SCIENCE CHINA PRESS

SCIENCE CHINA Life Sciences
life.scichina.com link.springer.com
https://doi.org/10.1007/s11427-024-2859-1

NEWS AND VIEWS

Biomedical data and AI

Hao Xu[1,2†], Shibo Zhou[2†], Zefeng Zhu[2,3], Vincenzo Vitelli[4*], Liangyi Chen[1*], Ziwei Dai[5*], Ning Yang[6*], Luhua Lai[2,6,7*], Shengyong Yang[8*], Sergey Ovchinnikov[9*], Zhuoran Qiao[10*], Sirui Liu[11*], Chen Song[1,2*], Jianfeng Pei[2*], Han Wen[12*], Jianfeng Feng[13*], Yaoyao Zhang[14*], Zhengwei Xie[15*], Yang-Yu Liu[16,17*], Zhiyuan Li[1,2*], Fulai Jin[18*], Hao Li[19*], Mohammad Lotfollahi[20*], Xuegong Zhang[21*], Ge Yang[22,23*], Shihua Zhang[24*], Ge Gao[25*], Pulin Li[26*], Qi Liu[27*] & Jing-Dong Jackie Han[1,2,6*]

[1]Peking-Tsinghua Center for Life Sciences (CLS), Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[2]Center for Quantitative Biology (CQB), Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[3]Peking University-Tsinghua University-National Institute of Biological Sciences Joint Graduate Program, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China
[4]Department of Physics, University of Chicago, Chicago, IL 60637, USA
[5]School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China
[6]Peking University Chengdu Academy for Advanced Interdisciplinary Biotechnologies, Chengdu 610213, China
[7]College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China
[8]Department of Biotherapy, Cancer Center and State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, China
[9]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[10]Lambic Therapeutics, Inc., San Diego, CA 92121, USA
[11]Changping Laboratory, Beijing 102200, China
[12]AI for Science Institute, Beijing 100083, China
[13]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China
[14]Department of Obstetrics and Gynecology, West China Second University Hospital, Sichuan University, Chengdu 610041, China
[15]Peking University International Cancer Institute and Peking University-Yunnan Baiyao International Medical Institute and State Key Laboratory of Natural and Biomimetic Drugs, Department of Molecular and Cellular Pharmacology, School of Pharmaceutical Sciences, Peking University Health Science Center, Peking University, Beijing 100871, China
[16]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA
[17]Center for Artificial Intelligence and Modeling, the Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[18]Department of Genetics and Genome Sciences, School of Medicine and Department of Computer and Data Sciences and Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA
[19]Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94110, USA
[20]Sanger Institute, Cambridge, CB10 1SA, UK
[21]Department of Automation, Tsinghua University, Beijing 100084, China
[22]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[23]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[24]Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China
[25]Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing 100871, China
[26]Whitehead Institute and Eugene Bell Career Development and Tissue Engineering, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[27]State Key Laboratory of Cardiology and Medical Innovation Center, Shanghai East Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200070, China

†Contributed equally to this work
*Corresponding authors (Vincenzo Vitelli, email: vitelli@uchicago.edu; Liangyi Chen, email: lychen@pku.edu.cn; Ziwei Dai, email: daizw@sustech.edu.cn; Ning Yang, email: yn_biophy@pku.edu.cn; Luhua Lai, email: lhlai@pku.edu.cn; Shengyong Yang, email: yangsy@scu.edu.cn; Sergey Ovchinnikov, email: so3@mit.edu; Zhuoran Qiao, email: zrqiao0@gmail.com; Sirui Liu, email: liusirui@cpl.ac.cn; Chen Song, email: c.song@pku.edu.cn; Jianfeng Pei, email: jfpei@pku.edu.cn; Han Wen, email: wenh@aisi.ac.cn; Jianfeng Feng, email: jffeng@fudan.edu.cn; Yaoyao Zhang, email: yaoyaozhang@scu.edu.cn; Zhengwei Xie, email: xiezhengwei@bjmu.edu.cn; Yang-Yu Liu, email: yyl@channing.harvard.edu; Zhiyuan Li, email: zhiyuanli@pku.edu.cn; Fulai Jin, email: fulai.jin@case.edu; Hao Li, email: haoli@genome.ucsf.edu; Mohammad Lotfollahi, email: ml19@sanger.ac.uk; Xuegong Zhang, email: zhangxg@tsinghua.edu.cn; Ge Yang, email: ge.yang@ia.ac.cn; Shihua Zhang, email: zsh@amss.ac.cn; Ge Gao, email: gaog@pku.edu.cn; Pulin Li, email: pli@wi.mit.edu; Qi Liu, email: qiliu@tongji.edu.cn; Jing-Dong Jackie Han, email: jackie.han@pku.edu.cn)

Received 31 December 2024; Accepted 4 February 2025; Published online 14 March 2025

The development of artificial intelligence (AI) and the mining of biomedical data complement each other. From the direct use of computer vision results to analyze medical images for disease screening, to now integrating biological knowledge into models and even accelerating the development of new AI based on biological discoveries, the boundaries of both are constantly expanding, and their connections are becoming closer. Therefore, the theme of the 2024 Annual Quantitative Biology Conference is set as "Biomedical Data and AI", and was held in Chengdu, China from July 15 to 17, 2024. The conference mainly focused on the following three topics: "AI for Quantitative Biology", "AI for Protein and Nucleic Acid Drug Development", and "AI for Multi-omics Data and Analysis" (Figure 1). Due to the time limitation of the conference, only a vignette of the topic of AI and biomedical data was discussed and presented; many other aspects such as brain-

Citation: Xu, H., Zhou, S., Zhu, Z., Vitelli, V., Chen, L., Dai, Z., Yang, N., Lai, L., Yang, S., Ovchinnikov, S., et al. (2025). Biomedical data and AI. Sci China Life Sci 68, 1536–1540. https://doi.org/10.1007/s11427-024-2859-1

CrossMark click for updates

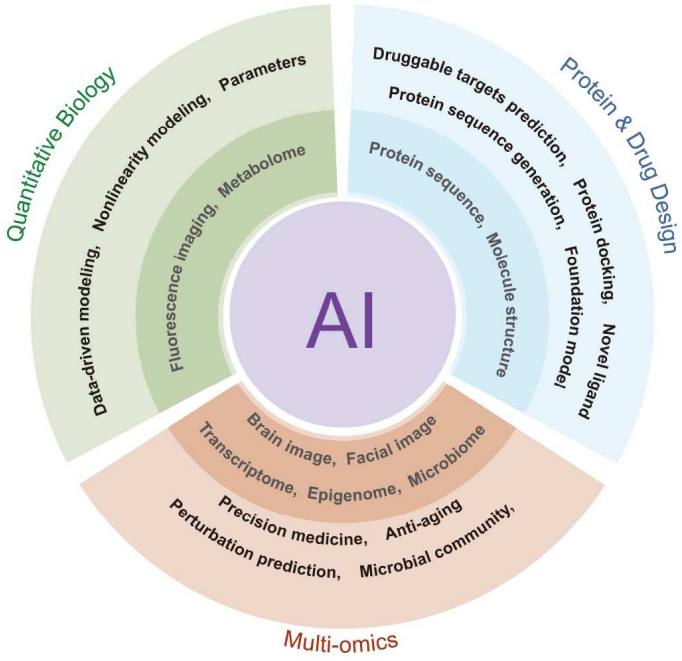

Figure 1. At the "QB2024: Biomedical Data and AI" conference, researchers shared the latest advancements in utilizing AI with various types of biomedical data (middle layer) to achieve various tasks (outer layer) in the fields of "Quantitative Biology", "Protein and Nucleic Acid Drug Development", and "Multi-omics Data and Analysis".

inspired intelligence, medical image processing, medical agents based on large language models (LLM), among others, on this topic are yet to be covered and expected in future events and publications.

Quantitative biology

Speakers in the Quantitative Biology section used systematic models to understand the principles of biological processes.

Vincenzo Vitelli introduced his data-driven AI modeling pipeline, which predicts cellular forces from fluorescence images of proteins like zyxin, and identified key cytoskeletal proteins mediating cellular mechanics. Additionally, by combining machine learning and physical modeling, they reconstructed the dynamic signaling system in *Drosophila* embryo development, unveiling crucial initial signals.

Liangyi Chen introduced his latest work COMET (Cortex-Wide Miniature Mesoscopic Technique). It is a head-mounted miniature mesoscope weighing 3.75 g that allows imaging of up to a 10 mm-diameter field of view in freely moving mice. He has recorded activity from 15,000 neurons within 30 min, helping to study fear responses, social behavior, and object recognition in mice.

Ziwei Dai introduced a theory to maximize the efficiency of linear metabolic pathways. The equilibrium constants of any tandem reactions must equal the square root of the ratio of enzyme activities. When testing this theory on the augmented dataset, issues arose, suggesting that real metabolic systems may be nonlinear or that maximizing flux is not the sole objective of metabolism.

Ning Yang introduced that initially, the loss landscape in deep learning exhibits multiple distinct solution valleys. As training progresses, these valleys gradually become flatter. However, the inherent noise in Stochastic Gradient Descent (SGD) makes it harder for the system to transition into these flatter valleys. This pattern is akin to Waddington's cell fate commitment landscape. These findings may have profound implications for the development of more efficient optimization algorithms.

Overall, these speakers demonstrate that the hybrid of AI and *de novo* mathematical modeling can complement each other, which may also be the next step for artificial intelligence in quantitative biology. On the one hand, deep learning is expected to reduce the complexity of modeling calculations by finding local optima; on the other hand, within the framework of traditional models, the parameters of deep learning are constrained by known knowledge, which diminishes the black-box nature of deep learning.

Protein and nucleic acid drug development

Speakers introduced new prospects in biomedicine.

Luhua Lai discussed how using AI-driven methods to generate novel molecules with desirable properties can accelerate the design-make-test-analyze cycle. They have developed a predictive algorithm for druggable protein targets, a novel algorithm to address the inability of traditional methods to predict flexible structural changes during protein docking. Their work also included tools enabling searches in larger drug-like small molecule spaces, achieving *de novo* design of small molecules and peptide drugs.

Shengyong Yang introduced a ligand-based generative model, which learns from known drug-like small molecules to create new compounds with similar properties but unique scaffolds. They also developed the receptor-based algorithm PocketFlow, which generates novel ligands directly within the active sites of target proteins. PocketFlow also incorporates chemical knowledge in its deep generative framework, achieving 100% chemical validity and high drug-likeness in generated molecules.

Sergey Ovchinnikov discussed the limitations of previous protein sequence design methods and a potential solution. Hallucination refers to constructing a feedback loop during training, where the difference between features extracted by the network and target features is not used to update network parameters but is instead used to directly update the input itself. Through continuous iteration, this process makes the original input increasingly resemble sequences with the target features. By using hallucination to sample sequence space and optimizing with neural networks, meaningful structural sequences are produced. This method ensures sequence diversity and reaches the global energy minimum efficiently, greatly shortening design time.

Zhuoran Qiao introduced NeuralPLexer which predicts the structures of protein-ligand complexes. This technique uses only protein sequences and ligand molecular graphs as inputs to directly predict the 3D structure and conformational

changes of binding complexes. The team is also developing NeuralPLexer2, which features a larger training dataset, increased model parameters, and enhanced structural smoothness, achieving state-of-the-art performance in structure prediction and molecular docking tasks.

Sirui Liu presented a protein-protein docking tool, ColabDock, which surpasses traditional methods in predicting complex structures restrained by residues and surfaces, as well as structures under nuclear magnetic resonance chemical shift perturbations and covalent labeling. Additionally, ColabDock assists in predicting antibody-antigen interfaces by integrating sparse interface restraints from simulated deep mutational scanning into the optimization framework.

Zefeng Zhu introduced FoldPolicy, which treats the backbone as a sequence of relative 3D rotations between peptide units and translates amino acid sequence to rotation sequence to perform protein structure prediction. Trained on protein ensembles and without explicit evolutionary information as input, FoldPolicy achieves accurate structure predictions of the size of typical protein domains and can also generate alternative conformations besides the ground state. This work demonstrates that appropriate geometric representations and probabilistic forms significantly reduce the complexity of protein conformation learning and sampling.

Jianfeng Pei highlighted that traditional ligand-based small molecule design methods often face intellectual property disputes. To address this, Pei developed TransPharmer, a generative model that combines small molecule structure information with pharmacophore data, achieving significant breakthroughs in generating novel scaffold molecules. Subsequently, a larger generative model based on the GPT (Generative Pre-trained Transformer) framework, PharmGPT, was trained on a larger database. PharmGPT has not only produced various novel small molecules recorded in the FDA (Food and Drug Administration) database but also confirmed their efficacy through experimental validation.

Han Wen introduced Uni-RNA, a large language model applied to RNA (Ribonucleic Acid) sequence analysis. Uni-RNA was developed to predict RNA secondary structures using high-quality training data and fine-tuning for downstream tasks. The model successfully enhances mRNA (messenger Ribonucleic Acid) vaccine development by optimizing 5′-UTR (Untranslated Region) sequences to increase protein translation efficiency and designing codon mutations to increase ribosome binding rates. Additionally, Uni-RNA, combined with Uni-MOL, identifies small molecules that can bind to RNA, enriching drug development across various dimensions. This approach harnesses the power of pre-trained models to unravel hidden knowledge in RNA sequences, significantly advancing research and nucleic acid drug discovery.

Overall, these speakers demonstrate that AI is not only efficiently solving the issue of structure prediction given sequences, but is now also actively and effectively generating novel compounds, sequences and structures based on desired properties, with the help of transfer learning and generative models (Table S1).

Deep learning in the field of biopharmaceuticals will continue to follow the development model based on big data. However, what is different is that the emergence of generative artificial intelligence, especially the generative large models, will further increase the demand for clean data volumes in this field, as well as produce new molecules and sequences beyond human cognition.

Multi-omics data and analysis

Multi-omics data aids in achieving precision medicine.

Jianfeng Feng discussed the challenges of analyzing structured, non-continuous data and reviewed existing statistical methods. Then he shared their large-scale, multi-scale brain science datasets, which span subcellular, cellular, and tissue levels. These datasets, capturing multi-spatial and multi-temporal scales, help elucidate the impact of lifestyle on brain function and contribute to the creation of a ‘digital twin’ of the human brain. He introduced a digital twin approach to modeling the human brain with 86 billion neurons and 100 trillion parameters.

Yaoyao Zhang utilized clinical resources to conduct multi-omics sequencing on embryonic samples. They established a comprehensive database incorporating single-cell and spatial transcriptomics. Through their collaboration, they explore molecular regulatory mechanisms in embryonic development and adverse pregnancy exposures. Their efforts include developing an integrated innovation platform for early screening technologies to reduce birth defect incidence.

Jing-Dong Jackie Han has been harnessing deep learning to study aging across various levels. Her team developed a 3D facial aging clock that deduces the molecular impact of lifestyle and predicts stroke risk. Their Thermal Face aging clock further uncovered ties between aging and metabolic diseases. At the cellular level, they developed the SenCID (Senescence Identification) tool to analyze single-cell senescence trajectories and discover transcription regulators that modulate cell senescence. They also developed single-cell aging clocks that revealed a molecular link between ribosomal levels and super longevity.

Zhengwei Xie focused on predicting drug effects at the transcriptomic level to address challenges in target-based drug design, such as multiple targets, off-target effects, and compensatory pathways. His DLEPS (Deep Learning–based Efficacy Prediction System) algorithm has successfully identified effective treatments for aging and metabolic diseases in mouse models. Xie is developing new metrics, Ordered Values (OV) for network inference and iterating on the DLEPS framework by incorporating protein-small molecule binding data to predict drug targets.

Yang-Yu Liu introduced his work on the application of deep learning techniques in microbiome research. For example, cNODE (compositional neural ordinary differential equation) can implicitly learn the community assembly rules and predict microbial compositions based on species presence/absence patterns. mNODE (Metabolomic profile predictor using Neural Ordinary Differential Equations) integrates microbiome data with individual diets to predict metabolomic profiles and effectively study the interactions between microbial species and metabolites. McMLP (Metabolite response predictor using coupled Multilayer Perceptrons) predicts metabolite response using baseline gut microbiome data before dietary interventions, which can be leveraged to prescribe personalized diets to achieve optimal metabolic response. These approaches highlight the significant role of deep learning in microbial ecology, metabolic modeling, and precision nutrition.

https://doi.org/10.1007/s11427-024-2859-1

Zhiyuan Li introduced siderophores, the iron-binding molecules produced by microbes to scavenge iron from the environment. By creating an extensive database of iron carriers and receptors, Li's team developed the 'iron.net' to map siderophore production, recognition, and uptake across various microbes. They discovered significant differences in iron acquisition behaviors between non-pathogenic and pathogenic microbes, the latter often avoid iron sharing, leading to severe infection outcomes.

Fulai Jin introduced DeepLoop, an advanced algorithm designed to efficiently analyze 3D chromatin structures and gene regulation from Hi-C (High-throughput chromosome conformation capture) sequencing data. DeepLoop improves the resolution of chromatin loops by correcting biases and enhancing signals through deep learning, even with low-depth Hi-C data. It enables single-cell Hi-C analysis and achieves consistency across different Hi-C protocols and micro-C methods. The algorithm has facilitated the mapping of genetic and epigenetic factors influencing allele-specific chromatin interactions, identifying new loci affected by imprinting and DNA (Deoxyribonucleic Acid) methylation. Additionally, DeepLoop detects heterozygous SNPs (Single Nucleotide Polymorphisms) and structural variants affecting chromatin loops, which impact gene expression.

Hao Li shared his team's work on identifying transcription factors with potential rejuvenation effects through large-scale experimental screening. Their research pinpointed several key factors that, when tested, were shown to be able to reverse a number of aging hallmarks in old human fibroblast cells. These findings provide promising avenues for developing therapies aimed at reversing aging and improving cell function.

In the single-cell field, several scholars shared their latest work.

Mohammad Lotfollahi discussed the advancements in single-cell atlases and multimodal, multiscale alignment algorithms that accelerate the identification of key cell types and biological processes in disease progression. He highlighted several tools they developed to predict the impact of novel perturbations on transcriptome expression and cell morphology, including the effects of new small-molecule drugs. Additionally, Mohammad introduced a novel spatial transcriptomics tool that leverages cell-cell communication data to identify spatial functional domains, offering new insights into cellular interactions and spatial organization.

Xuegong Zhang first introduced the hECA (human Ensemble Cell Atlas) database, a standardized cell-centric assembled human single-cell transcriptome database. Recognizing that databases alone are just the start for achieving cell digital twins, his team developed the scFoundation to learn the underlying logic of cell systems. They also designed the generative model scMulan on optimized datasets to simulate cells' responses to varied conditions as an attempt toward the final goal of building AI patients.

Ge Yang initially focused on biomedical image processing, and observed interactions between the endoplasmic reticulum and mitochondria in their intracellular segmentation work. However, the advent of large models inspired his team to embrace a new research paradigm. They developed GeneCompass, a knowledge-informed foundation model trained on 120 million single-cell RNA-seq datasets from humans and mice, to decipher gene regulatory mechanisms.

Shihua Zhang introduced a series of STA-tools for spatial and single-cell transcriptomics analysis. His team developed STAGATE to identify spatial domains by learning low-dimensional latent embeddings through the integration of spatial information and gene expression profiles; STAligner and BrainAlign to integrate and align spatial transcriptomics datasets across different conditions, technologies, developmental stages and species; STA-Marker to identify spatially domain-specific variable genes using saliency maps; STAGE, a spatial location-supervised auto-encoder generator for generating high-density spatial transcriptomics; STASCAN to decipher fine-resolution cell-distribution maps in spatial transcriptomics; and STALocator to achieve spatial transcriptomics-aided localization for single-cell transcriptomics.

Ge Gao introduced their deep learning tools for cross-sample, cross-platform, and cross-modality alignment. They developed Cell BLAST to align single-cell transcriptomes from different sources, GLUE (Graph-Linked Unified Embedding) to align different modalities and learn regulatory relationships, and SLAT (Spatial-Linked Alignment Tool) to integrate spatial transcriptomics by aligning different slices. Aiming at a holistic model for the gene regulatory system, Gao's team is working on a Causality-oriented Regulatory Language Model that will enable a genome-wide *in silico* simulation and, eventually, a rational design for cellular programming/reprogramming.

Pulin Li developed IRIS (Intracellular Response to Infer Signaling state), a semi-supervised deep learning method to infer individual cell signaling histories. By conducting a multiplexed scRNA-seq screen on hESCs (human Embryonic Stem Cells) and training IRIS on mESC (mouse Embryonic Stem Cell) data, they achieved accurate predictions for hESC-derived cell types and recovered signaling histories in mouse embryos. This revealed that signaling pathways induce transcriptional response signatures shared across cell types and species, enabling cross-dataset inference.

Qi Liu introduced the concept of weak supervision and its prevalence in adaptive immune receptor analysis. His team has developed several computational tools: PanPep addresses incomplete antigen-TCR (T Cell Receptor) binding annotations, TCRBagger tackles inexactly and inaccurate TCR annotations, and Uni-TCR leverages multimodal data, enhancing TCR sequencing analysis with transcriptomic data. These tools demonstrate AI solutions to decode the immune system and predict antigen-TCR bindings based on immunomics data.

Overall, these speakers demonstrated that deep learning models can be applied to various types of biological big data, from the epigenome, transcriptome, microbiome to phenome, in particular single-cell transcriptome data have provided unprecedented opportunities for training large models, including foundation and generative models, to uncover design principles of life at and across multi-modalities (Table S2). These foundation models, pretrained on tens of millions of human scRNA-seq data, have demonstrated the ability to capture complex gene expression relationships and achieve state-of-the-art performance in various single-cell analysis tasks, paving the way for advancements in biomedical research and healthcare applications.

Although there have been many large models in the field of single-cell research, most are limited to transcriptomics and are confined to two species: humans and mice. At the same time, the lack of detailed cleaning and processing of vast amounts of data could lead models to

learn incorrect knowledge. Therefore, it is necessary to have more modalities of data and to integrate accumulated knowledge into model training to give large models the opportunity to learn the true language of biological regulation and achieve digital twinning (Hao et al., 2024).

Conclusion

The conference provided valuable insights into the future development trends and potential challenges in the application of AI to biomedical big data.

Interpretable Artificial Intelligence. Deep learning is profoundly enhancing our grasp of fundamental biological principles. For example, Professor Vincenzo Vitelli demonstrated how integrating artificial intelligence with physical modeling can offer deeper insights into complex biological systems. However, the inherent 'black-box' nature of AI poses significant challenges, creating a gap between theoretical understanding and practical application. This opacity in AI models can obscure how predictions are generated, potentially hindering the interpretability of results. Looking ahead, Professor Luhua Lai is particularly interested in exploring whether AI can advance our understanding of protein fold spaces and the intricate relationships between sequence, structure, and function. This exploration aims to unlock new dimensions in protein biology by leveraging AI to uncover patterns and insights that might otherwise remain hidden.

Big Data and Large Models. The rise of big data and large models is revolutionizing research paradigms across various biological fields. Large models are now being employed to analyze small molecules, nucleic acids, proteins (Zhang et al., 2025), and single-cell transcriptomics (Hao et al., 2024), pushing the boundaries of traditional research approaches. This shift challenges smaller-scale research teams that may lack the resources to keep pace with these advancements. However, this transformation introduces several new issues. For instance, managing and sharing vast amounts of data becomes increasingly complex, and computational resource constraints can limit the ability to perform extensive analyses. Meanwhile, the sequential nature of the current LLM design may not necessarily align well with the sophisticated nonlinearity structure of multi-omics data (Chen et al., 2024). Additionally, there is a lack of comprehensive evaluation standards to effectively compare different models, which complicates the assessment of model performance and accuracy. Addressing these challenges is crucial for advancing the field and ensuring that the benefits of big data and large models can be fully realized in quantitative biology research.

Compliance and ethics

The authors declare that they have no conflict of interest.

Supporting information

The supporting information is available online at https://doi.org/10.1007/s11427-024-2859-1. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Chen, Z., Wei, L., and Gao, G. (2024). Foundation models for bioinformatics. Quant Biol 12, 339–344.
Hao, M., Wei, L., Yang, F., Yao, J., Theodoris, C.V., Wang, B., Li, X., Yang, G., and Zhang, X. (2024). Current opinions on large cellular models. Quant Biol 12, 433–443.
Zhang, Q., Ding, K., Lv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., et al. (2025). Scientific large language models: a survey on biological & chemical domains. ACM Comput Surv 57, 1–38.