrate four biological hierarchies. However, there are significant differences in the temporal dynamic characteristics of each hierarchy, thus forming the challenge of temporal mismatch between various biological scales. To address this challenge, it is necessary for AIVC modeling to draw on existing temporal matching algorithms. In this aspect, the temporal mismatch problem of AIVC could benefit from both existing mathematical solutions and recent data-driven time-series models.

First, in terms of classic mathematical solutions, optimal transport (OT) [133] is a practical tool for time-series data integration. Specific computational methods have been developed based on

different OT formulations to address various biological scenarios. For instance, methods such as scEGOT [134] adopt static OT to achieve basic matching of cell states between discrete-time points. In contrast, models like TrajectoryNet [135] adopt dynamical OT, which recovers the temporal evolution of cell trajectories by modeling a continuous flow field that connects distributions across time. Furthermore, to account for scenarios where the total number of cells changes, methods such as TIGON [130] adopt an unbalanced OT framework, which introduces a growth/death term (source–sink term) to explicitly model mass imbalances arising from cellular processes like proliferation and apoptosis. These together provide a mathematical foundation for AIVC to handle dynamic changes at the cell scale. In another case, stochastic differential equations and Fokker–Planck equations [136] are used to characterize random fluctuations in biological processes, while the Schrödinger Bridge framework integrates deterministic regulation and random perturbations to accurately capture probabilistic trajectories in cell fate decisions [137, 138].

In terms of data-driven time-series modeling, representative approaches such as MTPNet [139], MSFformer [140], and HiMTM [141] offer direct technical references for cross-scale temporal synchronization. For instance, the multi-scale decomposition paradigm proposed by MTPNet can split complex biological processes into trend components and transient components, adapting to modeling needs of different temporal granularities respectively. MSFformer's pyramid-shaped feature convolution method achieves efficient aggregation of fine-grained to coarse-grained features, providing an engineering approach for converting small-scale fast events to large-scale slow responses. Additionally, HiMTM's hierarchical masked reconstruction and cross-scale attention fine-tuning strategy demonstrate advantages in strengthening the correlation of cross-scale temporal nodes and accurately capturing temporal dependencies, which can provide key technical support for the association and fusion of temporal features across AIVC's hierarchies.

The above time-series models can be integrated into the dynamic function predictor of AIVC to enable its ability of temporal synchronization across scales. Here, we propose several potential ways to achieve this goal. The first is the adoption of the multi-scale decomposition concept of MTPNet to establish a cross-scale temporal correlation framework from the rapid conformational changes of proteins (transient events) to slow gene expression (long-term trends). It connects biological processes with distinct occurrence rates via adaptive time steps. The second is the utilization of the pyramid-shaped feature convolution structure of MSFformer to construct a learnable cross-scale information aggregation interface. It achieves effective information transmission and temporal synchronization between fast and slow scales through cross-scale feature fusion and resolution conversion between slow- and fast-rate scales. The third is the introduction of the cross-scale masked attention mechanism of HiMTM to conduct cross-scale attention focusing and selective mask verification on key biological events. This thereby calibrates and enhances the accuracy of temporal causal alignment of events across different scales to ensure the biological rationality of dynamic simulations. These technologies and optimization approaches do not require reconstructing the core architecture of the dynamic function predictor; instead, through the reasonable integration of existing methods, it can meet the conditions for cross-scale temporal synchronization, thereby enhancing the dynamic simulation accuracy of AIVC.

(2) **Multi-component dynamic regulation mechanism.** AIVC's multi-component dynamic regulation mechanism leverages a bioinspired programmable architecture to construct a closed-loop system. This system comprises a fine-tuning component library, a dynamic classification controller, and a meta-learning optimizer. The bioinspired design of the component library forms the foundation of this mechanism. The fine-tuning component library contains two types of core regulatory units. The first type is the computational basic operators. They achieve dimensional alignment of cross-scale features through the linear projection layer, dynamically adjust the intensity of the information flow in combination with the gated activation function, and complete task driven feature filtering using the feature selection gate [142]. The second type is the bioinspired operators. They realize mechanism embedding by simulating biological regulation. For example, the Sigmoid function is used to quantify the dose synergistic effect of transcription factors, the Softmax weight distribution is employed to simulate the competitive inhibition relationship of molecular binding sites, or the cyclic skip connection is used to reproduce the steady-state regulation function of the biological negative feedback loop [143, 144].

At the level of collaborative optimization of the dynamic regulation strategy, the system achieves intelligent adaptation through a three-layer paradigm.(i) The representational perception routing mechanism relies on a lightweight controller network to dynamically analyze the multimodal characteristics of the input data (such as protein sequence representation and protein structure representation), and autonomously generate the optimal representational combination path. (ii) The meta learning driven "model agnostic" framework initializes the parameter space by pretraining fine-tuning components. After pretraining the general patterns of protein interaction prediction with large-scale data, it only requires a small amount of fine-tuning to be transferred to the drug resistance analysis task of specific cancers, significantly improving the cross-task transfer efficiency. (iii) The task aware classification mechanism, on the one hand, designs hot swappable independent classification heads for heterogeneous targets such as classification tasks, regression tasks, and prediction tasks at different functional scales. It dynamically assigns gradient back propagation weights through the gating network to avoid conflicts in the optimization direction. On the other hand, it establishes cross module knowledge sharing channels, converting the intermediate parameters of each module into soft labels to promote collaborative optimization among modules.

Through the above flexible architecture, AIVC can autonomously reconstruct the computational path according to the task type and feature distribution. Just like real cells responding to changes in the microenvironment, it achieves a dynamic balance between biological rationality and computational efficiency in virtual cell modeling, providing an engineering solution close to the biological essence for cross-scale life system simulation.

Evaluation criteria and validation framework

The development of AIVC models necessitates a robust, multi-scale evaluation framework to rigorously assess their predictive accuracy, biological relevance, and practical utility. Since the core goal of AIVC is to integrate multi-level biological entities from genes to entire cells, covering multi-dimensional biological processes, relying solely on evaluation metrics for a single level or specific task is insufficient to fully characterize its performance. Therefore, a systematic validation strategy must be highly compatible with the multi-scale modeling architecture of AIVC, enabling comprehensive quantification of modeling accuracy across different biological levels. This section will systematically sort out the core evaluation metrics applicable to AIVC and further explore other key dimensions that need to be emphasized during the evaluation process, providing practical references for the performance validation and method optimization of AIVC.

(1) **Gene-scale evaluation metrics.** For the gene scale, Virtual Cell Challenge [31] is a representative and trending evaluation benchmark. It offers three metrics with a focus on the prediction of gene expression responses to genetic perturbations, which are namely differential expression score (DES), perturbation discrimination score (PDS), MAE.

- *DES*. This metric assesses the accuracy of predicting differentially expressed genes after perturbations. It is defined as the intersection between the predicted and true sets:

$$\mathrm{DES}_k = \frac{G_{k,\mathrm{pred}} \cap G_{k,\mathrm{true}}}{n_{k,\mathrm{true}}} \tag{1}$$

where k indicates the perturbation, $G_{k,\mathrm{true}}$ is the ground truth gene sets determined in experiments, $G_{k,\mathrm{pred}}$ is the predicted gene sets.

- *PDS*. This metric measures the model's capacity to distinguish between different perturbation types. It is defined as:

$$\mathrm{PDS}_p = 1 - \frac{\mathbf{argind}\left\{d_{pt}\right\}_{t=p} - 1}{N} \tag{2}$$

where $\mathbf{argind}\{d_{pt}\}_{t=p}$ is the true perturbation in the ordered list, d_{pt} is the Manhattan distance between the predicted and real perturbation delta. An idealized model with minimal distance will get a PDS score of 1.

- *MAE* (MAE). This metric evaluates the prediction accuracy of the full gene expression profile, ensuring the overall reliability of the model beyond just key differential genes. It is defined as:

$$\mathrm{MAE}_k = \frac{1}{G}\sum_{g=1}^{G}\left|\hat{y}_{kg} - y_{kg}\right| \tag{3}$$

where g denotes gene index. It is calculated across all G genes.

(2) **Protein-scale evaluation metrics.** For the protein scale, based on previous protein-related research and recommendations from the Critical Assessment of Functional Annotation (CAFA) challenge [145], three widely adopted metrics are adopted, namely area under the precision-recall curve (AUPR), the maximum F1-score (F-max), and Pearson correlation coefficient (PCC).

- *AUPR*. This metric evaluates the ability to balance precision and recall in imbalanced protein-related tasks such as protein function prediction. It is defined as:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{AUPR} = \sum_{t=1}^{T}\left(\mathrm{Recall}(t) - \mathrm{Recall}(t-1)\right)\cdot \mathrm{Precision}(t) \tag{6}$$

where TP, FP, and FN denote the sample numbers of true positive, false positive, and false negative, respectively, and t denotes the classification threshold.

- *F1-max*. This metric evaluates the optimal comprehensive performance of binary classification tasks such as protein function prediction. It is defined as:

$$\text{F1-max} = \max\left\{\frac{2\cdot \mathrm{Precision}\cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}\right\} \tag{7}$$

- *PCC*. This metric evaluates the linear consistency between predicted continuous values and experimental measurements. It is adopted in the task of predicting mutation effects on protein stability. It is defined as:

$$\mathrm{PCC} = \frac{\mathrm{cov}(\mathrm{y}, \hat{\mathrm{y}})}{\sigma(\mathrm{y})\,\sigma(\hat{\mathrm{y}})} \tag{8}$$

where *cov* and σ represent the covariance and standard deviation, respectively; y and $\hat{\mathbf{y}}$ refer to the true values and predicted values.

(3) **Pathway-scale evaluation metrics.** For the pathway scale, drawing on previous pathway-related research, e.g. protein–protein interaction (PPI) prediction, DNA–protein interaction (DPI) prediction, three key metrics—Matthews correlation coefficient (MCC), F1-Score, and Spearman rank correlation (Spearman)—are determined to assess model performance.

- *MCC*. This metric is tailored for extremely imbalanced pathway-related binary classification tasks, comprehensively balancing precision, recall, and the prediction accuracy of both positive and negative samples. It is defined as:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{9}$$

where TN denotes the number of true negative samples.

- *F1-Score*. This metric integrates precision and recall through harmonic mean, evaluating the robust performance of

pathway-related binary classification tasks and avoiding bias from a single indicator. It is defined as:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

- *Spearman*. This metric assesses the rank correlation consistency between predicted continuous values and experimental measurements, with no strict requirements on data distribution, making it suitable for predicting non-normally distributed pathway-related continuous variables. It is defined as:

$$\text{Spearman} = 1 - \frac{6 \sum_{i=1}^{n} \left(R(y_i) - R(\hat{y}_i)\right)^2}{n\left(n^2 - 1\right)} \tag{11}$$

where $R(y_i)$ denotes the rank of the ith sample value in y, and n denotes the sample size.

(4) **Cell-scale evaluation metrics.** For the cell-scale, drawing on previous cell-related research (e.g. cell type classification, cell fate trajectory inference, cell state prediction, etc.), three key metrics—Adjusted Rand Index (ARI), Silhouette Coefficient, and Macro F1 Score—are determined to assess model performance.

- *ARI*. This metric is tailored for cell clustering tasks with known true cell type labels, comprehensively balancing the consistency between predicted clusters and real cell types while correcting for false positives from random clustering. Given the ground truth cell type assignment C and the clustering result K, the ARI is defined as:

$$\text{RI} = \frac{a + b}{C_2^{n_{\text{samples}}}} \tag{12}$$

$$\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]} \tag{13}$$

where a is the number of pairs of elements that are in the same set in both C and K, b is the number of pairs of elements that are in different sets in both C and K, n_{samples} samples is the total number of samples, and $E[\text{RI}]$ is the expected Rand Index (RI) of random labelings.

- *SCS*. This metric integrates the average intra-cluster distance and nearest inter-cluster distance through a normalized formula, evaluating the quality of unsupervised cell clustering without relying on true labels. It is defined as:

$$s = \frac{b - a}{\max(a, b)} \tag{14}$$

where a is the mean distance between a sample and all other points in the same class, and b is the mean distance between a sample and all other points in the next nearest cluster.

- *Macro F1-Score* (F1-macro). This metric calculates the F1-Score for each cell type individually and takes the average, evaluating the robust performance of cell type classification tasks and avoiding bias from imbalanced cell type distributions. It is defined as:

$$\text{F1-macro} = \frac{1}{k} \sum_{i=1}^{k} \text{F1-score}_i \tag{15}$$

where k represents the total number of cell types.

The above metrics form a comprehensive evaluation criteria for multi-scale AIVC. In addition to these well-defined metrics, further validations are warranted in terms of both biological and computational aspects. Specifically, AIVC is expected to reveal full-chain biological or pathological mechanism, which may involve multiple scales. A joint cross-scale evaluation metric (but focusing on a particular mechanism) is needed to reflect disease-specific performance. This would be well-received in practical applications. Next, from a computational aspect, metrics reflecting model efficiency are to be considered so that to better accommodate the high-throughput demands. Cross-domain scalability is another computational factor since it reflects the model's capability in adapting to new data modalities, a common scene in the field of cell research.

Potential applications of cross-scale virtual cell

Disease research

Research on disease mechanisms is a core topic in the field of biomedicine. Traditional disease research methods have obvious scale gaps when analyzing complex diseases [146]. Although tissue sections and imaging examinations can show macroscopic pathological changes, they cannot reveal the pathogenic mechanisms at the molecular level [147]. Single-cell sequencing technology provides gene expression profiles but lacks information on dynamic processes [148]. More importantly, it is difficult for existing technologies to establish a full-chain association from gene mutations to abnormal cell functions and then to tissue pathological changes.

In disease mechanism research, AIVC reveals the dynamic laws of complex pathological processes through multi-scale modeling. AIVC could achieve multi-scale integration from underlying mutations to cell phenotypes, and simultaneously display the conformational changes of mutant proteins and their impacts on cell functions. For example, in the field of neurodegenerative diseases, the model constructs a cross-scale interaction network between neurons and glial cells, simulates the spatial–temporal propagation laws of β-amyloid protein deposition and Tau protein phosphorylation, and predicts the temporal changes of cerebrospinal fluid biomarkers in Alzheimer's disease patients [149].

In the field of clinical diagnosis and treatment, AIVC could reshape the diagnostic paradigm through virtual simulation. The individualized models generated based on patient's multi-omics data could correlate the CT imaging features of early-stage cancer with driver gene mutations, and assist in achieving noninvasive molecular typing [150]. At the treatment level, dynamic simulation targeting T cells can be used to predict the cell expansion kinetics and the risk of cytokine release syndrome [151].

For personalized medicine, AIVC promotes the innovation of treatment paradigms by constructing cell digital twins. By integrating patient's genomic, proteomic, and real-time dynamic data, the model could predict the drug metabolism rate and dynamically optimize

the combination strategy of targeted drugs. Its core potential lies in upgrading the traditional trial-and-error treatment to a closed loop of prediction–intervention–verification. However, it still needs to overcome bottlenecks such as the standardization of multi-omics data integration and the design of clinical translation pathways.

Application pipeline of AIVC for disease research. In this scenario, the input data to AIVC may include patient multi-omics data, clinical imaging, and single-cell transcriptomic data. For the model selection, the combination of a series of scale-specific models is needed, including a gene-scale model like EVE [152] for mutation pathogenicity prediction, a protein-scale model like AlphaFold [115] for conformational change simulation, a pathway-scale model such as BioNetGen [34] for downstream pathway activation analysis, and finally a cell-scale model like CellFM [42] for phenotypic change prediction. The anticipated outputs include full-chain pathogenic mechanism from gene mutation to cellular phenotype, correlation maps between clinical imaging features and molecular mechanisms, and personalized disease risk scores with potential therapeutic targets.

Drug discovery

Despite substantial annual investments ($1 billion) in drug research and development, the number of new drug approvals has not seen a significant increase [153]. The reasons are mainly attributed to the following limitations of existing drug discovery pipelines. First, *in vitro* experiments are not adequate to reflect the complicated biological environments in the human body, and there are significant species differences between animal models and humans. Besides, clinical trials are limited by sample size and ethical constraints. Especially in the target discovery stage, more than half of the candidate targets are eliminated due to lack of evidence of efficacy or safety [154–156].

AIVC could construct a test population consisting of millions of virtual patients, covering different ages, genders, and genotypes, and support digital clinical trials, significantly reducing R & D risks. In addition, it could analyze the mechanisms of action of existing drugs through reverse engineering to facilitate the new uses of old drugs. For example, in the early stage of drug discovery, AIVC could build a virtual screening system. By combining molecular dynamics and deep learning algorithms, it could efficiently identify target binding molecules from a compound library of hundreds of millions [157]. In the development of KRAS G12C inhibitors, it could predict the allosteric binding sites of AMG 510 [158]. Meanwhile, the model could issue early warnings of cardiotoxicity by simulating drug-off-target protein interactions, replacing traditional *in vitro* experiments. In the drug efficacy evaluation stage, the multi-scale modeling ability of AIVC enables full-chain simulation from molecular binding to tissue distribution. For instance, a hepatocyte immune cell interaction network could be constructed through the hepatotoxicity prediction module to simulate the damage process caused by NAPQI, a metabolite of acetaminophen.

Application pipeline of AIVC for drug discovery. In this scenario, the input data cover compound structure datasets, protein–ligand binding affinity data, and single-cell perturbation transcriptomic profiles (e.g. Tahoe-100M). Similarly, the combination of scale-specific models include a gene-scale model like PERD [159] for identifying drug-responsive regulatory elements, a protein-scale model like HiSIF-DTA [97] for predicting ligand binding poses and allosteric effects, pathway-scale models like DT-LEMBAS [160] for simulating drug-induced signaling pathway perturbations, and cell-scale models like TIGER [161] for evaluating cell viability and off-target effects. The anticipated outputs include optimized lead compound structures, drug–target interaction mechanism interpretations, toxicity risk early warnings, and virtual clinical trial efficacy prediction results for specific patient cohorts.

Bioengineering

The bioengineering field is currently facing a bottleneck of low efficiency in design and manufacturing, especially in key areas such as protein design and antimicrobial peptide development [162]. Traditional methods mainly rely on directed evolution and rational design. These techniques have limitations such as long experimental cycles, high costs, and difficulty in predicting the synergistic effects of complex mutation combinations [163]. In the research and development of antimicrobial peptides, traditional methods are even more restricted by the limited natural template library and the inefficient activity screening process [164].

In the field of synthetic biology, AIVC could guide the design of artificial life systems through dynamic metabolic modeling. Using geometric deep learning algorithms, AIVC could perform intelligent sampling in the known protein structure space, generating millions of brands new protein backbones that comply with physical laws. At the same time, it could optimize multiple key parameters such as the stability, activity of proteins. Based on the above technologies, AIVC could directly generate the optimal design scheme for catalytic pockets by simulating protein folding kinetics and substrate binding patterns. In addition, the simulation ability of AIVC is breaking through the bottlenecks of organoid culture technology. AIVC could predict the combination of regulatory factors for the formation of the crypt villus structure in intestinal organoids by establishing a model of the stem-cell differentiation microenvironment.

AIVC is reshaping the research paradigm of bioengineering. In the future, with the continuous upgrading of algorithms and the continuous expansion of databases, AIVC is expected to achieve a leap from single-molecule design to whole-cell reprogramming. Especially in addressing major challenges such as the threat of new pathogens and solving environmental governance problems, AIVC-driven bioengineering will play an irreplaceable role.

Application pipeline of AIVC for bioengineering. In this scenario, the input data contain synthetic gene sequence libraries, protein structure design templates, and microbial metabolic pathway databases. The scale-specific model selection involves a gene-scale model like RankGAN [165] for optimizing promoter and enhancer sequences, a protein-scale model like CoSaNN [166] for *de novo* enzyme design, a pathway-scale model like scFEA [167] for balancing metabolic flux distribution, and a cell-scale model like JCVI-syn3A [24] for simulating the growth dynamics of engineered cells/organoids. The anticipated outputs include high-activity synthetic enzymes, stable engineered microbial strains, organoid culture condition optimization schemes, and predictive models for cell reprogramming efficiency.

Challenges and prospects

A key challenge in AIVC development lies in the integration of mechanistic and data-driven models, as well as the propagation of uncertainty inherent to both paradigms. In detail, mechanistic models conduct bottom-up simulations based on known biological mechanisms, featuring strong interpretability but limited generalization ability. In contrast, data-driven models achieve end-to-end

prediction based on multi-omics data, with excellent generalization but suffering from the issue of weak interpretability. The inherent differences between the two model types lead to two core challenges in their integration: one is the difficulty in reconciling deterministic mechanistic rules with probabilistic data-driven outputs, and the other is the problem of quantifying the propagation of uncertainty from both models throughout the entire AIVC pipeline.

In response to these challenges, we propose an integration framework as a potential solution. Specifically, mechanistic models can serve as prior constraints to limit biological interactions to feasible ranges, narrowing the prediction space of data-driven models. While data-driven models, as an adaptive explorer, seek to discover unreported regulatory relationships from multi-omics data, complementing unknown mechanisms. Meanwhile, a bidirectional feedback channel can be constructed——on the one hand, the prediction results of data-driven models are validated against mechanistic rules to exclude biologically implausible conclusions; on the other hand, mechanistic constraints can be incorporated as regularization terms into the training process of data-driven models to enhance prediction reliability. As for the uncertainty evaluation, a possible way is to maintain a mutual consistency validator to assess the prediction confidence. Consistent predictions across both paradigms indicate high reliability, while inconsistencies trigger further analysis such as sensitivity testing for mechanistic model parameters or data quality validation for data-driven models. Although this framework provides ideas for the integration of the two types of models, the complete resolution of the aforementioned core challenges still requires continuous exploration in the field, and these directions have been highlighted as key topics for future research.

The quality and heterogeneity of data constrain the accuracy of the model. The construction of AIVC is highly dependent on the integration of cross-scale and multi-modal data, yet multiple issues at the data level have become the primary obstacles. On the one hand, although there are numerous datasets available at various levels, dedicated datasets tailored to the full-chain simulation needs of AIVC remain scarce. Additionally, there is a lack of unified public evaluation criteria, especially at the cellular level, where existing data fail to meet the requirements for standardized and scenario-specific data in cell dynamic simulations. On the other hand, data heterogeneity is a prominent problem. For instance, gene detection is widely absent in single-cell transcriptome data, and the insufficient resolution of spatial omics data makes it difficult to capture the dynamic changes of subcellular structures. Furthermore, the dynamic behaviors of cells themselves remain extremely challenging to depict accurately to this day, as they involve complex temporal processes such as molecular interactions and signal transduction. In addition, the annotation of pathological mechanisms relies on wet-lab experiments for verificationthat is costly and time-consuming. Differences in cross-species data further exacerbate the issue of model generalization. For example, the expression profiles of drug-metabolizing enzymes in preclinical mouse models differ significantly from those in humans, ultimately leading to biases in predictions.

AIVC driven by deep learning often triggers a trust crisis due to its black box nature, and there is an urgent need to improve the model's interpretability and generalization ability. For example, the binding sites of KRAS inhibitors predicted by graph neural networks do not overlap with the known allosteric domains, and the lack of mechanism interpretability hinders clinical translation. In addition, in rare disease research, the scarcity of samples leads to model overfitting, resulting in poor stability of prediction results. The generalization ability of existing models in cross-tissue and cross-disease scenarios still need to be verified.

Computational resources and timeliness limit large-scale applications. Whole-cell dynamic simulation is a highly complex process that involves the precise simulation of thousands of biochemical reactions within cells. These reactions are usually described by differential equations, which require powerful computational resources for real-time solution, far exceeding the computing power reserves of regular laboratories.

Despite facing numerous challenges, the development prospects of AIVC technology remain broad. In terms of technological breakthroughs, the next-generation AIVC will exhibit three major development trends. First, with the maturity of the quantum computing, the advanced quantum-based supercomputing system is expected to provide critical computing power support for building a more refined multi-scale AIVC, though it remains to be seen whether it could accommodate the explosion of dimensions caused by multi-scale simulations. The latest research results show that quantum algorithms have overcome the difficulties of classical computers in dealing with the folding problems of proteins with complex and subtly changing 3D shapes, which lays the foundation for real-time simulation at the whole-cell scale [168]. Second, the progress of interpretable AI technology is opening up the black box of the model. For example, ExplainableFold developed by Tan *et al.* can visualize the key interaction paths between protein residues, greatly enhancing the credibility of the model [169]. Finally, the application of privacy computing technologies such as federated learning makes it possible to achieve multi-center collaborative modeling without sharing the original data [170].

Key points

- We propose a cross-scale coupling mechanism for AIVC that encompasses multiple biological hierarchies of "gene–protein–pathway-cell," and decompose its technical logic.
- We systematically collate existing models and datasets in the AIVC field, establishing a directly referable resource system to provide clear guidance and foundational support for subsequent research.
- We deeply analyze core challenges for AIVC development and elaborate on its application value in disease research, drug discovery, and bioengineering.

Author contributions

H.J., X.H., X.B., and S.Z. wrote the manuscript. H.J., W.M., H.N., P.S., and H.Z. collected and processed the data. H.J., Z.W., and S.Z. acquired funding and provided resources. All authors reviewed and approved the final manuscript.

Conflicts of interest

None declared.

Funding

This work was supported by the National Natural Science Foundation of China (no. 62306293), the Natural Science Foundation of Shandong Province (no. ZR2025MS1069), the Fundamental Research

Funds for the Central Universities (no. 202561013), and the Youth Innovation Technology Project of Higher School in Shandong Province (no. 2025KJH006).

Data availability

No datasets have been utilized in this review paper.

References

1. Mazzarello P. A unifying concept: the history of cell theory. *Nat Cell Biol* 1999;**1**:E13–5.
2. Zeng H. What is a cell type and how to define it? *Cell* 2022;**185**:2739–55. https://doi.org/10.1016/j.cell.2022.06.031
3. Sadler F, Ma N, Ritt M *et al.* Autoregulation of GPCR signalling through the third intracellular loop. *Nature* 2023;**615**:734–41. https://doi.org/10.1038/s41586-023-05789-z
4. Wesley BT, Ross ADB, Muraro D *et al.* Single-cell atlas of human liver development reveals pathways directing hepatic cell fates. *Nat Cell Biol* 2022;**24**:1487–98. https://doi.org/10.1038/s41556-022-00989-7
5. Regev A, Teichmann SA, Lander ES *et al.* The human cell atlas. *elife* 2017;**6**:e27041.
6. Grifoni A, Sidney J, Vita R *et al.* SARS-CoV-2 human T cell epitopes: adaptive immune response against COVID-19. *Cell Host Microbe* 2021;**29**:1076–92. https://doi.org/10.1016/j.chom.2021.05.010
7. Burdziak C, Alonso-Curbelo D, Walle T *et al.* Epigenetic plasticity cooperates with cell-cell interactions to direct pancreatic tumorigenesis. *Science* 2023;**380**:eadd5327. https://doi.org/10.1126/science.add5327
8. Zhuo C, Zhang J, Jung-Hwan Lee J *et al.* Spatiotemporal control of CRISPR/Cas9 gene editing. *Signal Transduct Target Ther* 2021;**6**:238. https://doi.org/10.1038/s41392-021-00645-w
9. Alber AB, Suter DM. Dynamics of protein synthesis and degradation through the cell cycle. *Cell Cycle* 2019;**18**:784–94. https://doi.org/10.1080/15384101.2019.1598725
10. Skinnider MA, Squair JW, Courtine G. Enabling reproducible reanalysis of single-cell data. *Genome Biol* 2021;**22**:215. https://doi.org/10.1186/s13059-021-02422-y
11. Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. *Database* 2020;**2020**:baaa073.
12. Craig Venter J, Adams MD, Myers EW *et al.* The sequence of the human genome. *Science* 2001;**291**:1304–51.
13. Johnson GT, Agmon E, Akamatsu M *et al.* Building the next generation of virtual cells to understand cellular biology. *Biophys J* 2023;**122**:3560–9. https://doi.org/10.1016/j.bpj.2023.04.006
14. Marx V. How to build a virtual embryo. *Nat Methods* 2023;**20**:1838–43. https://doi.org/10.1038/s41592-023-02094-5
15. Goldberg AP, Szigeti B, Chew YH *et al.* Emerging whole-cell modeling principles and methods. *Curr Opin Biotechnol* 2018;**51**:97–102. https://doi.org/10.1016/j.copbio.2017.12.013
16. Georgouli K, Yeom J-S, Blake RC *et al.* Multi-scale models of whole cells: progress and challenges. *Front Cell Dev Biol* 2023;**11**:1260507.
17. Viana MP, Chen J, Knijnenburg TA *et al.* Integrated intracellular organization and its variations in human iPS cells. *Nature* 2023;**613**:345–54. https://doi.org/10.1038/s41586-022-05563-7
18. Johnson GT, Autin L, Al-Alusi M *et al.* Cellpack: a virtual mesoscope to model and visualize structural systems biology. *Nat Methods* 2015;**12**:85–91. https://doi.org/10.1038/nmeth.3204
19. Loew LM, Schaff JC. The virtual cell: a software environment for computational cell biology. *Trends Biotechnol* 2001;**19**:401–6.
20. Yugi K, Tomita M. A general computational model of mitochondrial metabolism in a whole organelle scale. *Bioinformatics* 2004;**20**:1795–6. https://doi.org/10.1093/bioinformatics/bth125
21. Cuellar AA, Nielsen PF, Bullivant DP *et al.* CellML 1.1 for the definition and exchange of biological models. *IFAC Proc Vol* 2003;**36**:451–6.
22. Varga MJ, Yiben F, Loggia S *et al.* NERDSS: a nonequilibrium simulator for multibody self-assembly at the cellular scale. *Biophys J* 2020;**118**:3026–40.
23. Hoffmann M, Fröhner C, Noé F. ReaDDy 2: fast and flexible software framework for interacting-particle reaction dynamics. *PLoS Comput Biol* 2019;**15**:e1006830. https://doi.org/10.1371/journal.pcbi.1006830
24. Thornburg ZR, Bianchi DM, Brier TA *et al.* Fundamental behaviors emerge from simulations of a living minimal cell. *Cell* 2022;**185**:345–360.e28. https://doi.org/10.1016/j.cell.2021.12.025
25. Hao M, Gong J, Zeng X *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91. https://doi.org/10.1038/s41592-024-02305-7
26. Fu X, Mo S, Buendia A *et al.* A foundation model of transcription across human cell types. *Nature* 2025;**637**:965–73. https://doi.org/10.1038/s41586-024-08391-z
27. Bunne C, Roohani Y, Rosen Y *et al.* How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 2024;**187**:7045–63. https://doi.org/10.1016/j.cell.2024.11.015
28. Yang T, Wang Y-Y, Ma F *et al.* Build the virtual cell with artificial intelligence: a perspective for cancer research. *Mil Med Res* 2025;**12**:4.
29. Katsoulakis E, Wang Q, Huanmei W *et al.* Digital twins for health: a scoping review. *NPJ Digit Med* 2024;**7**:77.
30. CZI Cell Science Program, Abdulla S, Aevermann B *et al.* CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2025;**53**:D886–900.
31. Roohani YH, Hua TJ, Tung P-Y *et al.* Virtual cell challenge: toward a turing test for the virtual cell. *Cell* 2025;**188**:3370–4.
32. Müssel C, Hopfensitz M, Kestler HA. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* 2010;**26**:1378–80. https://doi.org/10.1093/bioinformatics/btq124
33. Majarian TD, Cao-Berg I, Ruan X *et al.* CellOrganizer: learning and using cell geometries for spatial cell simulations. In: Hlavacek WS (ed.), *Modeling Biomolecular Site Dynamics: Methods and Protocols*. New York, NY: Springer, 2019, 251–64. https://doi.org/10.1007/978-1-4939-9102-0_11
34. Blinov ML, Faeder JR, Goldstein B *et al.* BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics* 2004;**20**:3289–91.
35. Shen H, Shen X, Feng M *et al.* A universal approach for integrating super large-scale single-cell transcriptomes by exploring gene rankings. *Brief Bioinform* 2022;**23**:bbab573. https://doi.org/10.1093/bib/bbab573
36. Shen H, Liu J, Jiani H *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering. *Iscience* 2023;**26**:106536. https://doi.org/10.1016/j.isci.2023.106536
37. Theodoris CV, Xiao L, Chopra A *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9
38. Yang F, Wang W, Wang F *et al.* ScBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;**4**:852–66. https://doi.org/10.1038/s42256-022-00534-z

39. Cui H, Wang C, Maan H *et al.* ScGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80.
40. Rosen Y, Roohani Y, Agarwal A *et al.* Universal cell embeddings: a foundation model for cell biology. *bioRxiv* 2023; **2023–11**:28.568918. https://doi.org/10.1101/2023.11.28.568918
41. Yang X, Liu G, Feng G *et al.* GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res* 2024;**34**:830–45. https://doi.org/10.1038/s41422-024-01034-y
42. Zeng Y, Xie J, Shangguan N *et al.* CellFM: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nat Commun* 2025;**16**:4679.
43. Karr JR, Sanghvi JC, Macklin DN *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* 2012;**150**:389–401. https://doi.org/10.1016/j.cell.2012.05.044
44. Maritan M, Autin L, Karr J *et al.* Building structural models of a whole mycoplasma cell. *J Mol Biol* 2022;**434**:167351. https://doi.org/10.1016/j.jmb.2021.167351
45. Mateo-Victoriano B, Samaranayake GJ, Pokharel S *et al.* Oncogenic KRAS addiction states differentially influence MTH1 expression and 8-oxodGTPase activity in lung adenocarcinoma. *Redox Biol* 2025;**82**:103610. https://doi.org/10.1016/j.redox.2025.103610
46. Awad MM, Liu S, Rybkin II *et al.* Acquired resistance to KRASG12C inhibition in cancer. *New Engl J Med* 2021;**384**:2382–93. https://doi.org/10.1056/NEJMoa2105281
47. Newport TD, Sansom MSP, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. *Nucleic Acids Res* 2019;**47**:D390–7.
48. Meersche YV, Cretin G, Gheeraert A *et al.* Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res* 2024;**52**:D384–92.
49. Meersche YV, Cretin G, de Brevern AG *et al.* MEDUSA: prediction of protein flexibility from sequence. *J Mol Biol* 2021;**433**:166882.
50. Mund A, Brunner A-D, Mann M. Unbiased spatial proteomics with single-cell resolution in tissues. *Mol Cell* 2022;**82**:2335–49. https://doi.org/10.1016/j.molcel.2022.05.022
51. Petrosius V, Aragon-Fernandez P, Üresin N *et al.* Exploration of cell state heterogeneity using single-cell proteomics through sensitivity-tailored data-independent acquisition. *Nat Commun* 2023;**14**:5910.
52. Kustatscher G, Collins T, Gingras A-C *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat Methods* 2022;**19**:774–9. https://doi.org/10.1038/s41592-022-01454-x
53. Chen P, Wang H, Zhang W *et al.* Loss of BAP1 results in growth inhibition and enhances mesenchymal–epithelial transition in kidney tumor cells. *Mol Cell Proteomics* 2019;**18**:1320–9. https://doi.org/10.1074/mcp.RA119.001457
54. Kalous J, Aleshkina D, Anger M. A role of PI3K/Akt signaling in oocyte maturation and early embryo development. *Cells* 2023;**12**:1830. https://doi.org/10.3390/cells12141830
55. Fan T-J, Han L-H, Cong R-S *et al.* Caspase family proteases and apoptosis. *Acta Biochim Biophys Sin* 2005;**37**:719–27. https://doi.org/10.1111/j.1745-7270.2005.00108.x
56. Zhang SEN, Sun W-Y, Jing-Jing W *et al.* Decreased expression of the type III TGF-β receptor enhances metastasis and invasion in hepatocellullar carcinoma progression. *Oncol Rep* 2016;**35**:2373–81. https://doi.org/10.3892/or.2016.4615
57. Mijit M, Caracciolo V, Melillo A *et al.* Role of p53 in the regulation of cellular senescence. *Biomolecules* 2020;**10**:420. https://doi.org/10.3390/biom10030420
58. Landrum MJ, Lee JM, Benson M *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**:D862–8.
59. Stenson PD, Mort M, Ball EV *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* 2020;**139**:1197–207. https://doi.org/10.1007/s00439-020-02199-3
60. Abdulla Bava K, Michael Gromiha M, Uedaira H *et al.* ProTherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res* 2004;**32**:D120–1.
61. Stourac J, Dubrava J, Musil M *et al.* FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res* 2021;**49**:D319–24. https://doi.org/10.1093/nar/gkaa981
62. Jankauskaitė J, Jiménez-García B, Dapkūnas J *et al.* SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* 2019;**35**:462–9. https://doi.org/10.1093/bioinformatics/bty635
63. Bycroft C, Freeman C, Petkova D *et al.* The UKBiobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–9. https://doi.org/10.1038/s41586-018-0579-z
64. MacArthur J, Bowler E, Cerezo M *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;**45**:D896–901. https://doi.org/10.1093/nar/gkw1133
65. Piñero J, Bravo À, Queralt-Rosinach N *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2016;**45**:D833–9. https://doi.org/10.1093/nar/gkw943
66. Barrett T, Wilhite SE, Ledoux P *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2012;**41**:D991–5.
67. Burley SK, Berman HM, Kleywegt GJ *et al.* Protein data bank (PDB): the single global macromolecular structure archive. In: Wlodawer A, Dauter Z, Jaskolski M (eds.), *Protein Crystallography: Methods and Protocols*. New York, NY: Springer New York, 2017:627–641. https://doi.org/10.1007/978-1-4939-7000-1_26
68. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**:D506–15.
69. Von Mering C, Jensen LJ, Snel B *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res* 2005;**33**:D433–7.
70. Oughtred R, Rust J, Chang C *et al.* The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 2021;**30**:187–200. https://doi.org/10.1002/pro.3978
71. Kerrien S, Aranda B, Breuza L *et al.* The intact molecular interaction database in 2012. *Nucleic Acids Res* 2012;**40**:D841–6. https://doi.org/10.1093/nar/gkr1088
72. Salwinski L, Miller CS, Smith AJ *et al.* The database of interacting proteins: 2004 update. *Nucleic Acids Res* 2004;**32**:D449–51.
73. Wang R, Fang X, Yipin L *et al.* The PDBbind database: methodologies and updates. *J Med Chem* 2005;**48**:4111–9.
74. Siebenmorgen T, Menezes F, Benassou S *et al.* MISATO: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nat Comput Sci* 2024;**4**:367–78. https://doi.org/10.1038/s43588-024-00627-2
75. Tang J, Szwajda A, Shakyawar S *et al.* Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J Chem Inf Model* 2014;**54**:735–43. https://doi.org/10.1021/ci400709d

76. Davis MI, Hunt JP, Herrgard S *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat Biotechnol* 2011;**29**:1046–51. https://doi.org/10.1038/nbt.1990
77. Liu H, Sun J, Guan J *et al.* Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 2015;**31**:i221–9. https://doi.org/10.1093/bioinformatics/btv256
78. Lindeboom RGH, Regev A, Teichmann SA. Towards a human cell atlas: taking notes from the past. *Trends Genet* 2021;**37**:625–30. https://doi.org/10.1016/j.tig.2021.03.007
79. The Tabula Sapiens Consortium∗, Jones RC, Karkanias J *et al.* The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* 2022;**376**:eabl4896.
80. Zhang J, Ubas AA, de Borja R *et al.* Tahoe-100m: a giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv* 2025;**2025–02**:20.639398. https://doi.org/10.1101/2025.02.20.639398
81. Youngblut ND, Carpenter C, Prashar J *et al.* scBaseCount: an AI agent-curated, uniformly processed, and continually expanding single cell data repository. *bioRxiv* 2025;**2025–02**:27.640494. https://doi.org/10.1101/2025.02.27.640494
82. Williams E, Moore J, Li SW *et al.* Image data resource: a bioimage data integration and publication platform. *Nat Methods* 2017;**14**:775–81. https://doi.org/10.1038/nmeth.4326
83. Efremova M, Vento-Tormo M, Teichmann SA *et al.* CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;**15**:1484–506.
84. Gunasekaran H, Ramalakshmi K, Arokiaraj RAM *et al.* Analysis of DNA sequence classification using CNN and hybrid models. *Comput Math Methods Med* 2021;**2021**:1835056. https://doi.org/10.1155/2021/1835056
85. Han G-S, Li Q, Li Y. Nucleosome positioning based on DNA sequence embedding and deep learning. *BMC Genomics* 2022;**23**:301. https://doi.org/10.1186/s12864-022-08508-6
86. Sharma S, Kalra N, Rani R. Exploring genomic sequence encoding methods for 1D convolutional neural networks: a comparative study. In: Soni B, Saini P, Verma GK, Gupta BB (eds.), *Beyond Artificial Intelligence*. Singapore: Springer Nature Singapore, 2025, 651–63. https://doi.org/10.1007/978-981-96-4170-3_51
87. Almuhaideb S, Altwaijry N, Al-Turaiki I *et al.* A comparative study of data representation techniques for deep learning-based classification of promoter and histone-associated DNA regions. *Computers, Materials and Continua* 2025;**85**:3095–128. https://doi.org/10.32604/cmc.2025.067390
88. Li G, Yuan Y, Zhang R. Predicting protein–ligand binding affinity using fusion model of spatial-temporal graph neural network and 3d structure-based complex graph. *Interdiscip Sci-Comput Life Sci* 2025;**17**:257–76. https://doi.org/10.1007/s12539-024-00644-9
89. Nguyen T, Le H, Quinn TP *et al.* GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* 2021;**37**:1140–7. https://doi.org/10.1093/bioinformatics/btaa921
90. Jha K, Saha S, Singh H. Prediction of protein–protein interaction using graph neural networks. *Sci Rep* 2022;**12**:8360. https://doi.org/10.1038/s41598-022-12201-9
91. Wang L, Liu H, Liu Y *et al.* Learning hierarchical protein representations via complete 3D graph networks. *arXiv* 2022;**2207**:12600. https://doi.org/10.48550/arXiv.2207.12600
92. Zhang Z, Xu M, Jamasb A *et al.* Protein representation learning by geometric structure pretraining. *arXiv* 2022;**2203**:06125. https://doi.org/10.48550/arXiv.2203.06125
93. Chatzipantazis E, Pertigkiozoglou S, Dobriban E *et al.* SE (3)-equivariant attention networks for shape reconstruction in function space. *arXiv* 2022;**2204**:02394. https://doi.org/10.48550/arXiv.2204.02394
94. Li J, Bi X, Ma W *et al.* MHAN-DTA: a multiscale hybrid attention network for drug-target affinity prediction. *IEEE J Biomed Health Inform* 2024;**29**:7910–21. https://doi.org/10.1109/JBHI.2024.3518619
95. Ma W, Bi X, Jiang H *et al.* Annotating protein functions via fusing multiple biological modalities. *Commun Biol* 2024;**7**:1705. https://doi.org/10.1038/s42003-024-07411-y
96. Nair A, Chauhan P, Saha B *et al.* Conceptual evolution of cell signaling. *Int J Mol Sci* 2019;**20**:3292. https://doi.org/10.3390/ijms20133292
97. Bi X, Zhang S, Ma W *et al.* HiSIF-DTA: a hierarchical semantic information fusion framework for drug-target affinity prediction. *IEEE J Biomed Health Inform* 2023;**29**:1579–90. https://doi.org/10.1109/JBHI.2023.3334239
98. Ma W, Zhang S, Li Z *et al.* Predicting drug-target affinity by learning protein knowledge from biological networks. *IEEE J Biomed Health Inform* 2023;**27**:2128–37. https://doi.org/10.1109/JBHI.2023.3240305
99. Ge F, Li C-F, Zhang C-M *et al.* PRITrans: a transformer-based approach for the prediction of the effects of missense mutation on protein–RNA interactions. *Int J Mol Sci* 2024;**25**:12348.
100. Mitra R, Li J, Sagendorf JM *et al.* Geometric deep learning of protein–DNA binding specificity. *Nat Methods* 2024;**21**:1674–83. https://doi.org/10.1038/s41592-024-02372-w
101. Pires DEV, Ascher DB. mCSM-NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Res* 2017;**45**:W241–6. https://doi.org/10.1093/nar/gkx236
102. Li W, Zhang S, Yang G. Dynamic organization of intracellular organelle networks. *WIREs Mech Dis* 2021;**13**:e1505. https://doi.org/10.1002/wsbm.1505
103. Evers TMJ, Holt LJ, Alberti S *et al.* Reciprocal regulation of cellular mechanics and metabolism. *Nat Metab* 2021;**3**:456–68. https://doi.org/10.1038/s42255-021-00384-w
104. Sun H, Xuecong F, Abraham S *et al.* Improving and evaluating deep learning models of cellular organization. *Bioinformatics* 2022;**38**:5299–306. https://doi.org/10.1093/bioinformatics/btac688
105. Huang C-J, Lee Y-J, Wei A-C. Cell cycle phase classification from deep learning-predicted images of cell organelles. In: Hu RM (ed.), *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. Corfu, Greece: IEEE, 2022, 199–203. https://doi.org/10.1109/BIBE55377.2022.00050
106. Li M, Xing X, Yuan J *et al.* Research progress on the regulatory role of cell membrane surface tension in cell behavior. *Heliyon* 2024;**10**:e29923. https://doi.org/10.1016/j.heliyon.2024.e29923
107. Dai Z, Yang Z, Yang Y *et al.* Transformer-XL: attentive language models beyond a fixed-length context. *arXiv* 2019;**1901**:02860. https://doi.org/10.48550/arXiv.1901.02860
108. Chiliński M, Plewczynski D. HiCDiffusion-diffusion-enhanced, transformer-based prediction of chromatin interactions from DNA sequences. *BMC Genomics* 2024;**25**:964. https://doi.org/10.1186/s12864-024-10885-z
109. Sun J, Guo J, Liu J. CRISPR-M: predicting sgRNA off-target effect using a multi-view deep learning network. *PLoS Comput Biol* 2024;**20**:e1011972. https://doi.org/10.1371/journal.pcbi.1011972
110. Zheng Y, Zou Q, Li J *et al.* CRISPR-MFH: a lightweight hybrid deep learning framework with multi-feature encoding for improved CRISPR-Cas9 off-target prediction. *Genes* 2025;**16**:387.

111. Emtiaj AH, Rafi RH, Nayeem MA *et al.* CRISMER: a transformer-based interpretable deep learning approach for genome-wide CRISPR Cas-9 off-target prediction and optimization. *bioRxiv* 2025;**2025–05**:03.652008. https://doi.org/10.1101/2025.05.03.652008
112. Zeng Z, Mao C, Vo A *et al.* Deep learning for cancer type classification and driver gene identification. *BMC Bioinformatics* 2021;**22**:491. https://doi.org/10.1186/s12859-021-04400-4
113. Brandes N, Goldman G, Wang CH *et al.* Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* 2023;**55**:1512–22. https://doi.org/10.1038/s41588-023-01465-0
114. Wu L, Tian Y, Lin H *et al.* Learning to predict mutation effects of protein-protein interactions by microenvironment-aware hierarchical prompt learning. *arXiv* 2024;**2405**:10348. https://doi.org/10.48550/arXiv.2405.10348
115. Senior AW, Evans R, Jumper J *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10. https://doi.org/10.1038/s41586-019-1923-7
116. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2
117. Abramson J, Adler J, Dunger J *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 2024;**630**:493–500. https://doi.org/10.1038/s41586-024-07487-w
118. Yu L, Gao Z, Tan C *et al.* Protein-SE (3): benchmarking SE (3)-based generative models for protein structure design. *arXiv* 2025;**2507**:20243. https://doi.org/10.48550/arXiv.2507.20243
119. Vladimir Gligorijević P, Renfrew D, Kosciolek T *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:3168.
120. Min Y, Wei Y, Wang P *et al.* From static to dynamic structures: improving binding affinity prediction with graph-based deep learning. *Adv Sci* 2024;**11**:2405404.
121. Shah A, Dixit S. DynPred: A multimodal deep learning approach for protein dynamics prediction. In: *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA, Shenzhen, China: Association for Computing Machinery (ACM), 2024, 80. https://doi.org/10.1145/3698587.3701429
122. Wang W, Shuai Y, Zeng M *et al.* DPFunc: accurately predicting protein function via deep learning with domain-guided structure information. *Nat Commun* 2025;**16**:70.
123. Zambaldi V, La D, Chu AE *et al.* De novo design of high-affinity protein binders with alphaproteo. *arXiv* 2024;**2409**:08022. https://doi.org/10.48550/arXiv.2409.08022
124. Zheng W, Wuyun Q, Li Y *et al.* Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods* 2024;**21**:279–89. https://doi.org/10.1038/s41592-023-02130-4
125. Zhouxin Y, Huang F, Zhao X *et al.* Predicting drug–disease associations through layer attention graph convolutional network. *Brief Bioinform* 2021;**22**:bbaa243.
126. Luo Y, Shi L, Xiao-Ming W. Classic GNNs are strong baselines: Reassessing GNNs for node classification. *Adv Neural Inf Proces Syst* 2024;**37**:97650–69.
127. Scarselli F, Marco Gori A, Tsoi C *et al.* The graph neural network model. *IEEE Trans Neural Netw* 2008;**20**:61–80. https://doi.org/10.1109/TNN.2008.2005605
128. Li M, Cai X, Li L *et al.* Heterogeneous graph attention network for drug-target interaction prediction. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery (ACM), New York, NY, USA, Atlanta, GA, USA, 2022, 1166–76. https://doi.org/10.1145/3511808.3557346
129. Ashish V. Attention is all you need. *Adv Neural Inf Proces Syst* 2017;**30**:I
130. Sha Y, Qiu Y, Zhou P *et al.* Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nat Mach Intell* 2024;**6**:25–39. https://doi.org/10.1038/s42256-023-00763-w
131. Zhenyu W, Lawrence PJ, Ma A *et al.* Single-cell techniques and deep learning in predicting drug response. *Trends Pharmacol Sci* 2020;**41**:1050–65. https://doi.org/10.1016/j.tips.2020.10.004
132. Yang W, Wang P, Luo M *et al.* DeepCCI: a deep learning framework for identifying cell–cell interactions from single-cell RNA sequencing data. *Bioinformatics* 2023;**39**:btad596.
133. Yang KD, Damodaran K, Venkatachalapathy S *et al.* Predicting cell lineages using autoencoders and optimal transport. *PLoS Comput Biol* 2020;**16**:e1007828. https://doi.org/10.1371/journal.pcbi.1007828
134. Yachimura T, Wang H, Imoto Y *et al.* scEGOT: single-cell trajectory inference framework based on entropic Gaussian mixture optimal transport. *BMC bioinformatics* 2024;**25**:388.
135. Tong A, Huang J, Wolf G *et al.* TrajectoryNet: a dynamic optimal transport network for modeling cellular dynamics. *Proc Mach Learn Res* 2020;**119**:9526–36.
136. Yeo GHT, Saksena SD, Gifford DK. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nat Commun* 2021;**12**:3222. https://doi.org/10.1038/s41467-021-23518-w
137. Zhang Z, Wang Z, Sun Y *et al.* Deciphering cell-fate trajectories using spatiotemporal single-cell transcriptomic data. *npj Syst Biol Appl* 2026;**12**:2. https://doi.org/10.1038/s41540-025-00624-9
138. Tong A, Malkin N, Fatras K *et al.* Simulation-free Schrödinger bridges via score and flow matching. *arXiv* 2023;**2307**:03672. https://doi.org/10.48550/arXiv.2307.03672
139. Zhang Y, Rui W, Dascalu SM *et al.* Multi-scale transformer pyramid networks for multivariate time series forecasting. *IEEE Access* 2024;**12**:14731–41. https://doi.org/10.1109/ACCESS.2024.3357693
140. Zhang R, Hao Y. Time series prediction based on multi-scale feature extraction. *Mathematics* 2024;**12**:973. https://doi.org/10.3390/math12070973
141. Zhao S, Jin M, Hou Z *et al.* HiMTM: hierarchical multi-scale masked time series modeling with self-distillation for long-term forecasting. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery (ACM), New York, NY, USA, Boise, ID, USA, 2024, 3352–62. https://doi.org/10.1145/3627673.3679741
142. Ramasinghe S, Lucey S. Beyond periodicity: towards a unifying framework for activations in coordinate-MLPs. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds.), *European Conference on Computer Vision*. Lecture Notes in Computer Science, Vol. 13693, Cham, Switzerland: Springer, 2022, 142–58. https://doi.org/10.1007/978-3-031-19827-4_9
143. Elfwing S, Uchibe E, Doya K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw* 2018;**107**:3–11. https://doi.org/10.1016/j.neunet.2017.12.012
144. Liu W, Wen Y, Yu Z *et al.* Large-margin softmax loss for convolutional neural networks. In: *Proceedings of the 33rd International*

Conference on Machine Learning, Vol. 48, JMLR.org, New York, NY, USA, 2016, 507–16. https://doi.org/10.5555/3045390.3045445
145. Radivojac P, Clark WT, Oron TR *et al.* A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**:221–7. https://doi.org/10.1038/nmeth.2340
146. Fischer DS, Villanueva MA, Winter PS *et al.* Adapting systems biology to address the complexity of human disease in the single-cell era. *Nat Rev Genet* 2025;**26**:514–31. https://doi.org/10.1038/s41576-025-00821-6
147. Shah M, Polónia A, Curado M *et al.* Impact of tissue thickness on computational quantification of features in whole slide images for diagnostic pathology. *Endocr Pathol* 2025;**36**:1–11. https://doi.org/10.1007/s12022-025-09872-1
148. Chen W, Guillaume-Gentil O, Rainer PY *et al.* Live-seq enables temporal transcriptomic recording of single cells. *Nature* 2022;**608**:733–40. https://doi.org/10.1038/s41586-022-05046-9
149. Ennerfelt H, Frost EL, Shapiro DA *et al.* SYK coordinates neuroprotective microglial responses in neurodegenerative disease. *Cell* 2022;**185**:4135–4152.e22. https://doi.org/10.1016/j.cell.2022.09.030
150. Wang S, Shi J, Ye Z *et al.* Predicting egfr mutation status in lung adenocarcinoma on computed tomography image using deep learning. *Eur Respir J* 2019;**53**:1800986. https://doi.org/10.1183/13993003.00986-2018
151. Cess CG, Finley SD. Data-driven analysis of a mechanistic model of CAR T cell signaling predicts effects of cell-to-cell heterogeneity. *J Theor Biol* 2020;**489**:110125. https://doi.org/10.1016/j.jtbi.2019.110125
152. Frazer J, Notin P, Dias M *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5. https://doi.org/10.1038/s41586-021-04043-8
153. Sertkaya A, Beleche T, Jessup A *et al.* Costs of drug development and research and development intensity in the US, 2000-2018. *JAMA Netw Open* 2024;**7**:e2415445–5. https://doi.org/10.1001/jamanetworkopen.2024.15445
154. Izat N, Bolleddula J, Abbasi A *et al.* Challenges and opportunities for in vitro–in vivo extrapolation of aldehyde oxidase-mediated clearance: toward a roadmap for quantitative translation. *Drug Metab Dispos* 2023;**51**:1591–606. https://doi.org/10.1124/dmd.123.001436
155. Han P, Li X, Yang J *et al.* Advancing toxicity predictions: a review on in vitro to in vivo extrapolation in next-generation risk assessment. *Environment & Health* 2024;**2**:499–513. https://doi.org/10.1021/envhealth.4c00043
156. Sun D, Gao W, Hongxiang H *et al.* Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B* 2022;**12**:3049–62. https://doi.org/10.1016/j.apsb.2022.02.002
157. Gentile F, Yaacoub JC, Gleave J *et al.* Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nat Protoc* 2022;**17**:672–97. https://doi.org/10.1038/s41596-021-00659-2
158. Nerin-Fonz F, Cournia Z. Machine learning approaches in predicting allosteric sites. *Curr Opin Struct Biol* 2024;**85**:102774. https://doi.org/10.1016/j.sbi.2024.102774
159. Wang Y, Wang Y. Identification of drug responsive enhancers by predicting chromatin accessibility change from perturbed gene expression profiles. *NPJ Syst Biol Appl* 2024;**10**:62. https://doi.org/10.1038/s41540-024-00388-8
160. Meimetis N, Lauffenburger DA, Nilsson A. Inference of drug off-target effects on cellular signaling using interactome-based deep learning. *Iscience* 2024;**27**:109509. https://doi.org/10.1016/j.isci.2024.109509
161. Wessels H-H, Stirn A, Méndez-Mancilla A *et al.* Prediction of on-target and off-target activity of CRISPR-Cas13d guide RNAs using deep learning. *Nat Biotechnol* 2024;**42**:628–37. https://doi.org/10.1038/s41587-023-01830-8
162. Gupta A, Garg P, Tiwari V, Palit R. Integrating AI with synthetic biology for custom enzyme design. In: Bansal JC, Jamwal PK, Hussain S (eds.), *International Conference on Sustainable Computing and Intelligent Systems*. Lecture Notes in Networks and Systems, Vol. 1295, Singapore, Bangalore, India: Springer, 2024, 105–16. https://doi.org/10.1007/978-981-96-3311-1_9
163. Kortemme T. De novo protein design—From new structures to programmable functions. *Cell* 2024;**187**:526–44. https://doi.org/10.1016/j.cell.2023.12.028
164. Tan P, Huiyang F, Ma X. Design, optimization, and nanotechnology of antimicrobial peptides: from exploration to applications. *Nano Today* 2021;**39**:101229.
165. Geng Q, Yang R, Zhang L. A deep learning framework for enhancer prediction using word embedding and sequence generation. *Biophys Chem* 2022;**286**:106822. https://doi.org/10.1016/j.bpc.2022.106822
166. Zimmerman L, Alon N, Levin I *et al.* Context-dependent design of induced-fit enzymes using deep learning generates well-expressed, thermally stable and active enzymes. *Proc Natl Acad Sci USA* 2024;**121**:e2313809121. https://doi.org/10.1073/pnas.2313809121
167. Alghamdi N, Chang W, Dang P *et al.* A graph neural network model to estimate cell-wise metabolic flux using single-cell RNA-seq data. *Genome Res* 2021;**31**:1867–84. https://doi.org/10.1101/gr.271205.120
168. Chandarana P, Hegade NN, Montalban I *et al.* Digitized counterdiabatic quantum algorithm for protein folding. *Phys Rev Appl* 2023;**20**:014024. https://doi.org/10.1103/PhysRevApplied.20.014024
169. Tan J, Zhang Y. Explainablefold: understanding alphafold prediction with explainable AI. In: Singh A, Sun Y, Akoglu L et al. (eds.), *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 2023, 2166–76. https://doi.org/10.1145/3580305.3599337
170. Calvino G, Peconi C, Strafella C *et al.* Federated learning: breaking down barriers in global genomic research. *Genes* 2024;**15**:1650.

© The Author(s) 2026. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
Briefings in Bioinformatics, 2026, **27**, bbag104
https://doi.org/10.1093/bib/bbag104
Review